2019年度「専修学校による地域産業中核的人材養成事業」

AIシステム開発教材

Society5.0実現のためのIT技術者養成モデルカリキュラム開発と実証事業

2019年度「専修学校による地域産業中核的人材養成事業」

AIシステム開発教材

Society5.0 実現のための IT 技術者養成モデルカリキュラム開発と実証事業

目 次

第1回：AIシステムの構成要素

システム全体像の理解

アジェンダ

- AIシステムの事例
- AIシステムの構成要素
  - データ収集機能
  - モデル学習機能
  - リリース管理機能
  - 判定実施機能
  - 判定後業務機能への連携

全15回の講義について

- 具体的なケーススタディ（I：農作物市況予測、II：対話分類）を通して、機械学習モデルだけではないAIシステムの構成要素について学習する。またAIシステム構築のステップを実践する。
  - プログラミング言語としてはPython
  - Pythonの各種ライブラリを利用してデータ分析に必要なスキルの習得を目指す

AIシステムの事例

人事分野への応用

- 自然言語処理により、応募者が記載したエントリーシートの内容を評価する事例です。
- 人事担当者の負担の軽減、評価基準の統一などの効果があります。

プレスリリース 2017年

いいね！ ツイート　一覧へ戻る | < 前のリリース | 次のリリース >

新卒採用選考におけるIBM Watsonの活用について

2017年5月29日
ソフトバンク株式会社

ソフトバンク株式会社は、応募者をより客観的に、また適正に評価することを目的に、2017年5月29日より新卒採用選考のエントリーシート※評価にIBM Watson日本語版（以下「IBM Watson」）を活用します。

過去のデータを学習させたIBM Watsonに応募者のエントリーシートデータを読み込ませると、IBM WatsonのAPIの一つであるNLC（Natural Language Classifier、自然言語分類）により、エントリーシートの内容が認識され、項目ごとに評価が提示されます。合格基準を満たす評価が提示された項目については、選考通過とし、それ以外の項目については人事担当者が内容を確認し、合否の最終判断を行います。IBM Watsonによる評価をエントリーシート選考の合否判断に活用することで、統一された評価軸でのより公平な選考を目指します。

出展：https://www.softbank.jp/corp/group/sbm/news/press/2017/20170529_01/

業務効率化

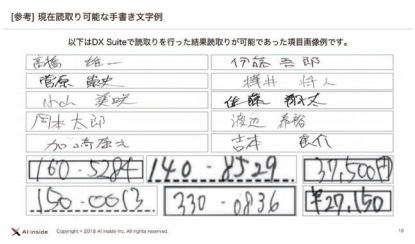
- OCR（Optical Character Recognition：光学的文字認識）により手書き文字をAIが認識する事例です。
- 事務作業の効率化ができます。

AI inside沿革

2017年

- 2017年1月 電通テック様とのキャンペーンサポートサービスにおける協業開始
- 2017年1月 東京海上日動火災様保険支払いの迅速化を目的として、AI（当社Intelligent OCR）による手書き請求書の読取りを実施
- 2017年3月 「金融イノベーションビジネスカンファレンスFIBC2017」 にて審査員特別賞を受賞
- 2017年5月 株式会社レオパレス21様 AIを活用したIntelligence OCR技術を導入
  - ・年間20,900時間の作業時間の削減
  - ・4,200万円のコスト削減

出展：https://rpa-bank.com/pdf/document-dl/ai_inside.pdf



新商品開発

- 保険会社が新商品開発にAIを使用した事例です。
- 保険会社は新しい商品の売上が増え、保険を契約する顧客は機器の保守コストを抑えることができる事例です。

ニュース

東京海上日動と日立が保険サービスを共同開発、IoTとAIで予兆診断

山端 宏実＝日経 xTECH／日経コンピュータ　日経 XTECH

この記事を評価する

この記事は　仕事に役立った 0　人に勧めたい 0　難しい 0　易しい 0

東京海上日動火災保険と日立製作所は2019年1月16日、データを活用した保険サービスを共同開発することで合意したと発表した。東京海上日動が持つ事故データや保険サービスと、日立のIoT（インターネット・オブ・シングズ）やAI（人工知能）などの技術を組み合わせ、新たな保険サービスを生み出す狙いだ。

第1弾として月内に、日立のIoTやAIによる予兆診断技術を活用し、製造設備の異常の兆候をつかみ、検査などに必要な費用を補償する保険の提供を始める。従来の保険は物的損壊を補償の要件にしていた。設備の異常の兆候を検知し、あらかじめ対策を講じることができれば企業は損失を最小限に抑えられ、保険会社が支払う補償額も少なく済む利点がある。

出展：https://tech.nikkeibp.co.jp/atcl/nxt/news/18/03854/

販促活動

- アサヒビールが、自社製品を販売してくれる小売企業（スーパー）に対して値付けサービスを提供するという事例です。
- アサヒビールを取り扱う小売企業が潤えば、アサヒビールも潤うというサプライチェーン全体を見据えた取り組みの事例です。

アサヒ、AIでビール売価指南　販売てこ入れ

2018/8/21 20:58 | 日本経済新聞　電子版

保存　共有　印刷　その他

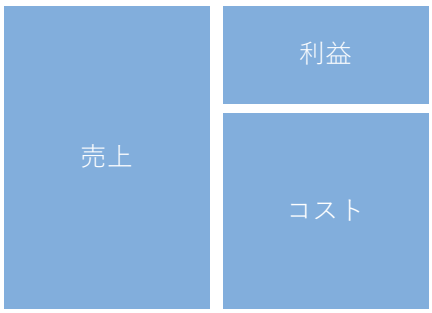
アサヒビールは、スーパーなどがビール系飲料の収益を効率良く上げるため、最適な値付けを提案できるシステムを売り込む。人工知能（AI）が販売環境などを踏まえ予測する。2017年6月の酒の安売り規制強化後、販売競争は激しい。小売店はライバル店の動きを横目に原価を下回らずに値ごろ感をどう出すか悩んでいる。競争力のある価格設定を手助けし、小売りとの取引拡大につなげる。

システムはNECのAI技術を組み込み、同社と共同開発した。小売店からPOS（販売時点情報管理）データに加え、天気や気温、運動会といった周辺イベント情報の提供を受けAIに入力する。入力データをもとに小売店の売り上げや利益を効率良くあげるにはどんな商品を、どの時期にどんな価格で売ればよいかAIがはじき出す。

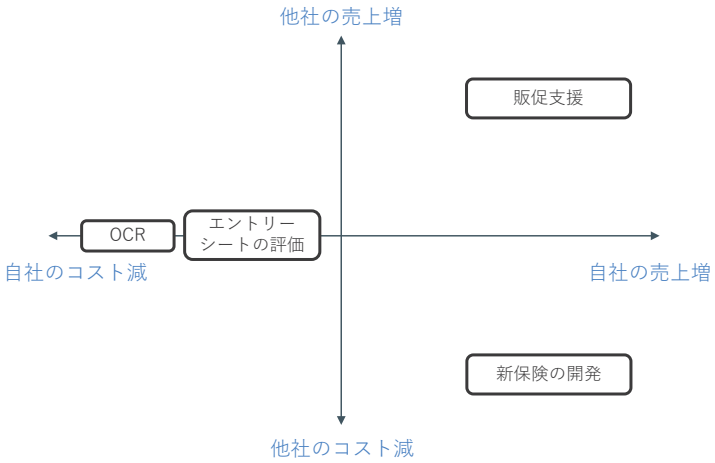
出展：https://www.nikkei.com/article/DGXMZO34407380210820189l6M00/

AIをどこで使うか

- 企業活動においてAIを適用する効果として①売上が増えるのか、もしくは②コストが減るのかのどちらかに大別できます。



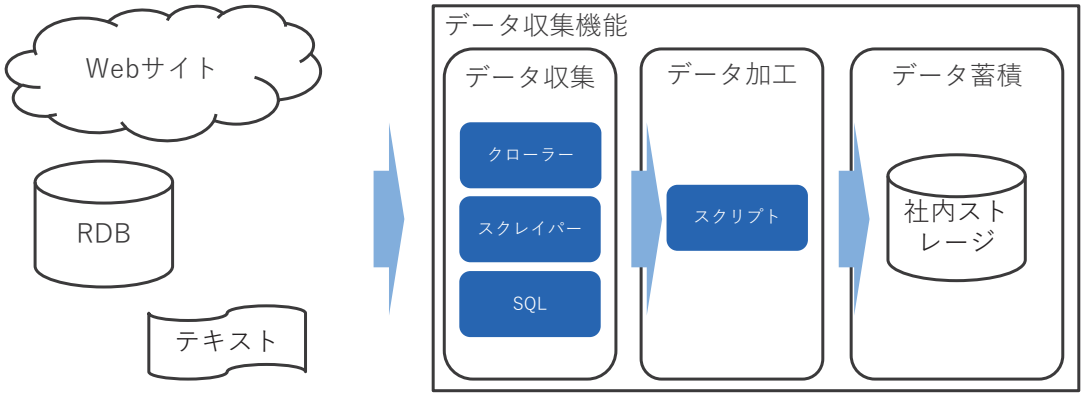
企業における収支のバランス



AIシステムの構成要素

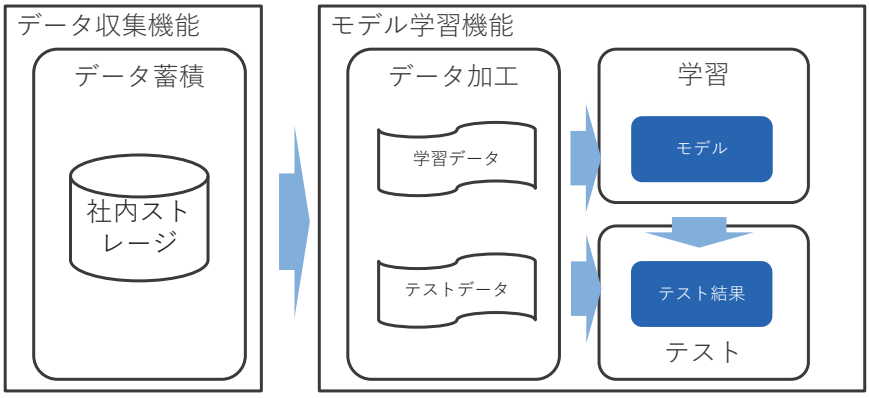
データ収集機能

- AIモデル作成に必要なデータを収集、加工、蓄積する機能です。



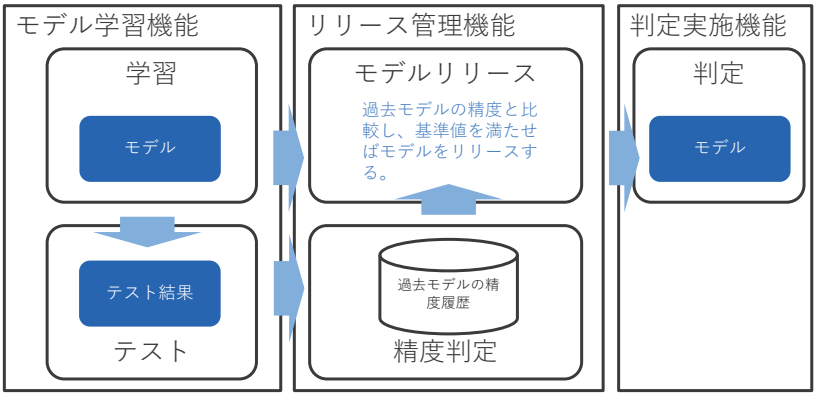
モデル学習機能

- 収集したデータを元に、AIモデルを作成します。
- 作成したモデルの精度を評価します。



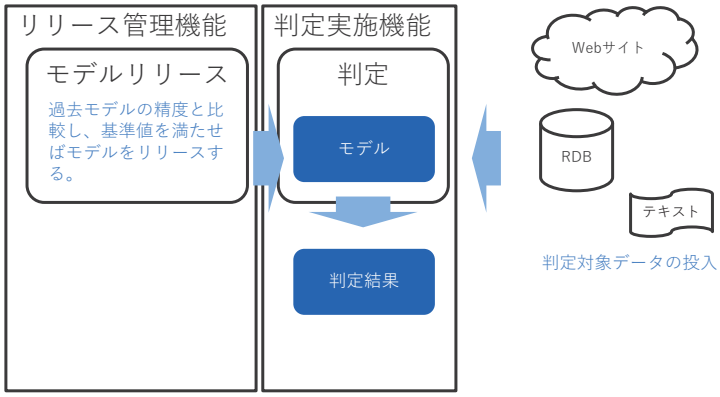
リリース管理機能

- 機械学習などで作成したモデルの精度を検証します。
- 実サービスへのリリース基準を持たしていれば、モデルをリリースします。



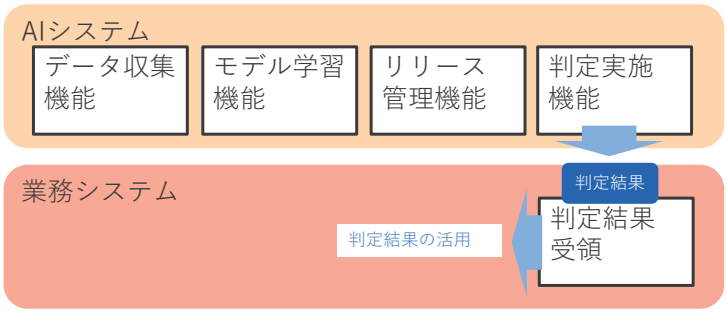
判定実施機能

- 作成したモデルを使用し、判定対象データの判定結果を出力します。



業務機能への連携

- 既存の業務システムにおいてAIを活用したい場合、業務システムの外側にAIシステムが構築されるケースが多いです。
- AIシステムの判定結果を業務システムに連携し、業務システムで活用します。
- 業務システムとAIシステムの関係を疎にすることによって、いずれかのシステムの改修、取替が容易になります。



第2回：AIシステム構築プロジェクトのステップ

ステップごとの実施事項と留意点

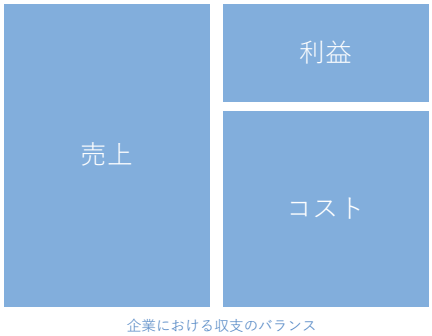
アジェンダ

- 第1回目講義の振り返り
- AIシステム構築プロジェクトのステップ
  - 営業
  - ヒアリング
  - 事前調査
  - 要件定義
  - 運用設計
  - 事前分析
  - PoC
  - システム設計
  - システム構築

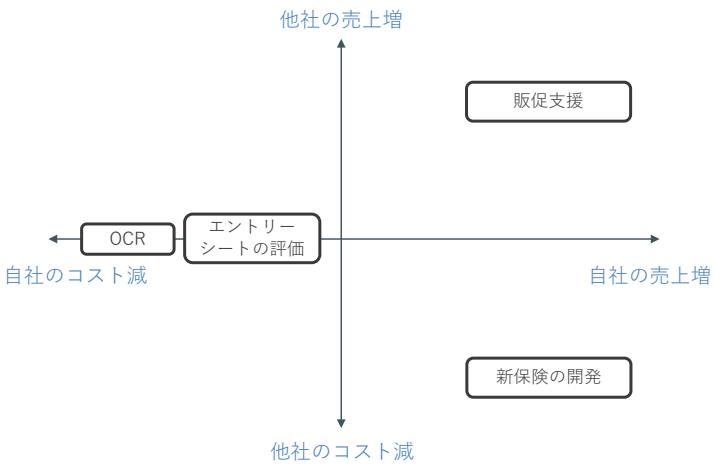
第1回講義の振り返り

AIをどこで使うか

- 企業活動においてAIを適用する効果として①売上が増えるのか、もしくは②コストが減るのかのどちらかに大別できます。

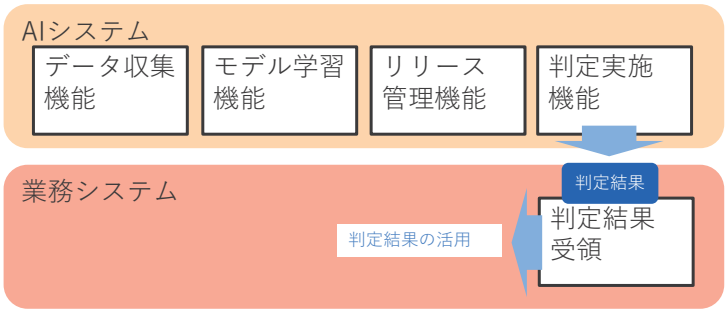


企業における収支のバランス



業務機能への連携

- 既存の業務システムにおいてAIを活用したい場合、業務システムの外側にAIシステムが構築されるケースが多いです。
- AIシステムの判定結果を業務システムに連携し、業務システムで活用します。
- 業務システムとAIシステムの関係を疎にすることによって、いずれかのシステムの改修、取替が容易になります。



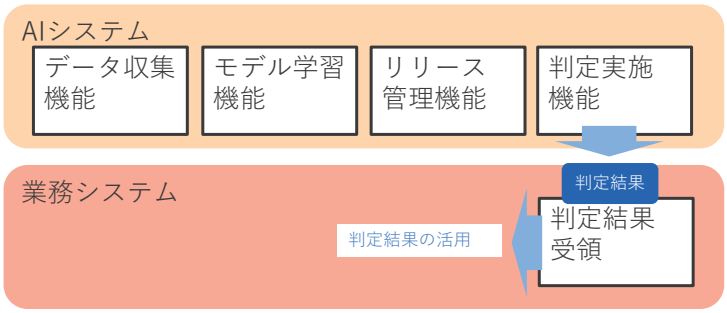
AIシステム構築プロジェクトのステップ

営業

- AIシステムを構築する場合、社内にAIを構築できる人材がいれば自社開発（Case1）を行います。社内にAIを構築できる人材がいない場合、社外に発注（Case2）することになります。
- Case2の場合、「AIを構築できる会社」は「AIを使いたい会社」に営業活動をすることになります。「AIを構築できる会社」の評判が良い場合、「AIを使いたい会社」からシステム構築を依頼されるケースもあります。

ヒアリング

- AIを使いたい会社・ユーザはAIを使って何を実現したいのか、ヒアリングを行います。
- 特に既存の業務システムとの連携を考えている場合は、業務システムの仕組み上実現が難しい場合や費用が大きくなる場合もあるため、導入コストと得られるメリットの双方を勘案する必要があります。
- AIを導入しなくてもユーザのやりたいことが実現できると判断する場合や、コストが大きすぎて導入を避けた方や良い場合は、プロジェクト自体の中断を決断することになります。



事前調査

- AI導入にメリットがあると判断した場合、データ・技術の事前調査へと進めます。
- 機械学習などを用いてAIモデルを作成するためには、データの量と質がモデルを構築するに足るかどうかを判断する必要があります。

- データの量と質が基準を満たしている場合でも個人情報が含まれていたり、コンプライアンス（法令遵守）の視点からデータの使用が困難であったりする場合があります。コストを投入する前に様々な角度からプロジェクトの実現性を検証する必要があります。

要件定義

- まずはシステムが満たすべき条件を決定します。例えばモデルのインプットデータの項目と量、モデルのアウトプットの項目、モデルの精度などです。
- モデル以外にも蓄積するデータの量と、蓄積の速度・期間などについて決定します。

- システム以外にも、サービスとして満たすべき条件を決定します。例えばユーザーからのリクエストに対して何秒以内に応答を返さないといけないか、一日何回の使用に耐えなければならないか、システム運用のコストと売上の損益分岐をどこに定めるかなどです。

- 実務においては後述する運用設計・事前分析・PoCを数回繰り返して要件を決定していきます（実際にやってみて発覚することがあるため、このような繰り返しのやり方が必要となります）。

運用設計

- AIモデルを実サービスとして展開し続けるためには、運用が必要になります。
- 増え続けるデータの蓄積、モデルの更新、不具合への対応にどれだけの人・機材・お金・時間を投入するのかを決定します。投入するこれらのコストを勘案し、ユーザに請求するサービス料の設定が必要となります。

- ユーザの予算が決まっていて多額の投資をできない場合、システム機能の削減や運用フローの簡素化を実施します。

事前分析

- モデル作成に使用するデータの詳細な調査を実施します。
- モデルの目的変数、説明変数の統計値を算出したり、目的変数と説明変数間で相関があるのかなどの検証を実施します。

- 本プロセスにおいてデータの大きな偏りが確認されたり、また説明変数と目的変数間に全く相関が確認されない場合、機械学習を導入しても効果が見られないと判断してプロジェクトを中断することもあります。このステップはプロジェクトの「勝ち目」を判断する重要なステップです。

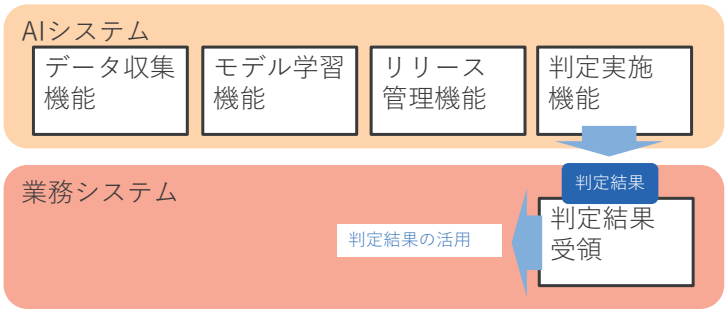
PoC（Proof of Concept：概念実証）

- AIモデルが本当に実現できるのか、実際に試してみるステップです。
- データ加工、機械学習アルゴリズムの試行、精度検証をいかに正確に積み上げていけるかがプロジェクト成否の分かれ目となります。機械学習の実施が効率化できるライブラリの使用や、独自スクリプトの構築が欠かせません。

- このステップで大事なのはモデルの精度向上ではありますが、要件定義で決定した事項を満たせているのか、ということを常にチェックしなければなりません。例えば「精度の向上を優先しすぎるあまりDeep Neural Networkをアルゴリズムとして選択してしまい、モデルの説明性が低下する。」という事態を避けなければなりません。

システム設計・システム構築

- モデルを作成した後は、データ収集からモデル判定結果の提供方法までを含めたシステム設計が必要となります。
- 設計後は実際にシステムを構築していきます・



第3回以降の講義について

- 実際のデータを用いた模擬AIシステム構築を実施します。
- テーマは2つです。前半は「農作物市況予測」について、後半は「対話分類」について実施します。

第3回：データを用いた事前調査

データ項目・量の確認

アジェンダ

- 農作物市況予測モデル構築プロジェクト（ケーススタディI）の設定について
- データを用いた事前調査

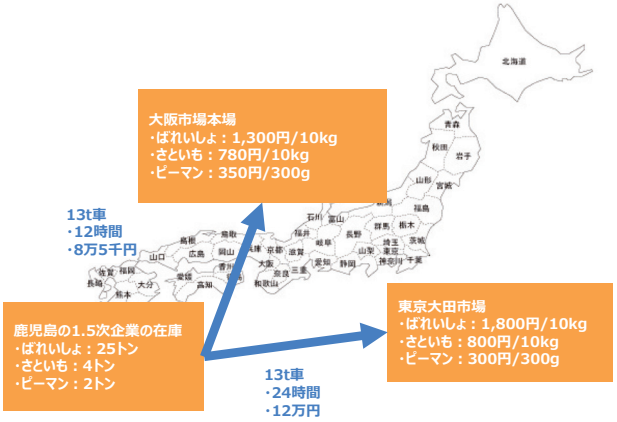
農作物市況予測モデル構築プロジェクト（ケーススタディⅠ）の設定について

仕事の依頼

- 株式会社Lは金融工学が得意なベンチャー企業で、数学や工学が得意なメンバーが在籍しています。株式会社Lは、機械学習を金融分野に適用する仕事の実績があります。
- そんな株式会社Lに、農業分野の仕事の依頼がありました。中央卸売市場における"ばれいしょ（じゃがいも）"の値動きを可視化できるツールが欲しいという依頼です。
- 野菜市場についての知見が無い株式会社Lは、ひとまず依頼先の会社を訪問して詳しい話を聞くことにしました。

ヒアリング

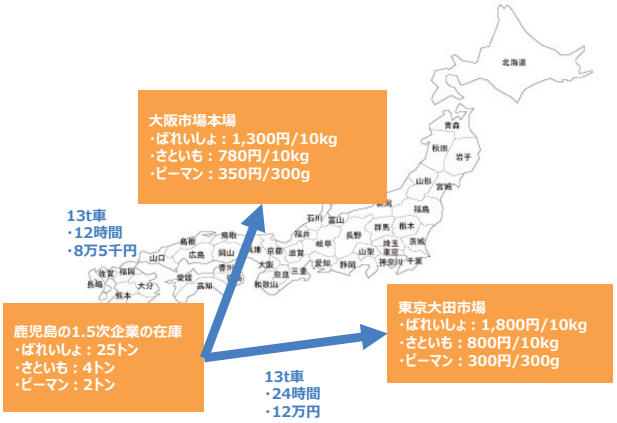
- 仕事を依頼してきたのは大手商社の子会社であるK株式会社でした。K株式会社は九州に拠点があり、九州で産地から買い集めた野菜を関東・関西の中央卸売市場に転売する仲卸業を行っているそうです。
- 例えば春先であれば、その時期に取れるばれいしょ・さといも・ピーマンを中央卸売市場に出荷します。それぞれの市場で買い取り価格が異なりますし、輸送コスト（時間と金額）も異なるそうです。



営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → PoC → システム設計 → システム構築

ヒアリング

- K株式会社の最終的な目的は、「いつどの市場に、どの野菜を何トン出荷すれば利益が最大化できるかを予測するモデルの構築」ということでした。
- その話を聞いた株式会社Lのプロジェクトリーダーは、スコープ（1つの契約で実施する作業の範囲）をいくつかに切ることを提案しました。
  - スコープ1：市場における野菜の数量・価格を見える化する仕組みを構築する。
  - スコープ2：市場における野菜の価格を予測するモデルを構築する。
  - スコープ3：出荷先と出荷量を最適化するモデルを構築する。



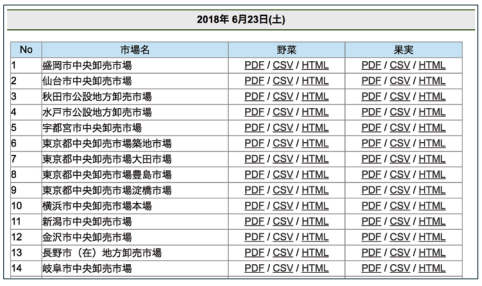
営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → PoC → システム設計 → システム構築

質問：スコープの設定

- 「いつどの市場に、どの野菜を何トン出荷すれば利益が最大化できるかを予測するモデルの構築」を実施したいと行ったK株式会社に対して、株式会社Lはスコープを3つに切ることを提案しました。
- これは、株式会社Lは1つの契約で大きな金額を得るチャンスがあったにも関わらず、わざわざ契約を3つに分けて受注のコストを増やしたことを意味します。
- なぜ株式会社Lはスコープを3つに切ったのか、議論してみてください。

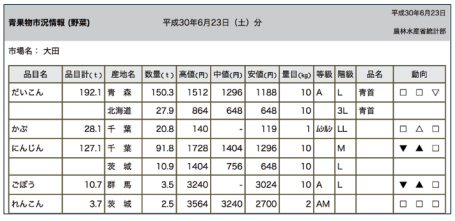
事前調査

- 株式会社Lは、最初のスコープである「市場における野菜の数量・価格を見える化する仕組みを構築する。」の実現のため、データの調査を開始しました。
- 農水省HPに中央卸売市場のデータが公開されていることを知った株式会社Lは、データの調査を開始することにしました。



2018年 6月23日(土)

No	市場名	野菜	果実
1	盛岡市中央卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
2	仙台市中央卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
3	秋田市公設地方卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
4	水戸市公設地方卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
5	宇都宮市中央卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
6	東京都中央卸売市場築地市場	PDF / CSV / HTML	PDF / CSV / HTML
7	東京都中央卸売市場大田市場	PDF / CSV / HTML	PDF / CSV / HTML
8	東京都中央卸売市場豊島市場	PDF / CSV / HTML	PDF / CSV / HTML
9	東京都中央卸売市場淀橋市場	PDF / CSV / HTML	PDF / CSV / HTML
10	横浜市中央卸売市場本場	PDF / CSV / HTML	PDF / CSV / HTML
11	新潟市中央卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
12	金沢市中央卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
13	長野市（在）地方卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
14	岐阜市中央卸売市場	PDF / CSV / HTML	PDF / CSV / HTML



青果物市況情報（野菜）　平成30年6月23日（土）分　平成30年6月23日　農林水産省統計部

市場名：大田

品目名	品目計(t)	産地名	数量(t)	高値(円)	中値(円)	安値(円)	量目(kg)	等級	階級	品名	動向
だいこん	192.1	青　森	150.3	1512	1296	1188	10	A	L	青首	□ □ ▽
		北海道	27.9	864	648	648	10		3L	青首	
かぶ	28.1	千　葉	20.8	140	-	119	1	ﾊﾞﾗ	LL		□ △ □
にんじん	127.1	千　葉	91.8	1728	1404	1296	10		M		▼ ▲ □
		茨　城	10.9	1404	756	648	10		L		
ごぼう	10.7	群　馬	3.5	3240	-	3024	10	A	L		▼ ▲ □
れんこん	3.7	茨　城	2.5	3564	3240	2700	2	AM			□ □ □

出展：農林水産省HPより

演習1：データ項目の確認

- 野菜市況データに含まれる項目を確認してください。

データの確認

- データを確認すると下図のような項目を持っていることがわかりました。市場ごとにCSVファイルが1つで、その中に複数の品目と産地、数量と金額（高値・中値・安値）が含まれていることが確認できました。

	年	月	日	曜日	市場名	市場コード	品目名	品目コード	産地名	産地コード	品目計	入荷量	高値	中値	安値	等級	階級	品名	量目	動向
0	2018	1	15	月	築地	13300	だいこん	30100	神奈川	14.0	36.8	20.8	3240	2376	1836	シユウ	L	青首	10.0	弱保合
1	2018	1	15	月	築地	13300	だいこん	30100	千　葉	12.0	NaN	7.9	3024	2365	1836	NaN	L	青首	10.0	NaN
2	2018	1	15	月	築地	13300	だいこん	30100	徳　島	36.0	NaN	3.8	3024	2592	2268	シユウ	L	青首	10.0	NaN
3	2018	1	15	月	築地	13300	だいこん	30100	鹿児島	46.0	NaN	2.9	-	1361	-	NaN	L	NaN	10.0	NaN
4	2018	1	15	月	築地	13300	かぶ	30200	千　葉	12.0	6.0	4.9	3564	-	2916	ベツ	LL	NaN	12.0	強い

演習2：品目と産地の確認

- 1つのCSVファイルの中に、どれくらいの品目・産地が含まれているのか確認してください。

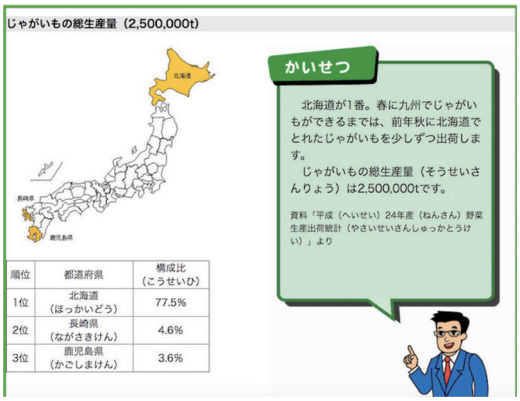
品目と産地の確認

- 品目と産地の組み合わせを確認してみると、1つの市場で100近くの種類の野菜を取り扱っていることがわかりました。
- 株式会社Lはこの結果をK株式会社に報告しました。ディスカッションの結果、まずは春先（1月～4月）に出荷されるばれいしょに焦点を当てて調査を続行することとなりました。

	品目名	産地名
0	えだまめ	静岡
1	えのきだけ	新潟
2	えのきだけ	長野
3	かぶ	千葉
4	かぶ	埼玉
5	かぼちゃ	メキシコ
6	かんしょ	千葉
7	かんしょ	徳島
8	きゅうり	埼玉
9	きゅうり	宮崎
51	はくさい	兵庫
52	はくさい	茨城
53	ばれいしょ	北海道
54	ばれいしょ	長崎
55	ほうれんそう	埼玉
56	ほうれんそう	群馬
57	ほうれんそう	茨城
88	レタス	兵庫
89	レタス	愛知
90	レタス	福岡
91	レタス	静岡
92	根しょうが	高知
93	生しいたけ	北海道
94	生しいたけ	岩手
95	生しいたけ	栃木
96	生しいたけ	秋田

ばれいしょ出荷の季節性

- 株式会社Lが調査したところ、日本におけるばれいしょの3大産地は北海道、長崎、鹿児島の3都県だということが明らかになりました。
- K株式会社の話では、春先（1月～4月）に出荷されるばれいしょは産地ごとに出荷時期がずれており、その差を利用して利益を上げることが可能だと言うことです。
- 株式会社Lは、K株式会社からの情報が正しいのか、実際のデータで確認することにしました。



出展：http://www.maff.go.jp/j/kids/crops/potato/farm.html

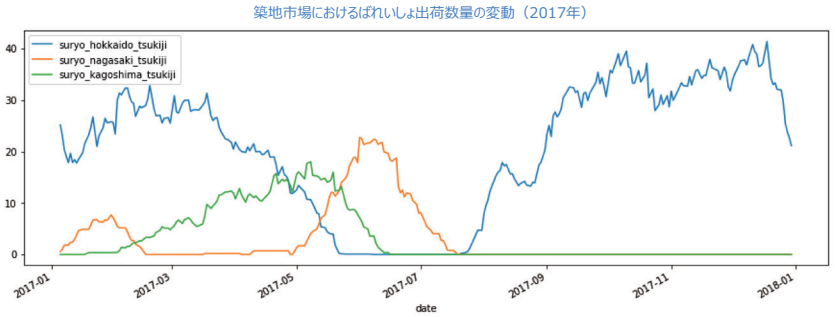
営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → PoC → システム設計 → システム構築

演習3：ばれいしょ出荷時期の確認

- 北海道産、長崎産、鹿児島産のばれいしょ出荷時期に差があるのか、データを確認してください。

ばれいしょ出荷時期の確認

- 株式会社Lがデータを確認した結果、北海道・長崎・鹿児島産ばれいしょの出荷時期には季節性があることが確認できました。
- 普段は金融分野の仕事をしている株式会社Lは、「市場に出回るばれいしょの量が変動するということは、価格の差益で利益を出すことができるのではないか」という感触を得て、次のステップに進むことにしました。



第4回：要件定義

システムが満たすべき性能とインターフェース

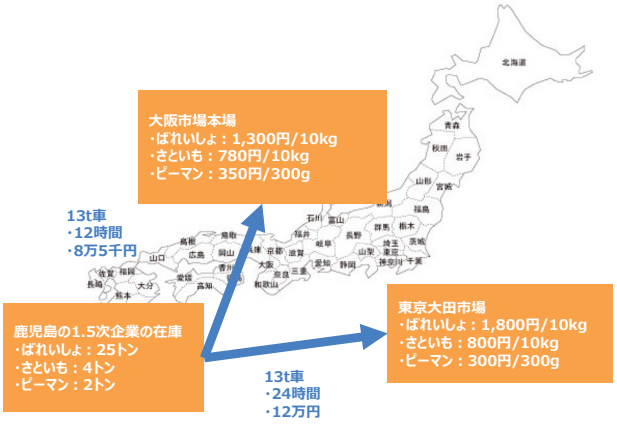
アジェンダ

- 前回までの講義の振り返り
- AIシステムの要件定義

前回までの講義の振り返り

前回までの振り返り

- 株式会社Lは、九州で産地から買い集めた野菜を関東・関西の中央卸売市場に転売する仲卸業を行っているK株式会社からの依頼を受け、市場における野菜の価格予測を行うシステムの構築を進めています。



営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → PoC → システム設計 → システム構築

前回までの振り返り

- K株式会社の最終的な目的は、「いつどの市場に、どの野菜を何トン出荷すれば利益が最大化できるかを予測するモデルの構築」ということでした。
- その話を聞いた株式会社Lのプロジェクトリーダーは、スコープ（1つの契約で実施する作業の範囲）をいくつかに切ることを提案しました。
  - スコープ1：市場における野菜の数量・価格を見える化する仕組みを構築する。
  - スコープ2：市場における野菜の価格を予測するモデルを構築する。
  - スコープ3：出荷先と出荷量を最適化するモデルを構築する。

- 株式会社Lは、データを用いた事前調査を終え、要件定義へと進むことにしました。

要件定義（第2回講義より）

- まずはシステムが満たすべき条件を決定します。例えばモデルのインプットデータの項目と量、モデルのアウトプットの項目、モデルの精度などです。
- モデル以外にも蓄積するデータの量と、蓄積の速度・期間などについて決定します。

- システム以外にも、サービスとして満たすべき条件を決定します。例えばユーザーからのリクエストに対して何秒以内に応答を返さないといけないか、一日何回の使用に耐えなければならないか、システム運用のコストと売上の損益分岐をどこに定めるかなどです。

- 実務においては後述する運用設計・事前分析・PoCを数回繰り返して要件を決定していきます（実際にやってみて発覚することがあるため、このような繰り返しのやり方が必要となります）。

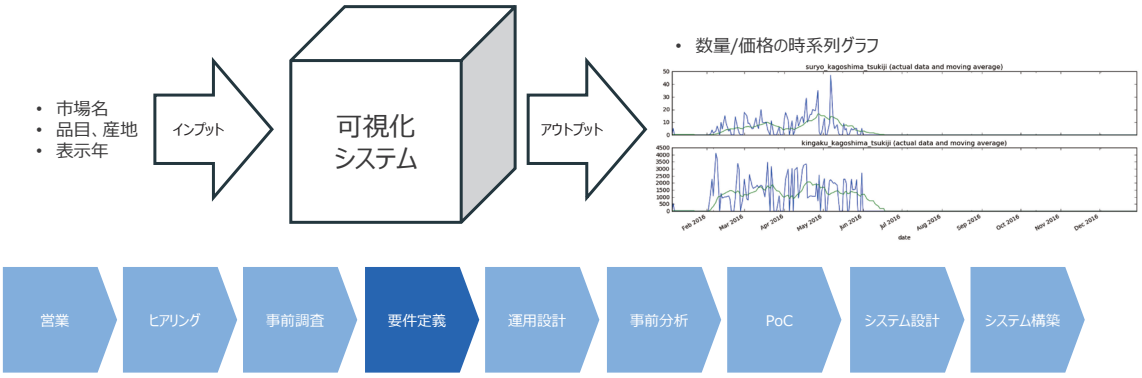
AIシステムの要件定義

使用するデータの決定

- 株式会社Lは事前調査において、農水省が公開している卸売市場の数量/価格データを確認しました。
- 農水省のデータは、市場・品目・産地ごとに数量/価格が取得できるデータでした。K株式会社からの要望を満たす可視化システムと（スコープ1）、野菜価格を予測するモデル（スコープ2）を構築するために最適なデータだと判断しました。

システムのインプットとアウトプット

- 「スコープ1：市場における野菜の数量・価格を見える化する仕組み」におけるシステムインプット/アウトプットの要件ついてディスカッションしています。
- アウトプットにおける「数量」と「価格」は、具体的にどのような数字を表示することが望ましいのか、株式会社Lが考えはじめました。



演習1：数量・価格の定義

- ばれいしょの数量・価格の表示パターンを確認してください。

数量・価格の定義

- 数量の指標として「入荷量」を適用することにしました。ただし、入荷量が記載れていないレコードは、その商品の入荷が少なかったことを表すことが判明したため、無視することにしました。
- 金額の指標として使用できそうな項目は高値・中値・安値の3項目存在します。調査の結果、中値が仕入量が最も多い製品の金額であることがわかったのですが、高値と安値を公開することが市場に与えるインパクトは無視できないということから、3項目のうち存在する数値の平均値を価格として採用することにしました。

	年	月	日	曜日	市場名	市場コード	品目名	品目コード	産地名	産地コード	品目計	入荷量	高値	中値	安値	等級	階級	品名	量目	動向
78	2018	1	15	月	築地	13300	ばれいしょ	36200	北海道	1.0	58.1	50.4	1080	-	972	シユウ	L	メクイン	10.0	弱保合
79	2018	1	15	月	築地	13300	ばれいしょ	36200	北海道	1.0	NaN	NaN	1404	1296	1188	シユウ	L	男爵	10.0	NaN
80	2018	1	15	月	築地	13300	ばれいしょ	36200	長　崎	42.0	NaN	7.6	-	864	-	シユウ	S	NaN	10.0	NaN
81	2018	1	15	月	築地	13300	ばれいしょ	36200	長　崎	42.0	NaN	NaN	756	-	648	シユウ	3S	NaN	10.0	NaN

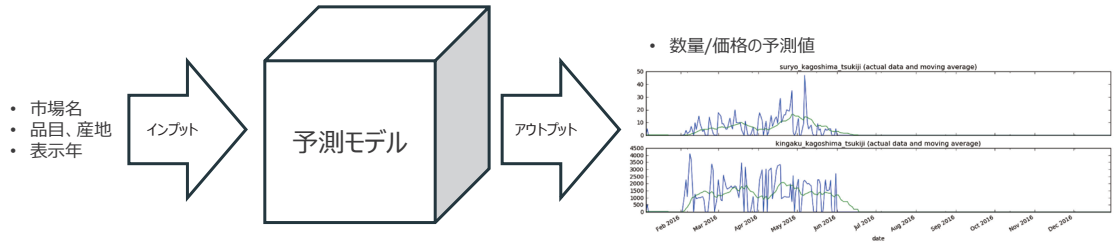
入荷量の記載がないレコードが存在する。

金額に関する数値は高値・中値・安値の3つある。

営業 → ヒアリング → 事前調査 → **要件定義** → 運用設計 → 事前分析 → PoC → システム設計 → システム構築

システムのインプットとアウトプット

- 「スコープ1：市場における野菜の数量・価格を見える化する仕組み」におけるアウトプットの定義は前ページのように決定しました。
- 「スコープ2：市場における野菜の価格を予測するモデル」のインプットはスコープ1と同じにし、アウトプットは、スコープ1の金額に対する予測値とすることが決まりました。



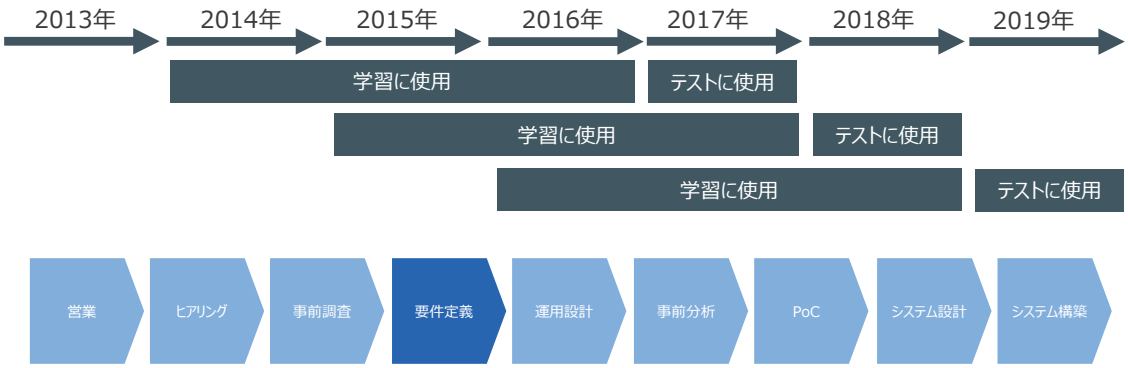
営業 → ヒアリング → 事前調査 → **要件定義** → 運用設計 → 事前分析 → PoC → システム設計 → システム構築

システムの応答速度

- 「スコープ1：市場における野菜の数量・価格を見える化する仕組み」については、K株式会社の担当者がインプットデータをシステムに投入してから数秒～10秒以内で画面表示できること、という目標を設定しました。
- 「スコープ2：市場における野菜の価格を予測するモデル」については、1日1回株式会社Kの担当者が閲覧できればいいということで、夜間バッチで予測結果を算出することになりました。

必要なデータ量について

- 「スコープ1：市場における野菜の数量・価格を見える化する仕組み」については、データを入手できた分だけ閲覧可能とする、と設定しました。
- 「スコープ2：市場における野菜の価格を予測するモデル」についても同様に、入手できる分のデータを全て使ってクロスバリデーションテストを実施し、モデルのパラメータをチューニングすることにしました。



第5回：データ準備

データ蓄積環境の整備

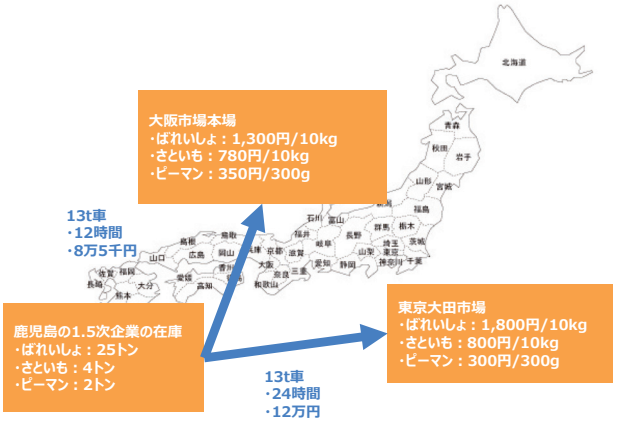
アジェンダ

- 前回までの講義の振り返り
- PoCのためのデータ準備

前回までの講義の振り返り

前回までの振り返り

- 株式会社Lは、九州で産地から買い集めた野菜を関東・関西の中央卸売市場に転売する仲卸業を行っているK株式会社からの依頼を受け、市場における野菜の価格予測を行うシステムの構築を進めています。



営業 → ヒアリング → 事前調査 → **要件定義** → 運用設計 → 事前分析 → PoC → システム設計 → システム構築

前回までの振り返り

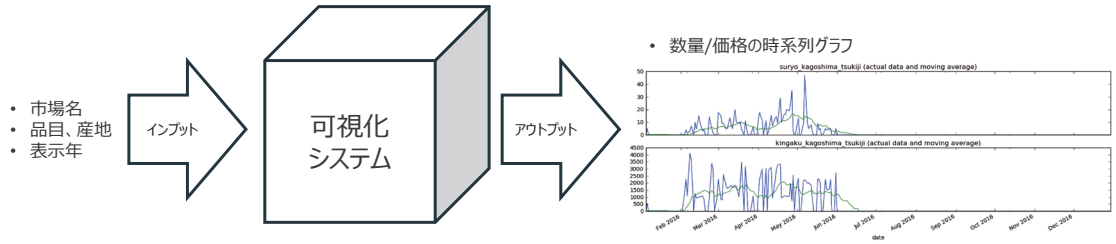
- K株式会社の最終的な目的は、「いつどの市場に、どの野菜を何トン出荷すれば利益が最大化できるかを予測するモデルの構築」ということでした。
- その話を聞いた株式会社Lのプロジェクトリーダーは、スコープ（1つの契約で実施する作業の範囲）をいくつかに切ることを提案しました。
  - スコープ1：市場における野菜の数量・価格を見える化する仕組みを構築する。
  - スコープ2：市場における野菜の価格を予測するモデルを構築する。
  - スコープ3：出荷先と出荷量を最適化するモデルを構築する。

- 株式会社Lは、スコープ1と2の検証を実施するために、データの蓄積・取り回し環境の準備へと進むことにしました。

PoCのためのデータ準備

データの蓄積

- CSVファイルは市場ごと、日毎に分割されていて、1ファイルの中に複数品目のデータが存在します。また、金額は高値・中値・安値から計算する必要があります。
- PoCにおいては様々な切り口でデータを抽出、加工する必要があるため、いったんCSVをRDB（リレーショナルデータベース）で管理することにしました。



演習1：データ蓄積量の見積もり

- 築地市場の1日のデータサイズを元に、必要なデータベースの容量を見積もってください。

SQLite

- データ蓄積量を見積もったところ、30市場・9年間のデータを蓄積しても700MB程度しか無いことが判明しました。
- ローカルPCでも簡易に構築できる軽量データベースであるSQLiteを採用することになりました。

SQLiteとは

SQLite(エスキューライト)： https://www.ossnews.jp/oss_info/SQLite

- 軽量コンパクトなリレーショナルデータベースシステムです。主に組み込み用途や、小規模システムのデータストアとして利用されます。
- 本体サイズは600KB程度と非常にコンパクトです。消費メモリも少ない点も特徴です。
- 外部依存関係がない自己完結型で、セットアップも不要です。

演習2：SQLite環境の準備

- SQLiteを利用してデータベースを作成し、データを蓄積するテーブルも作成してください。

演習3： テーブルへのデータ投入

- 作成したテーブルにデータを投入してください。
- SQLiteに投入したデータを確認してください。

第6回：データ蓄積とデータ加工

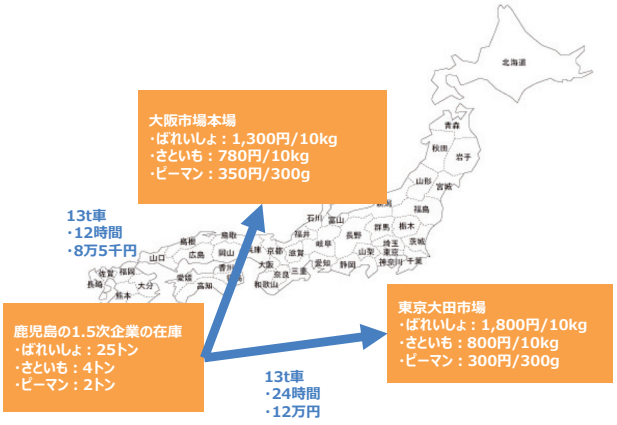
アジェンダ

- 前回までの講義の振り返り
- SQLを駆使したデータ加工のテクニック

前回までの講義の振り返り

前回までの振り返り

- 株式会社Lは、九州で産地から買い集めた野菜を関東・関西の中央卸売市場に転売する仲卸業を行っているK株式会社からの依頼を受け、市場における野菜の価格予測を行うシステムの構築を進めています。



営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → **事前分析** → PoC → システム設計 → システム構築

前回までの振り返り

- K株式会社の最終的な目的は、「いつどの市場に、どの野菜を何トン出荷すれば利益が最大化できるかを予測するモデルの構築」ということでした。
- その話を聞いた株式会社Lのプロジェクトリーダーは、スコープ（1つの契約で実施する作業の範囲）をいくつかに切ることを提案しました。
  - スコープ1：市場における野菜の数量・価格を見える化する仕組みを構築する。
  - スコープ2：市場における野菜の価格を予測するモデルを構築する。
  - スコープ3：出荷先と出荷量を最適化するモデルを構築する。

- 株式会社Lにて使用するデータ量を見積もったところ、30市場・9年間のデータを蓄積しても700MB程度しか無いことが判明しました。
- ローカルPCでも簡易に構築できる軽量データベースであるSQLiteを採用することになりました。

SQLを駆使したデータ加工のテクニック

データの確認

- SQLを使用したデータ加工を実施する際、まずはテーブルに格納されているデータを確認します。
- レコード数が多いテーブルに対して単純なselect文を実行してしまうと検索時間が長くなり、またCPUやメモリリソースを大量に消費してしまうので、好ましくありません。
- 検索レコード数にlimitを掛けることで、検索時間を抑え速やかにデータを確認することができます。

演習1：データの確認

- テーブルに格納されているデータを確認してください。
- 全件検索ではなく、limitで検索数を絞ってください。

検索データの絞り込み

- SQLのwhere句で、検索するデータを選別する条件を指定することができます。
- K株式会社とのミーティングにおいて、「まずはばれいしょの可視化システム・価格予測モデルを構築する」ということが決定しています。このような場合、まずはばれいしょのデータのみを抽出する条件分をwhereに記載します。
- 前回までの調査で明らかになっているばれいしょの3大産地（北海道・長崎・鹿児島）のみを扱うことにするため、産地を絞り込む条件もwhere句に記載します。

演習2：検索データの絞り込み

- ばれいしょのデータのみを抽出してください。

演習3：検索データの絞り込み

- ばれいしょの産地の数を確認してください。
- 産地ごとの数量を集計してください。

演習4：検索データの絞り込み

- 北海道、長崎、鹿児島産のばれいしょのみを抽出してください。

データ項目の横持ち

- 機械学習データセットを作成する際、説明変数は横持ち（1レコードに複数項目が存在すること）にすることが一般的です。
- SQLのselect句でcase構文を使用することで、項目を横持ちにすることができます。

	sanchicd	hinmokukei	suryo
北海道	1.0	30.6	16.3
長　崎	42.0		14.3

データ縦持ち

	sanchicd	hinmokukei	suryo		sanchicd	hinmokukei	suryo
北海道	1.0	30.6	16.3	長　崎	42.0		14.3

データ横持ち

演習5：データ項目の横持ち

- 北海道、長崎、鹿児島産のばれいしょ数量を横持ちにするSQLを作成し、検索結果を確認してください。

	sanchicd	hinmokukei	suryo
北海道	1.0	30.6	16.3
長　崎	42.0		14.3

データ縦持ち

	sanchicd	hinmokukei	suryo		sanchicd	hinmokukei	suryo
北海道	1.0	30.6	16.3	長　崎	42.0		14.3

データ横持ち

演習6：データ項目の横持ち

- 数量と金額を横持ちにして抽出してください。
- SQL中で条件絞り込み、group by、caseを多用しています。それぞれの書き方がどのようなデータ加工をするためのものなのか、確認しながら実行してください。

	date	suryo_hokkaido	kingaku_hokkaido	suryo_nagasaki	kingaku_nagasaki	suryo_kagoshima	kingaku_kagoshima
0	20180105	109.6	1512.0	0.0	0.0	0.0	0.0
1	20180106	0.2	1692.0	0.0	0.0	0.0	0.0
2	20180109	18.8	486.0	15.5	972.0	0.0	0.0
3	20180111	6.7	1332.0	15.2	342.0	0.0	0.0
4	20180112	29.8	1008.0	0.0	0.0	0.0	0.0

第7回：PoC

野菜数量と価格の考察

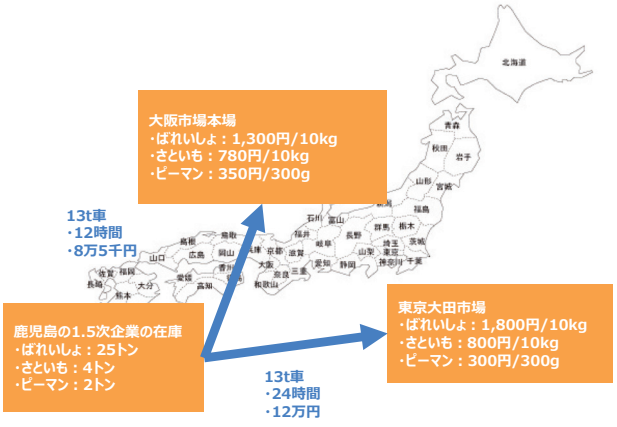
アジェンダ

- 前回までの講義の振り返り
- Pythonを用いたデータ可視化のテクニックと、可視化結果をうけての考察

前回までの講義の振り返り

前回までの振り返り

- 株式会社Lは、九州で産地から買い集めた野菜を関東・関西の中央卸売市場に転売する仲卸業を行っているK株式会社からの依頼を受け、市場における野菜の価格予測を行うシステムの構築を進めています。



営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → **事前分析** → PoC → システム設計 → システム構築

前回までの振り返り

- K株式会社の最終的な目的は、「いつどの市場に、どの野菜を何トン出荷すれば利益が最大化できるかを予測するモデルの構築」ということでした。
- その話を聞いた株式会社Lのプロジェクトリーダーは、スコープ（1つの契約で実施する作業の範囲）をいくつかに切ることを提案しました。
  - スコープ1：市場における野菜の数量・価格を見える化する仕組みを構築する。
  - スコープ2：市場における野菜の価格を予測するモデルを構築する。
  - スコープ3：出荷先と出荷量を最適化するモデルを構築する。

- スコープ1と2について、ばれいしょ（北海道、長崎、鹿児島産）についてのデータ加工を実施してきました。

Pythonを使用したデータ可視化のテクニックと、可視化結果をうけての考察

市場ごと数量の確認

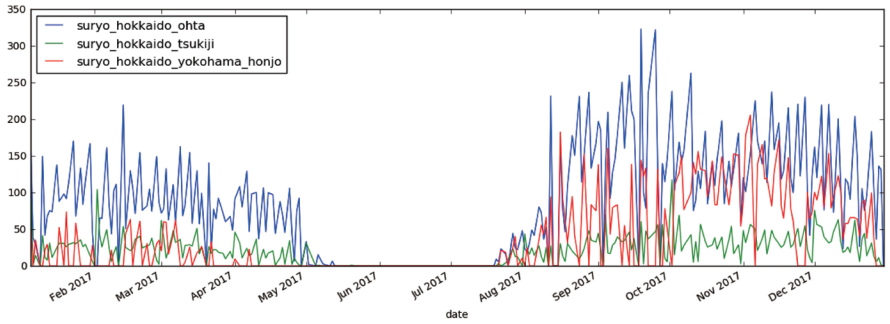
- 株式会社Lは北海道・長崎・鹿児島産ばれいしょの数量が季節でどのように変動するのか確認することにしました。
- まずは市場ごとに数量を表示してみることにしました。

演習1：市場ごと数量の確認

- ばれいしょの数量を可視化してください。

市場ごと数量の確認

- 株式会社Lは北海道産ばれいしょ数量を可視化してみましたが、先の上下が激しく、トレンドの考察などに不都合だと考えました。
- 移動平均線を表示すればスムーズな線になり、考察がしやすくなりそうです。

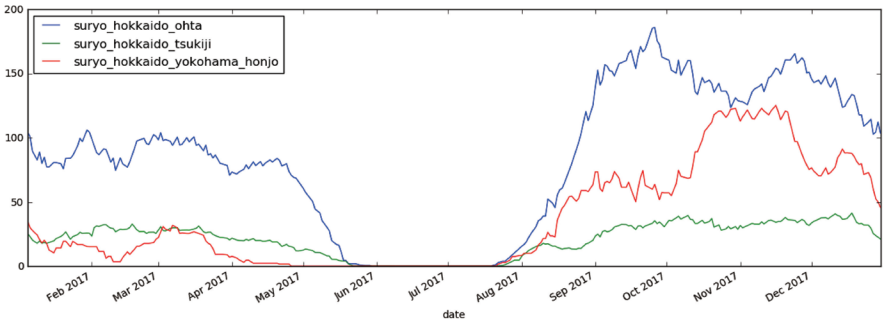


演習2：数量の移動平均線の確認

- ばれいしょ数量の移動平均線を可視化してください。

数量の移動平均線の確認

- 移動平均線を表示することにより、数量トレンドが見やすくなりました。株式会社Lは、今後の考察をする上で移動平均線を利用することに決定しました。
- 数量を可視化した後は、価格の移動平均線を可視化することにしました。

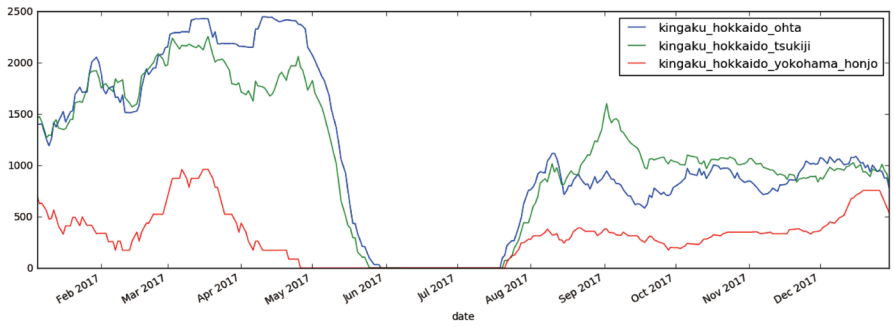


演習3：価格の移動平均線の確認

- ばれいしょ価格の移動平均線を可視化してください。

価格の移動平均線の確認

- 株式会社Lは、価格の移動平均線も可視化できました。
- 北海道産のばれいしょについて数量と価格の双方の可視化ができたため、数量と価格に関連性がないか考察することにしました。

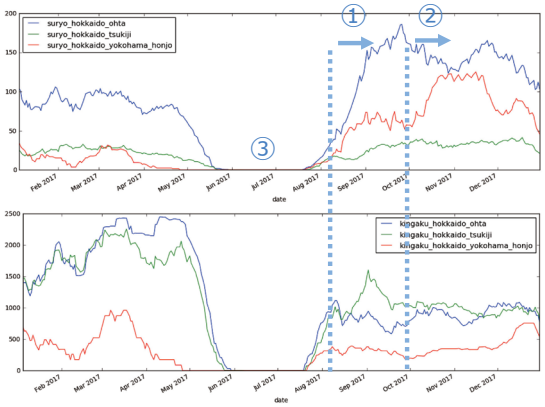


営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → **PoC** → システム設計 → システム構築

数量と価格の関連

- 株式会社Lは、まずは北海道産ばれいしょの数量・価格を比較してみました。
- ①のように、数量が多くなると市場にばれいしょがあふれるため、価格は低下することが確認できました。
- ②のように、数量が少なくなると市場のばれいしょが少なくなるため、価格が上昇することが確認できました。
- ③のように、そもそもシーズンではなくばれいしょの出荷がない時期は数量・価格ともにゼロになることも確認できました。

- 株式会社Lは個別の市場で数量・価格の関連を考察しましたが、全国平均ではどのような動きをするのかも考察することにしました。



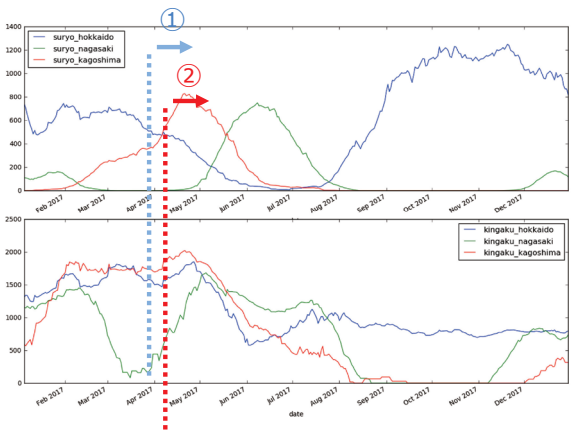
営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → **PoC** → システム設計 → システム構築

演習4：全国平均の数量と価格

- 全国平均のばれいしょ数量・価格の移動平均線を可視化してください。

全国平均の数量と価格の関連

- ①において北海道産の数量が少なくなると、北海道産の価格は上昇することが確認できました。
- ②においては鹿児島産の数量と価格は同時に上昇しています。これは価格が高いうちに利益をあげるために出荷量を増やしているのか、それとも需要に対して数量がまったく追いつかずに価格が上昇し続けているのか、のいずれかだと考察されます。
- これらの考察より、ばれいしょの価格変動は同じ産地の過去価格のみに依存するわけではなく、産地ごとの数量・価格の影響を受ける多変量自己回帰の挙動を示しているようです。



営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → **PoC** → システム設計 → システム構築

機械学習用データセットの作成

- ばれいしょ価格変動は多変量自己回帰の挙動を示しているように思われます。機械学習のデータセットにおける説明変数には産地ごとの数量・価格、また全国平均の数量・価格を横持ちにし、目的変数（将来価格）を説明するモデルの構築が必要になりそうです。

演習5：機械学習用データセットの作成

- 市場ごとデータと全国平均データを横持ちにしてください。

演習6：追加考察

- 数量・価格を市場や産地ごとにグラフ化し、数量・価格について皆さん独自の考察をしてください。

第8回：PoC

野菜数量と価格の予測モデル構築

アジェンダ

- 前回までの講義の振り返り
- LSTMによる野菜数量と価格の予測モデル構築

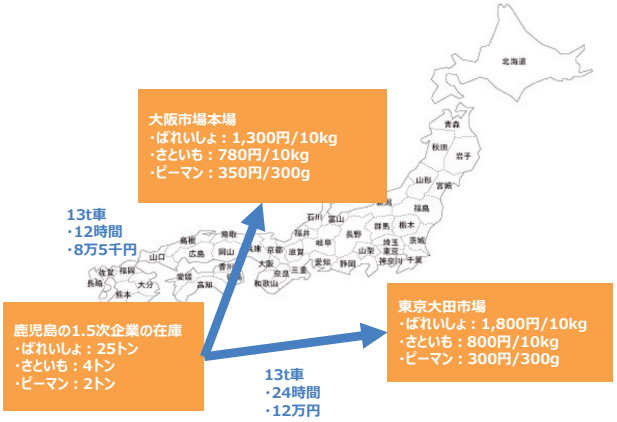
本講義で学習するテクニック

- 本講義ではLSTMによるモデリングを学習しますが、LSTMに投入するデータの加工（前処理）について多くの時間を割いています。
- データセットの結合、欠損値の補間、正規化、説明変数の追加、時系列データセットを構築する際の注意事項など、実社会におけるデータサイエンスの現場において頻繁に用いられるテクニックについて述べています。

前回までの講義の振り返り

前回までの振り返り

- 株式会社Lは、九州で産地から買い集めた野菜を関東・関西の中央卸売市場に転売する仲卸業を行っているK株式会社からの依頼を受け、市場における野菜の価格予測を行うシステムの構築を進めています。



営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → **PoC** → システム設計 → システム構築

前回までの振り返り

- K株式会社の最終的な目的は、「いつどの市場に、どの野菜を何トン出荷すれば利益が最大化できるかを予測するモデルの構築」ということでした。
- その話を聞いた株式会社Lのプロジェクトリーダーは、スコープ（1つの契約で実施する作業の範囲）をいくつかに切ることを提案しました。
  - スコープ1：市場における野菜の数量・価格を見える化する仕組みを構築する。
  - スコープ2：市場における野菜の価格を予測するモデルを構築する。
  - スコープ3：出荷先と出荷量を最適化するモデルを構築する。

- スコープ1と2について、ばれいしょ（北海道、長崎、鹿児島産）について数量と価格の関連性について考察を実施したところ、数量・価格は多変量自己回帰の挙動を示していることがわかりました。
- 機械学習によるモデルを構築するため、データセットの加工を実施してきました。

営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → **PoC** → システム設計 → システム構築

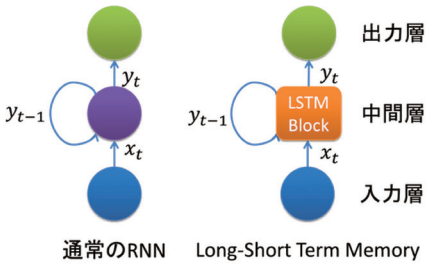
LSTMによる野菜数量と価格の予測モデル構築

数量・価格予測モデルのアルゴリズム

- 株式会社Lは、多変量自己回帰の挙動を示すばれいしょ数量・価格データから予測モデルを構築するために、LSTM（Long short-term memory）を採用することにしました。

LSTM (Long short-term memory)とは

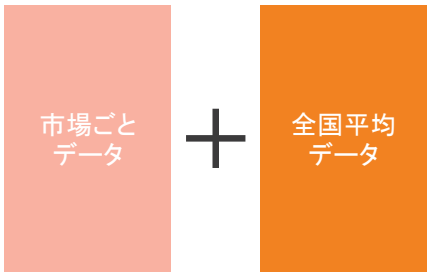
- RNN(Recurrent Neural Network)の拡張として1995年に登場した、時系列データ(sequential data)に対するモデル、あるいは構造(architecture)の1種です。その名は、Long term memory(長期記憶)とShort term memory(短期記憶)という神経科学における用語から取られています。LSTMはRNNの中間層のユニットをLSTM blockと呼ばれるメモリと3つのゲートを持つブロックに置き換えることで実現されています。



出展： https://qiita.com/t_Signull/items/21b82be280b46f467d1b

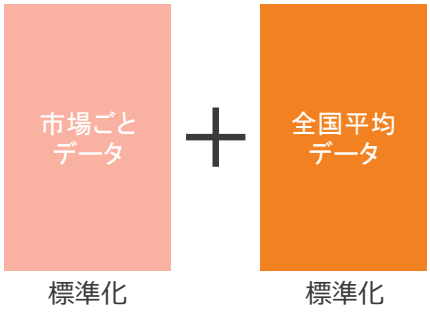
演習1：数量・価格データの準備

- 市場ごとのデータと全国平均のデータを横持ちにしてください。



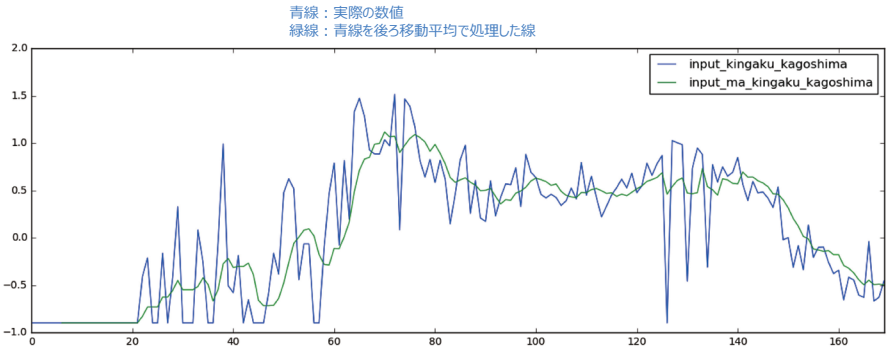
演習2：数量・価格データの標準化

- 結合した市場ごとデータと全国平均データに対して、標準化を実施してください。



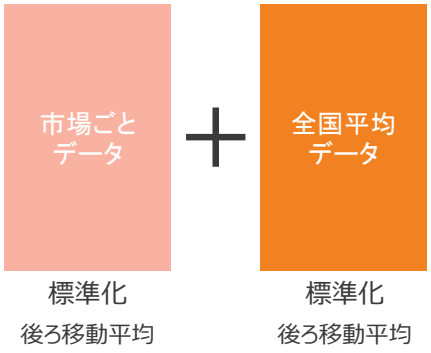
時系列データに対する移動平均処理の注意点

- 未来の数量・価格を予測するモデルを構築するので、移動平均処理が必要な場合には後ろ移動平均（ある時点の数値と、それ以前の数値のみを使用する）を選択する必要があります。
- 前移動平均や中央移動平均を使用すると、実際には知りえない未来のデータを説明変数に使用することになってしまい、正しいモデル構築、評価ができなくなってしまいます。



演習3：数量・価格データの移動平均

- 標準化後のデータに対して、後ろ移動平均を実施してください。



説明変数の追加

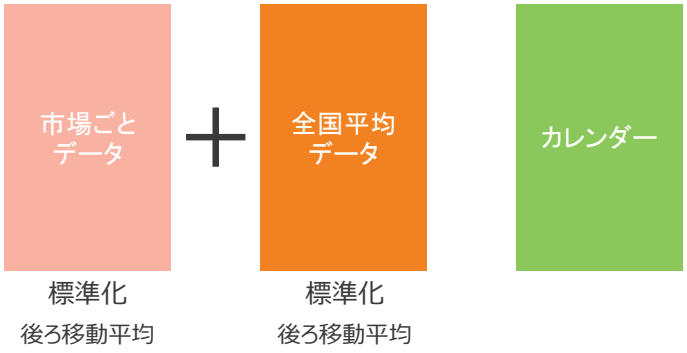
- 株式会社Lによる事前分析で、水曜日（隔週）と日曜日が休市日であることがわかっています。
- また、ばれいしょを始めとした野菜には旬があるので、月ごとに出荷総量が異なることもわかっています。

- 産地において収穫日の前に雨が降ると、掘った芋が泥で汚れてしまい値が下がるということもわかりました。出荷数日前の天気が、出荷量に影響を与えるであろうという仮説を立てました。
- また、出荷数週間前に気温が下がりすぎると芋のサイズが小ぶりになり値が下がるため、農家はわざと掘る時期を遅らせて芋を大きくするケースがあることもわかりました。

- 上記の情報より、数量・価格の予測モデルを構築するに当たりカレンダーと気象条件も説明変数として考慮したほうがモデル精度が向上すると思われるため、説明変数に追加することにしました。

演習4：カレンダーの追加

- カレンダーを読み込み、月・曜日のone hot vector部分のみを抽出してください。

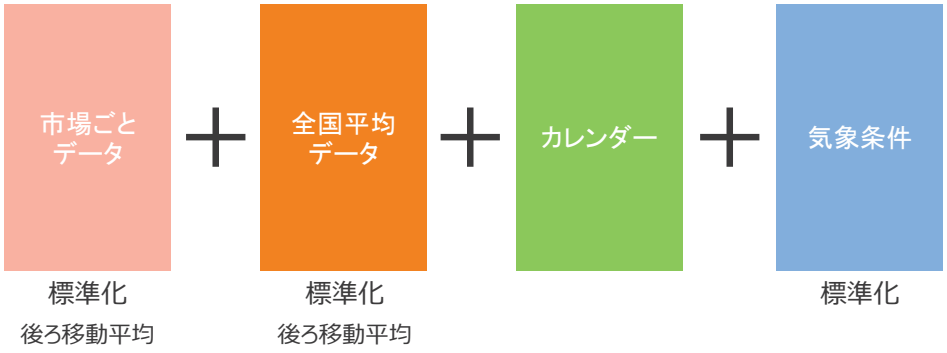


演習5：気象条件の追加

- 気象条件を読み込み、値を標準化してください。

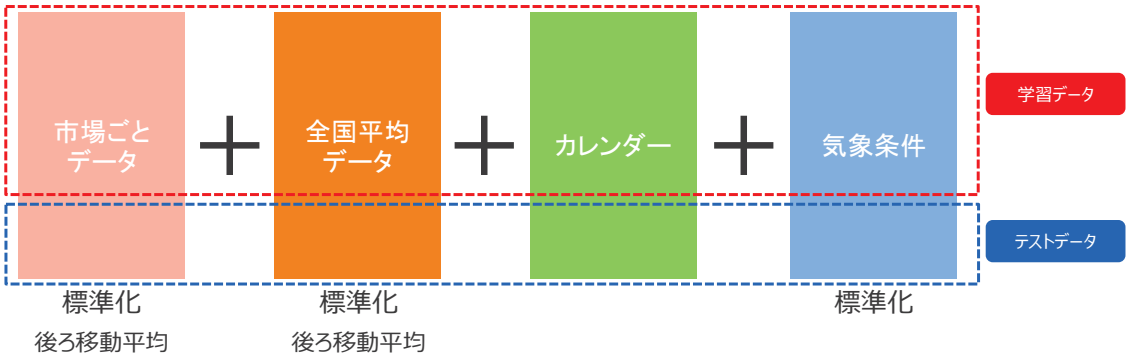
演習6：各種データの横持ち

- 気象条件を読み込み、値を標準化してください。



演習7：学習/テストデータの作成

- LSTM用の学習データ、テストデータを作成してください。



演習8：LSTMによるモデル構築

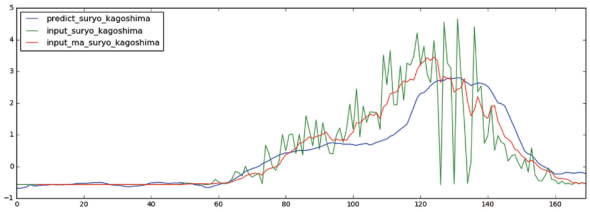
- LSTMモデルを作成してください。
- 作成したモデルを保存、読み込みできるように実装してください。

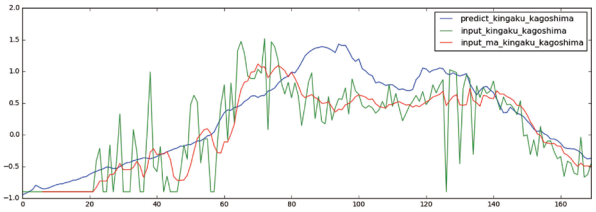
演習9：数量・価格の予測

- 予測を実行し、結果を描画してください。

（参考）モデルの予測結果をどのように活用するか

- モデルを構築して予測結果を得た後に、その予測結果を活用して利益を上げるビジネスの設計が重要となってきます。
- 例えば、数量と価格の予測値のみをK株式会社に提示したとして、K株式会社の購買担当者は野菜は買うか否か、判断に困ってしまうことが予想できます。
- 予測結果そのものを、担当者が行動を起こすための明確な指標にまで昇華させることができて初めて、人間をサポートするAIが構築できたと言えます。





第9回：PoC

モデルの高度化

アジェンダ

- 前回までの講義の振り返り
- LSTMによる野菜数量と価格の予測モデルのユースケース
- 数量・価格の長期予測モデル構築に向けて

本講義で学習すること

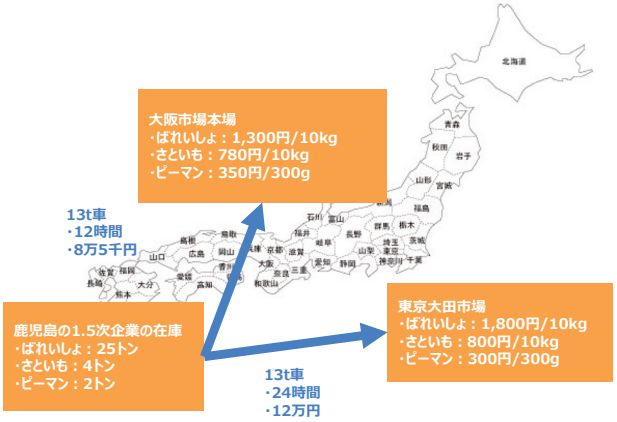
- PoCにおいて、AIモデルは目的を達成できるのか、どの程度達成できるのかということを明らかにしていきます。
- 例えば野菜数量・価格モデルを例にすると、
  - モデルは将来の値を予測できるのか？全くできないのか？
  - 予測できる場合、精度はどれくらいか？
  - 達成できる精度は、実用に耐えうるのか？
  - 活用できる場合、モデル単体で使用するのか？別のモデルや指標と組み合わせるのか？人による判断が必要なのか？

  ということを検討します。
- 本講義では、モデルのユースケース（モデルの活用事例）についての架空の議論を通して、モデル高度化にいたるプロセスを学習します。

前回までの講義の振り返り

前回までの振り返り

- 株式会社Lは、九州で産地から買い集めた野菜を関東・関西の中央卸売市場に転売する仲卸業を行っているK株式会社からの依頼を受け、市場における野菜の価格予測を行うシステムの構築を進めています。



営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → **PoC** → システム設計 → システム構築

前回までの振り返り

- K株式会社の最終的な目的は、「いつどの市場に、どの野菜を何トン出荷すれば利益が最大化できるかを予測するモデルの構築」ということでした。
- その話を聞いた株式会社Lのプロジェクトリーダーは、スコープ（1つの契約で実施する作業の範囲）をいくつかに切ることを提案しました。
  - スコープ1：市場における野菜の数量・価格を見える化する仕組みを構築する。
  - スコープ2：市場における野菜の価格を予測するモデルを構築する。
  - スコープ3：出荷先と出荷量を最適化するモデルを構築する。

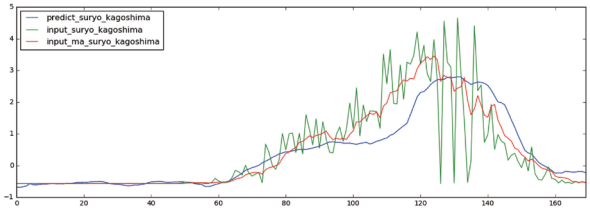
- スコープ1と2について、ばれいしょ（北海道、長崎、鹿児島産）について数量と価格の関連性について考察を実施したところ、数量・価格は多変量自己回帰の挙動を示していることがわかりました。
- 株式会社LはLSTMによる数量・価格予測モデルを構築し、モデルのユースケースについて具体的な検討をK株式会社とともに進めることにしました。

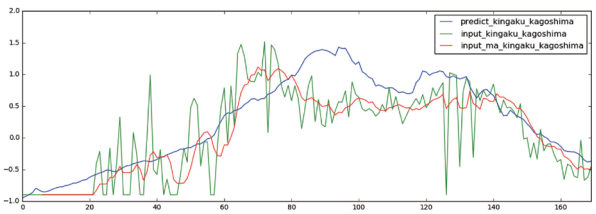
営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → **PoC** → システム設計 → システム構築

LSTMによる野菜数量と価格の予測モデルのユースケース

モデルの予測結果をどのように活用するか

- モデルを構築して予測結果を得た後に、その予測結果を活用して利益を上げるビジネスの設計が重要となってきます。
- 例えば、数量と価格の予測値のみをK株式会社に提示したとして、K株式会社の購買担当者は野菜は買うか否か、判断に困ってしまうことが予想できます。
- 予測結果そのものを、担当者が行動を起こすための明確な指標にまで昇華させることができて初めて、人間をサポートするAIが構築できたと言えます。



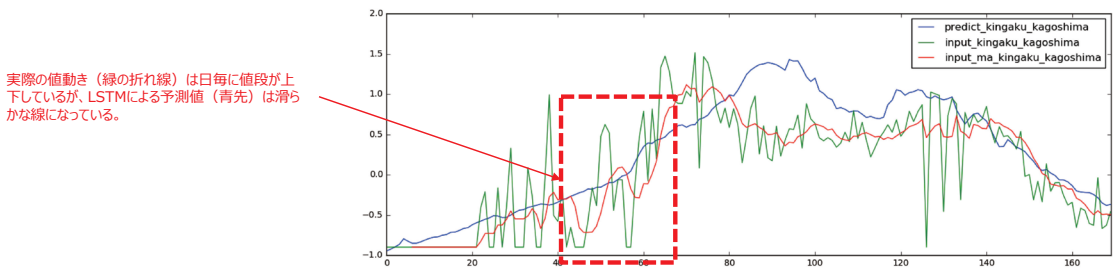


モデルの予測結果をどのように活用するか

- 株式会社Lのデータサイエンティストは、K株式会社の購買担当者とモデルの予測結果について議論しています。
- K株式会社の購買担当者は、当初このモデルを短期の投機目的で使用することを検討していました。例えば、「2日後の鹿児島産ばれいしょの値段が今日より1割上昇すると予測されているので、今日は市場に出さずに2日後に市場に出そう。」という判断に使うことを検討していたとのことです。

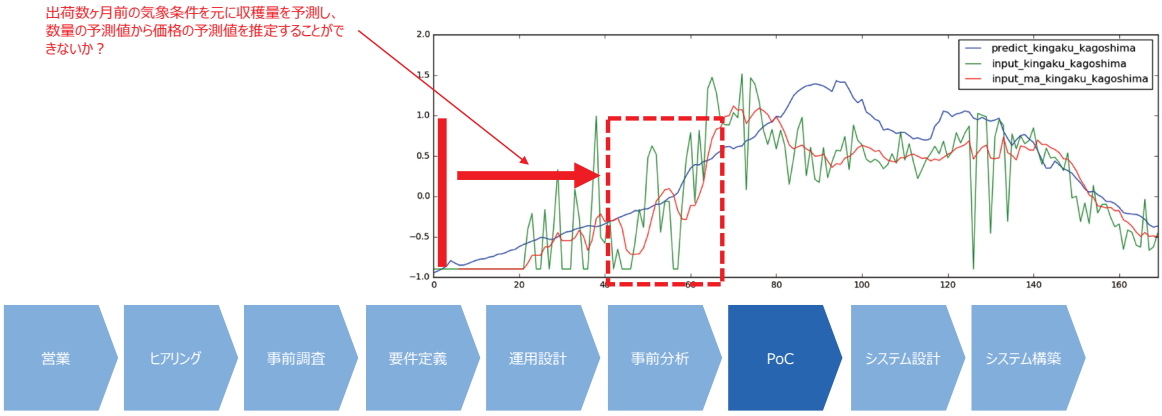
モデルの予測結果をどのように活用するか

- モデルの予測結果を観察してみると、実際の日毎の値動きは表現できていないが、前後数日の平均値は表現できていることがわかりました。
- 「数日先の出荷を想定したモデルの活用は困難である。」いう判断がされました。



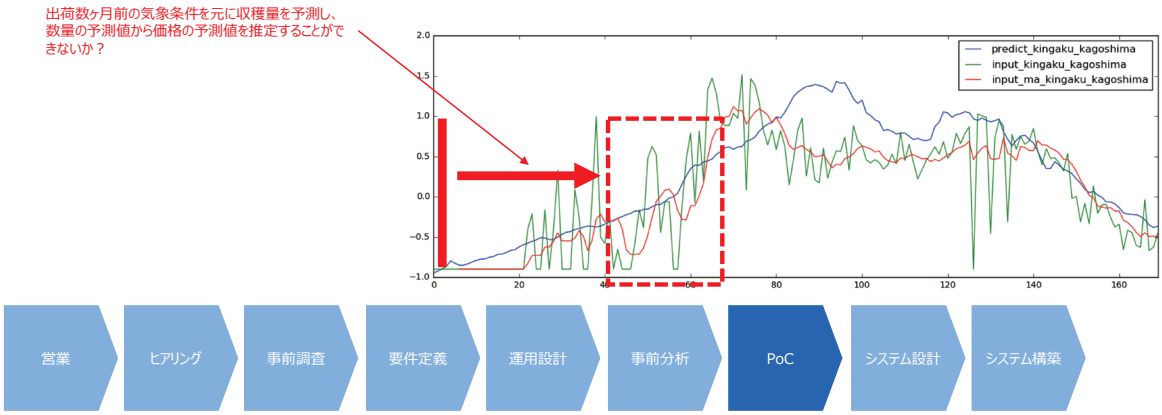
モデルの予測結果をどのように活用するか

- 「数日先の出荷を想定したモデルの活用は困難である。」いう判断がされましたが、「仮に数ヶ月前に価格が予測できれば、早期に農家と契約を交わし、値上がりする時期にばれいしょを大量に集めることはできる。数日単位の利益でなく、数週間平均の利益で評価すればいいのではないか。」という別のユースケースについて検討することになりました。



モデルの予測結果をどのように活用するか

- 数ヶ月後の数量を気象条件から予測することができれば、新しいユースケースが実現できる可能性があります。
- 株式会社Lは、気象条件から数量の予測が可能なのか検討することにしました。



演習1：ばれいしょデータの準備

- ばれいしょ全国のデータを年月で集計してください。

演習2：気象データの準備

- ばれいしょ産地の気象データを読み込み、年月で集計してください。

演習3：鹿児島の気象データの準備

- 3大産地の気象条件データから、鹿児島の気象データのみを抽出してください。

演習4：鹿児島の気象データの準備

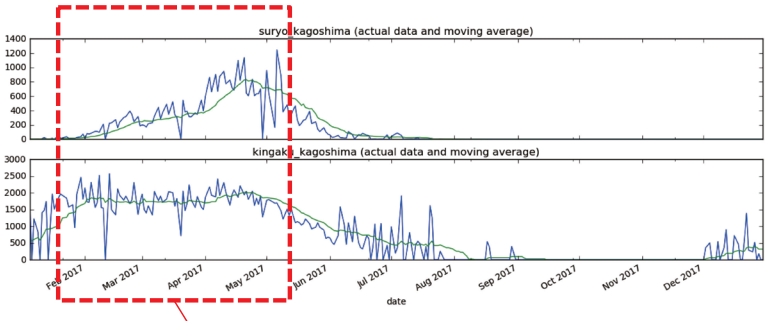
- 鹿児島内の気象条件の平均値を算出してください。

演習5：ばれいしょ数量と気象条件の相関

- 鹿児島産ばれいしょの数量と鹿児島の気象データの相関を観察してください。

2月～5月の数量と気象条件

- 鹿児島産ばれいしょの出荷時期全体において、数量と気象条件には相関があることがわかりました。
- 株式会社Lは、出荷時期の前半と後半で傾向に違いがあるのかを更に観察することにしました。

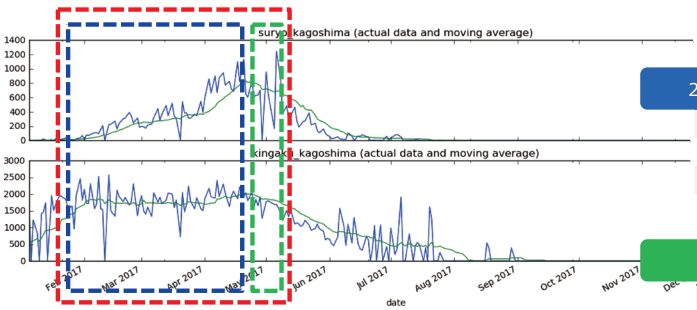


	suryo_kagoshima	kagoshima_kion	kagoshima_kousui	kagoshima_nissyo
suryo_kagoshima	1.000000	-0.602390	-0.415203	-0.567579
kagoshima_kion	-0.602390	1.000000	0.801528	0.766864
kagoshima_kousui	-0.415203	0.801528	1.000000	0.495284
kagoshima_nissyo	-0.567579	0.766864	0.495284	1.000000

演習6：出荷時期ごとの相関

- 相関係数を算出するデータ期間を変更して相関係数を算出してください。

出荷時期ごとの相関



2～5月

	suryo_kagoshima	kagoshima_kion	kagoshima_kousui	kagoshima_nissyo
suryo_kagoshima	1.000000	-0.602390	-0.415203	-0.567579
kagoshima_kion	-0.602390	1.000000	0.801528	0.766864
kagoshima_kousui	-0.415203	0.801528	1.000000	0.495284
kagoshima_nissyo	-0.567579	0.766864	0.495284	1.000000

2～4月

	suryo_kagoshima	kagoshima_kion	kagoshima_kousui	kagoshima_nissyo
suryo_kagoshima	1.000000	-0.937026	-0.686870	-0.683169
kagoshima_kion	-0.937026	1.000000	0.757176	0.791791
kagoshima_kousui	-0.686870	0.757176	1.000000	0.509180
kagoshima_nissyo	-0.683169	0.791791	0.509180	1.000000

5月

	suryo_kagoshima	kagoshima_kion	kagoshima_kousui	kagoshima_nissyo
suryo_kagoshima	1.000000	0.137484	0.270102	-0.551330
kagoshima_kion	0.137484	1.000000	0.790900	-0.273683
kagoshima_kousui	0.270102	0.790900	1.000000	-0.396731
kagoshima_nissyo	-0.551330	-0.273683	-0.396731	1.000000

質問：出荷時期ごとの相関係数の違い

- 2～4月において数量と気象条件は逆相関の傾向が強く、5月は弱い順相関の傾向が見られました。どうしてこのような傾向の違いが出ているのか、その理由を皆さんで議論してみてください。

第10回：データの可視化

低コストでデータ可視化の仕組みを構築する

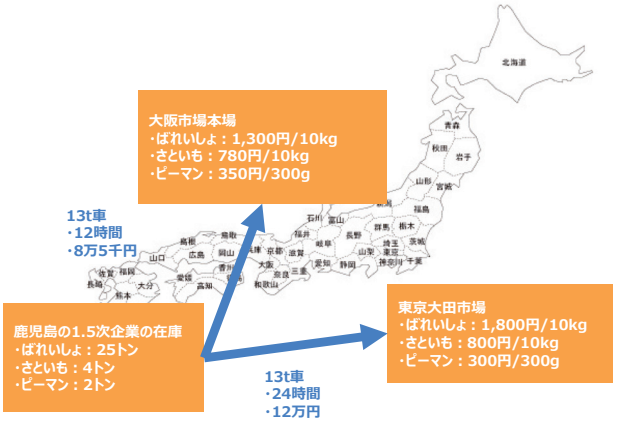
アジェンダ

- 前回までの講義の振り返り
- Jupyter notebookとipywidgetsを活用したデータ可視化機能の実装

前回までの講義の振り返り

前回までの振り返り

- 株式会社Lは、九州で産地から買い集めた野菜を関東・関西の中央卸売市場に転売する仲卸業を行っているK株式会社からの依頼を受け、市場における野菜の価格予測を行うシステムの構築を進めています。



営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → PoC → システム設計 → システム構築

前回までの振り返り

- K株式会社の最終的な目的は、「いつどの市場に、どの野菜を何トン出荷すれば利益が最大化できるかを予測するモデルの構築」ということでした。
- その話を聞いた株式会社Lのプロジェクトリーダーは、スコープ（1つの契約で実施する作業の範囲）をいくつかに切ることを提案しました。
  - スコープ1：市場における野菜の数量・価格を見える化する仕組みを構築する。
  - スコープ2：市場における野菜の価格を予測するモデルを構築する。
  - スコープ3：出荷先と出荷量を最適化するモデルを構築する。

- LSTMによる数量・価格予測モデルを構築した後、数カ月先の長期予測モデルを見据えた分析（気象データと数量の相関）を実施してきました。

Jupyter notebookとipywidgetsを活用したデータ可視化機能の実装

可視化がなぜ大事なのか

- 定量化のために相関係数などの指標を参考にすることがありますが、指標化することで失われてしまう情報があります。
- 例えば時系列データでは確認できるデータ列の傾き、散布図では確認できるデータ点数などが、指標化することで失われてしまいます。

2～5月

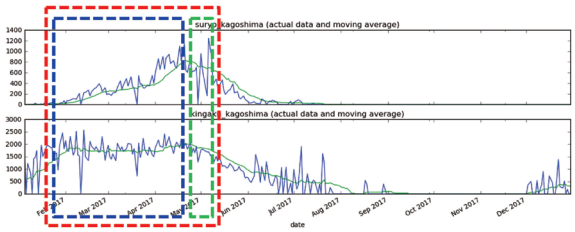
	suryo_kagoshima	kagoshima_kion	kagoshima_kousui	kagoshima_nissyo
suryo_kagoshima	1.000000	-0.602390	-0.415203	-0.567579
kagoshima_kion	-0.602390	1.000000	0.801528	0.766864
kagoshima_kousui	-0.415203	0.801528	1.000000	0.495284
kagoshima_nissyo	-0.567579	0.766864	0.495284	1.000000

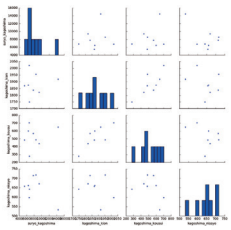
2～4月

	suryo_kagoshima	kagoshima_kion	kagoshima_kousui	kagoshima_nissyo
suryo_kagoshima	1.000000	-0.937026	-0.686870	-0.683169
kagoshima_kion	-0.937026	1.000000	0.757176	0.791791
kagoshima_kousui	-0.686870	0.757176	1.000000	0.509180
kagoshima_nissyo	-0.683169	0.791791	0.509180	1.000000

5月

	suryo_kagoshima	kagoshima_kion	kagoshima_kousui	kagoshima_nissyo
suryo_kagoshima	1.000000	0.137484	0.270102	-0.551330
kagoshima_kion	0.137484	1.000000	0.790900	-0.273683
kagoshima_kousui	0.270102	0.790900	1.000000	-0.396731
kagoshima_nissyo	-0.551330	-0.273683	-0.396731	1.000000





可視化のコストを小さくする

- Jupyter notebook、matplotlib、ipywidgetsなどの可視化に便利なツール、ライブラリを使用することで、可視化に必要なコストを大きく下げることができます。
- 本講義では、これらのツール、ライブラリを駆使してデータを可視化する手法を習得します。

演習1：可視化するデータの準備

- 各市場ごとのデータと全国のデータを横持ちにしてください。

演習2：可視化するデータの準備

- 移動平均を計算してください。

演習3：対話的な可視化

- 数量と価格を市場・産地ごとに表示する機能を実装してください。

演習4：対話的な可視化

- 産地ごとの数量と気象条件を表示する機能を実装してください。

第11回：AIモデルの運用

モデルの監視・高度化・人間との協業

アジェンダ

- ケーススタディI（農作物市況予測）の振り返り
- AIモデル運用時に検討が必要なこと

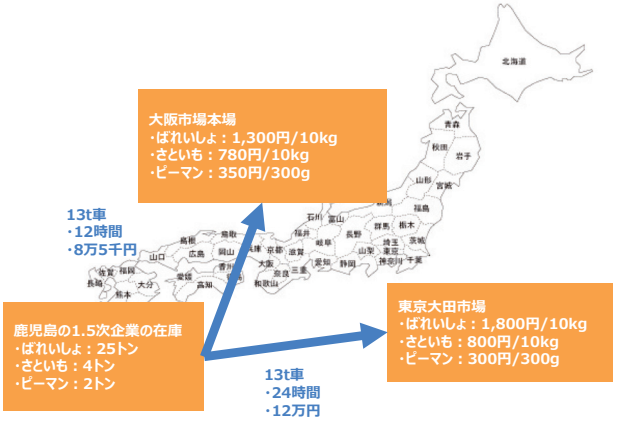
ケーススタディⅠ（農作物市況予測）の振り返り

仕事の依頼

- 株式会社Lは金融工学が得意なベンチャー企業で、数学や工学が得意なメンバーが在籍しています。株式会社Lは、機械学習を金融分野に適用する仕事の実績があります。
- そんな株式会社Lに、農業分野の仕事の依頼がありました。中央卸売市場における"ばれいしょ（じゃがいも）"の値動きを可視化できるツールが欲しいという依頼です。
- 野菜市場についての知見が無い株式会社Lは、ひとまず依頼先の会社を訪問して詳しい話を聞くことにしました。

ヒアリング

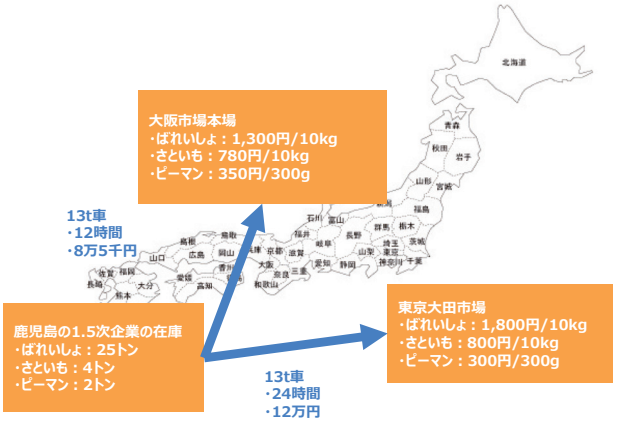
- 仕事を依頼してきたのは大手商社の子会社であるK株式会社でした。K株式会社は九州に拠点があり、九州で産地から買い集めた野菜を関東・関西の中央卸売市場に転売する仲卸業を行っているそうです。
- 例えば春先であれば、その時期に取れるばれいしょ・さといも・ピーマンを中央卸売市場に出荷します。それぞれの市場で買い取り価格が異なりますし、輸送コスト（時間と金額）も異なるそうです。



営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → PoC → システム設計 → システム構築

ヒアリング

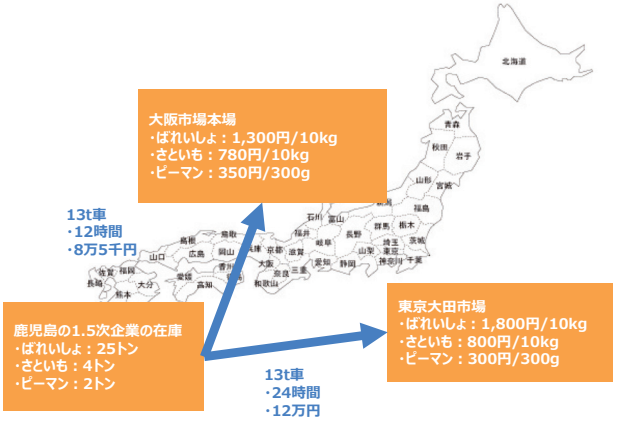
- K株式会社の最終的な目的は、「いつどの市場に、どの野菜を何トン出荷すれば利益が最大化できるかを予測するモデルの構築」ということでした。
- その話を聞いた株式会社Lのプロジェクトリーダーは、スコープ（1つの契約で実施する作業の範囲）をいくつかに切ることを提案しました。
  - スコープ1：市場における野菜の数量・価格を見える化する仕組みを構築する。
  - スコープ2：市場における野菜の価格を予測するモデルを構築する。
  - スコープ3：出荷先と出荷量を最適化するモデルを構築する。



営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → PoC → システム設計 → システム構築

ヒアリング

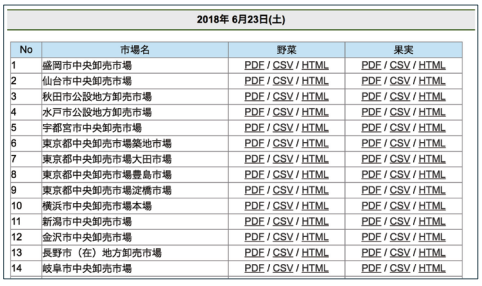
- K株式会社と株式会社Lの話し合いにより、まずはスコープ1とスコープ2について実施することとなりました。



営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → PoC → システム設計 → システム構築

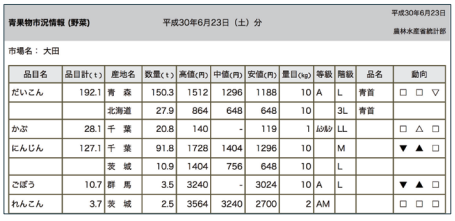
事前調査

- 株式会社Lは、最初のスコープである「市場における野菜の数量・価格を見える化する仕組みを構築する。」の実現のため、データの調査を開始しました。
- 農水省HPに中央卸売市場のデータが公開されていることを知った株式会社Lは、データの調査を開始することにしました。



2018年 6月23日(土)

No	市場名	野菜	果実
1	盛岡市中央卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
2	仙台市中央卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
3	秋田市公設地方卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
4	水戸市公設地方卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
5	宇都宮市中央卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
6	東京都中央卸売市場築地市場	PDF / CSV / HTML	PDF / CSV / HTML
7	東京都中央卸売市場大田市場	PDF / CSV / HTML	PDF / CSV / HTML
8	東京都中央卸売市場豊島市場	PDF / CSV / HTML	PDF / CSV / HTML
9	東京都中央卸売市場淀橋市場	PDF / CSV / HTML	PDF / CSV / HTML
10	横浜市中央卸売市場本場	PDF / CSV / HTML	PDF / CSV / HTML
11	新潟市中央卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
12	金沢市中央卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
13	長野市（在）地方卸売市場	PDF / CSV / HTML	PDF / CSV / HTML
14	岐阜市中央卸売市場	PDF / CSV / HTML	PDF / CSV / HTML



青果物市況情報（野菜）　平成30年6月23日（土）分　平成30年6月23日　農林水産省統計部

市場名：大田

品目名	品目計(t)	産地名	数量(t)	高値(円)	中値(円)	安値(円)	量目(kg)	等級	階級	品名	動向
だいこん	192.1	青　森	150.3	1512	1296	1188	10	A	L	青首	□ □ ▽
		北海道	27.9	864	648	648	10		3L	青首	
かぶ	28.1	千　葉	20.8	140	-	119	1	ﾊﾞﾗ	LL		□ △ □
にんじん	127.1	千　葉	91.8	1728	1404	1296	10		M		▼ ▲ □
		茨　城	10.9	1404	756	648	10		L		
ごぼう	10.7	群　馬	3.5	3240	-	3024	10	A	L		▼ ▲ □
れんこん	3.7	茨　城	2.5	3564	3240	2700	2	AM			□ □ □

出展：農林水産省HPより

営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → PoC → システム設計 → システム構築

数量・価格の定義

- 数量の指標として「入荷量」を適用することにしました。ただし、入荷量が記載れていないレコードは、その商品の入荷が少なかったことを表すことが判明したため、無視することにしました。
- 金額の指標として使用できそうな項目は高値・中値・安値の3項目存在します。調査の結果、中値が仕入量が最も多い製品の金額であることがわかったのですが、高値と安値を公開することが市場に与えるインパクトは無視できないということから、3項目のうち存在する数値の平均値を価格として採用することにしました。

	年	月	日	曜日	市場名	市場コード	品目名	品目コード	産地名	産地コード	品目計	入荷量	高値	中値	安値	等級	階級	品名	量目	動向
78	2018	1	15	月	築地	13300	ばれいしょ	36200	北海道	1.0	58.1	50.4	1080	-	972	シユウ	L	メクイン	10.0	弱保合
79	2018	1	15	月	築地	13300	ばれいしょ	36200	北海道	1.0	NaN	NaN	1404	1296	1188	シユウ	L	男爵	10.0	NaN
80	2018	1	15	月	築地	13300	ばれいしょ	36200	長　崎	42.0	NaN	7.6	-	864	-	シユウ	S	NaN	10.0	NaN
81	2018	1	15	月	築地	13300	ばれいしょ	36200	長　崎	42.0	NaN	NaN	756	-	648	シユウ	3S	NaN	10.0	NaN

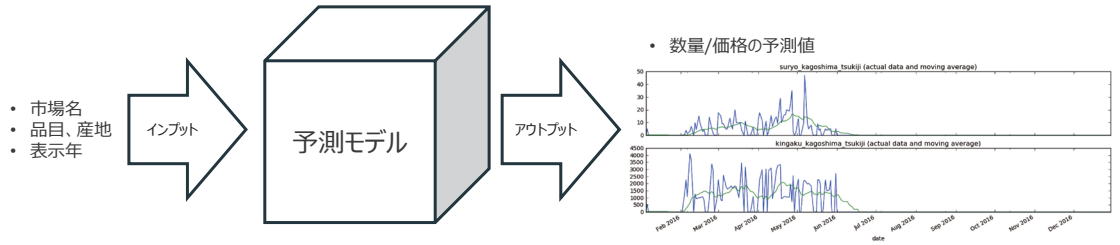
入荷量の記載がないレコードが存在する。

金額に関する数値は高値・中値・安値の3つある。

営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → PoC → システム設計 → システム構築

システムのインプットとアウトプット

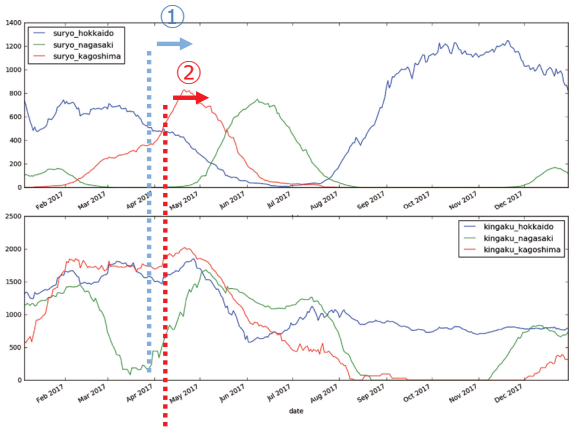
- 「スコープ1：市場における野菜の数量・価格を見える化する仕組み」におけるアウトプットの定義は前ページのように決定しました。
- 「スコープ2：市場における野菜の価格を予測するモデル」のインプットはスコープ1と同じにし、アウトプットは、スコープ1の金額に対する予測値とすることが決まりました。



営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → PoC → システム設計 → システム構築

ばれいしょ全国平均の数量と価格の関連

- ①において北海道産の数量が少なくなると、北海道産の価格は上昇することが確認できました。
- ②においては鹿児島産の数量と価格は同時に上昇しています。これは価格が高いうちに利益をあげるために出荷量を増やしているのか、それとも需要に対して数量がまったく追いつかずに価格が上昇し続けているのか、のいずれかだと考察されます。
- これらの考察より、ばれいしょの価格変動は同じ産地の過去価格のみに依存するわけではなく、産地ごとの数量・価格の影響を受ける多変量自己回帰の挙動を示しているようです。

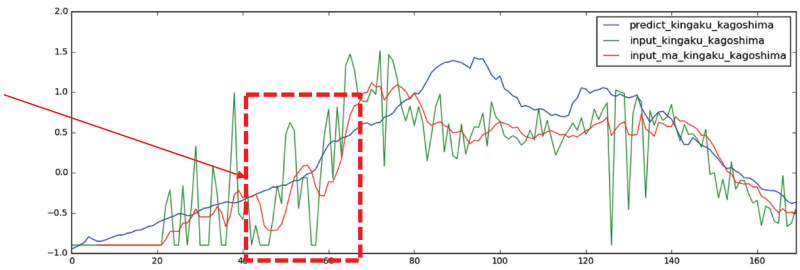


営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → **PoC** → システム設計 → システム構築

モデルの予測結果をどのように活用するか

- モデリングにはLSTMを用いることにしました。
- モデルの予測結果を観察してみると、実際の日毎の値動きは表現できていないが、前後数日の平均値は表現できていることがわかりました。
- 「数日先の出荷を想定したモデルの活用は困難である。」という判断がされました。

実際の値動き（緑の折れ線）は日毎に値段が上下しているが、LSTMによる予測値（青先）は滑らかな線になっている。

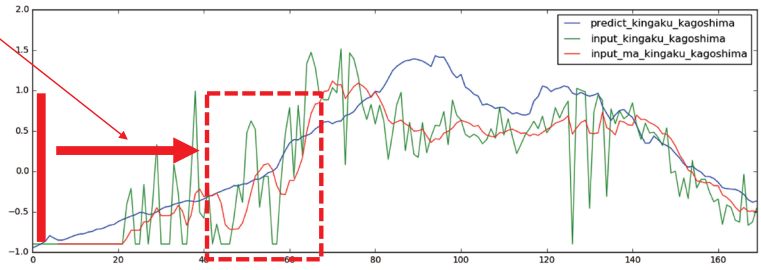


営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → **PoC** → システム設計 → システム構築

モデルの予測結果をどのように活用するか

- 「数日先の出荷を想定したモデルの活用は困難である。」いう判断がされましたが、「仮に数ヶ月前に価格が予測できれば、早期に農家と契約を交わし、値上がりする時期にばれいしょを大量に集めることはできる。数日単位の利益でなく、数週間平均の利益で評価すればいいのではないか。」という別のユースケースについて検討することになりました。

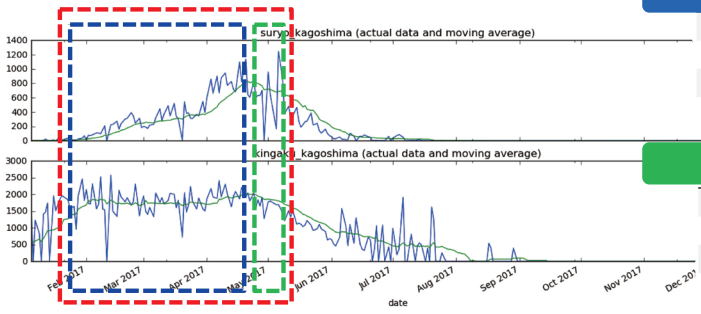
出荷数ヶ月前の気象条件を元に収穫量を予測し、数量の予測値から価格の予測値を推定することができないか？



営業 → ヒアリング → 事前調査 → 要件定義 → 運用設計 → 事前分析 → **PoC** → システム設計 → システム構築

出荷時期ごとの相関

- 数量価格の長期予測をするために、気象条件と数量の相関を分析することになりました。
- 気象条件と数量には一定の相関が見られ、モデル高度化の可能性が期待できる結果となりました。



2〜5月

	suryo_kagoshima	kagoshima_kion	kagoshima_kousui	kagoshima_nissyo
suryo_kagoshima	1.000000	-0.602390	-0.415203	-0.567579
kagoshima_kion	-0.602390	1.000000	0.801528	0.766864
kagoshima_kousui	-0.415203	0.801528	1.000000	0.495284
kagoshima_nissyo	-0.567579	0.766864	0.495284	1.000000

2〜4月

	suryo_kagoshima	kagoshima_kion	kagoshima_kousui	kagoshima_nissyo
suryo_kagoshima	1.000000	-0.937026	-0.686870	-0.683169
kagoshima_kion	-0.937026	1.000000	0.757176	0.791791
kagoshima_kousui	-0.686870	0.757176	1.000000	0.509180
kagoshima_nissyo	-0.683169	0.791791	0.509180	1.000000

5月

	suryo_kagoshima	kagoshima_kion	kagoshima_kousui	kagoshima_nissyo
suryo_kagoshima	1.000000	0.137484	0.270102	-0.551330
kagoshima_kion	0.137484	1.000000	0.790900	-0.273683
kagoshima_kousui	0.270102	0.790900	1.000000	-0.396731
kagoshima_nissyo	-0.551330	-0.273683	-0.396731	1.000000

可視化がなぜ大事なのか

- モデル高度化の一環として、データの可視化について学習しました。
- 例えば時系列データでは確認できるデータ列の傾き、散布図では確認できるデータ点数などが、相関係数などの指標とするで失われてしまいます。データ可視化はそれを回避する有効な手法です。

2～5月

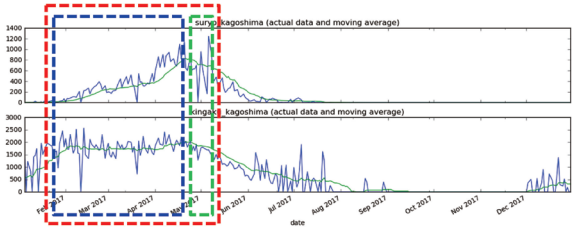
	suryo_kagoshima	kagoshima_kion	kagoshima_kousui	kagoshima_nissyo
suryo_kagoshima	1.000000	-0.602390	-0.415203	-0.567579
kagoshima_kion	-0.602390	1.000000	0.801528	0.766864
kagoshima_kousui	-0.415203	0.801528	1.000000	0.495284
kagoshima_nissyo	-0.567579	0.766864	0.495284	1.000000

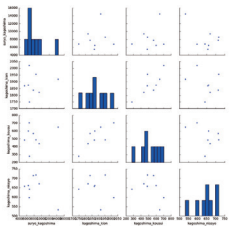
2～4月

	suryo_kagoshima	kagoshima_kion	kagoshima_kousui	kagoshima_nissyo
suryo_kagoshima	1.000000	-0.937026	-0.686870	-0.683169
kagoshima_kion	-0.937026	1.000000	0.757176	0.791791
kagoshima_kousui	-0.686870	0.757176	1.000000	0.509180
kagoshima_nissyo	-0.683169	0.791791	0.509180	1.000000

5月

	suryo_kagoshima	kagoshima_kion	kagoshima_kousui	kagoshima_nissyo
suryo_kagoshima	1.000000	0.137484	0.270102	-0.551330
kagoshima_kion	0.137484	1.000000	0.790900	-0.273683
kagoshima_kousui	0.270102	0.790900	1.000000	-0.396731
kagoshima_nissyo	-0.551330	-0.273683	-0.396731	1.000000





AIモデル運用時に検討が必要なこと

運用設計（第2回講義より）

- AIモデルを実サービスとして展開し続けるためには、運用が必要になります。
- 増え続けるデータの蓄積、モデルの更新、不具合への対応にどれだけの人・機材・お金・時間を投入するのかを決定します。投入するこれらのコストを勘案し、ユーザに請求するサービス料の設定が必要となります。

- ユーザの予算が決まっていて多額の投資をできない場合、システム機能の削減や運用フローの簡素化を実施します。

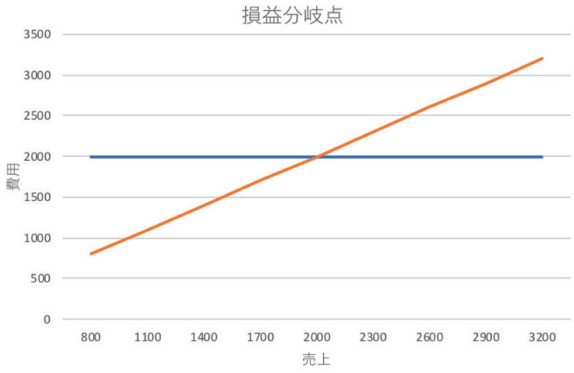
運用の重要性

- 本講義では、進行の都合上、「運用設計」の前に「事前分析」や「PoC」を学習しました。
- AIを使用して実現したいサービスが明確な場合、「運用設計」を優先して実施することが多くあります。仮にPoCで素晴らしいモデルが構築できたとしても、実際のサービスとしてモデルを運用できないケースや、運用コストが高すぎてサービスを維持できなくケースを回避するためです。

- これらの判断はデータサイエンティストだけで下せるわけではなく、企画・営業・システム構築を実施する部署とも連携して進められることが一般的です。
- データサイエンティストはデータやプログラムと向き合うだけでなく、チームメンバーとも向き合う必要があるという意味で、高いコミュニケーション力が求められます。

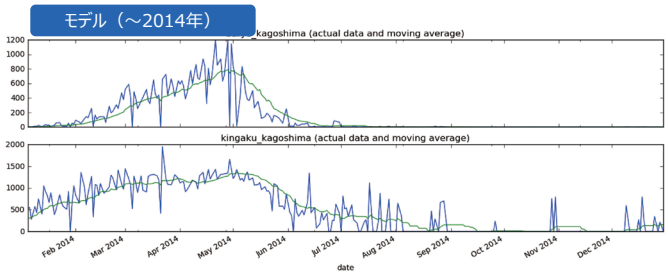
売上と費用（コスト）のバランス

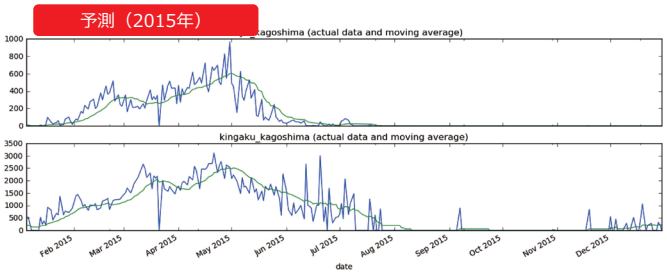
- システムを作るコスト、システムを維持するコストよりも売上が上回らなければ、ビジネスとしては成り立ちません。
- 例えばAIモデルを導入することで顧客に提供する価値が向上し、売上（オレンジの線）の傾きが右肩上がりになるとします。AIシステム構築や運用に掛かるコスト（青の線）を、どこかで売上線が追い越します（損益分岐点）。
- 損益分岐までの時間は適切なのか、そもそも損益分岐を迎えることができるのか（黒字になるのか）、という視点でサービスや運用を設計します。



モデルの適用範囲

- 運用においてAIモデルの適用範囲を抑えておくことと、適用範囲を超えた際に人間が介入する仕組みを考えておくことも重要です。
- 例えば、下記のように2014年までのばれいしょデータを使ってモデルを作成したとします。モデルが学習するデータには価格が1,500円程度までのデータしか存在しない場合、2015年のように価格が2,500円を超えるような状況においては、モデルの予測結果は予期しない値となる場合があります。
- 「モデルが学習したデータの範囲を超える状況になった場合にはアラートを上げ、予測結果を人間が判断する」などの対策を講じます。





第12回：独自コーパスの作成

アジェンダ

- 対話分類モデル構築プロジェクト（ケーススタディII）の設定について
- 独自コーパスの作成

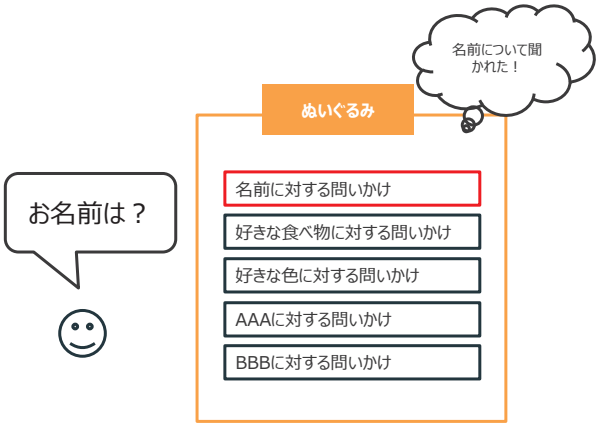
対話分類モデル構築プロジェクト（ケーススタディⅡ）の設定について

仕事の依頼

- 株式会社Lは金融工学が得意なベンチャー企業で、数学や工学が得意なメンバーが在籍しています。株式会社Lは、機械学習を金融分野に適用する仕事の実績があります。
- そんな株式会社Lに、玩具メーカーから自然言語処理分野の仕事の依頼がありました。玩具メーカーの人気商品に、人間と話す新機能を追加したいという依頼です。
- 「人間と話す」という先方の要望がどの程度の難易度の話なのか、ひとまず依頼先の会社を訪問して詳しい話を聞くことにしました。

ヒアリング

- 仕事を依頼してきたのは玩具メーカの株式会社Bでした。株式会社Bは、子供からの問いかけに対して答えることができるぬいぐるみを作りたいとのことです。
- 例えば「お名前は？」という子供の問いかけに対して、「私の名前はXXXだよ。」というように、簡単な会話が楽しめるようにしたいとのことです。
- 名前を聞かれる/好きな色を聞かれる/好きな食べ物を聞かれるなど、いくつかの問いかけに対して、まずは質問の分類ができるモデルを作成したいとのことでした。



独自コーパスの作成

コーパスの事前調査

- 株式会社L、株式会社Bともに、名前を聞かれる/好きな色を聞かれる/好きな食べ物を聞かれる、などを想定したコーパスは持っていませんでした。またライセンスの制約がないコーパスの調査をしましたが、適切なものが見つかりませんでした。
- そこで株式会社Lは、独自にコーパスを作成することにしました。

モデルの要件定義

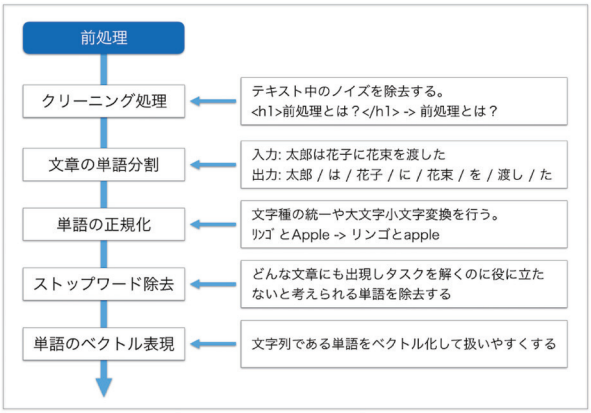
- 株式会社Bは音声をテキストに変換するプロダクトを持っていたため、子供の発話をテキストに変換する機能はそのプロダクトを使用することにしました。したがってモデルのインプットは「テキスト」と決定しました。
- 分類する発話内容は下記の5つと決まりました。
  1. 名前を聞かれる
  2. 好きな色を聞かれる
  3. 好きな食べ物を聞かれる
  4. 年齢を聞かれる
  5. 挨拶される

演習1：独自データ収集

- 下記の問いかけの発話事例を記載してください。
  class_0: 名前を聞かれる
  class_1: 好きな色を聞かれる
  class_2: 好きな食べ物を聞かれる
  class_3: 年齢を聞かれる
  class_4: 挨拶される

自然言語処理における前処理

- コンピュータは自然言語そのものを理解できないため、コンピュータが理解できるデータに自然言語を変換する必要があります。
- 自然言語処理においてモデルを作成するときは、単語や文章をベクトル（多次元の要素を持つ量）に変換し、コンピュータで処理します。



出展： https://qiita.com/Hironsan/items/2466fe0f344115aff177

文章のベクトル化：Bag of Words

- ベクトル表現の一種で、文章に単語が含まれるかどうかのみを考え、単語の並び方などは考慮しない形式のことです。

Step1：解析対象の文章群を準備します。

```
['天気 を 教え て ください 。',
 '明日 の 天気 は どう です か ?',
 '今日 の 天気 を 教え て よ 。',
 '新宿 の 天気 は どう なっ て いる ?',
 '横浜 の 明日 の 天気 は どう か の かな ?',
 '気温 を 教え て よー 。',
 '明日 の 気温 は どう な の ?',
 '今日 の 気温 は 低い ね !',
 '横浜 の 気温 は ?',
 '新宿 の 昨日 の 気温 を 教え て ください 。']
```

Step2：重複しない形態素のリストを作成し、次元を決定します。

```
{'いる': 0,
 'かな': 1,
 'ください': 2,
 'です': 3,
 'どう': 4,
 'なっ': 5,
 'よー': 6,
 '今日': 7,
 '低い': 8,
 '天気': 9,
 '教え': 10,
 '新宿': 11,
 '明日': 12,
 '昨日': 13,
 '横浜': 14,
 '気温': 15}
```

Step3：形態素リストを元に、解析対象の文書群をベクトルに変換します。

```
array([[0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0],
       [0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0],
       [1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0],
       [0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0],
       [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1],
       [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1],
       [0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1],
       [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1]], dtype=int64)
```

分類に寄与する文言

- 発話の種類ごとに、分類に寄与しそうな文言を入れるとモデルの精度が向上します。
- 例えば「class_0: 名前を聞かれる」の場合だと、"名前"という文言のことです。
- 「あなた」や「教えて」などの文言はどのクラスにも登場する可能性があり、発話の分類には寄与しないと考えられます。

```
sentences = [
    "名前は何",
    "名前をなんていうの",
    "名前教えて",
    "あなたのお名前は",
    "お名前教えてよ",

    "どんな色が好きなの",
    "何色が好き",
    "好きな色は何",
    "黄色は好き",
    "好きな色を教えてちょうだい",

    "どんな食べものが好きなの",
    "ピーマンは食べれる",
    "好きな食べものは",
    "食べものは何が好きなの",
    "何が美味しい",

    "歳はいくつですか",
    "歳はいくつになった",
    "何歳なの",
    "何歳か教えて",
    "何歳ですか",

    "おはよう",
    "おはようございます",
    "こんにちは",
    "こんばんは",
    "おやすみなさい",
]
```

演習2：BoWの実装

- 文章をベクトル化するプログラムを実装してください。

演習3：SVMによる分類モデル

- ベクトル化したデータでSVM（サポート・ベクター・マシン）のモデルを作成し、モデルの精度を確認してください。

SVMによる分類結果

- 下記の表はSVMによるモデルの分類結果の一例です。
- 「Class_2：好きな食べ物を聞かれる」の誤認識が多いことがわかります。
- 次回の講義では、モデル精度を向上させるための工夫について学んでいきます。

テストデータに対する正解率: 0.75

predict	0	1	2	3	4
class					
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	0	0	1
3	0	0	1	2	0
4	0	0	0	0	2

「Class_4：挨拶される」だとモデルは予測したが、実際は「Class_2：好きな食べ物を聞かれる」であった誤認識のケース

第13回：対話分類モデルの精度向上

コーパスの充実とベクトル化の工夫

アジェンダ

- 前回の振り返り
- モデル精度向上の施策

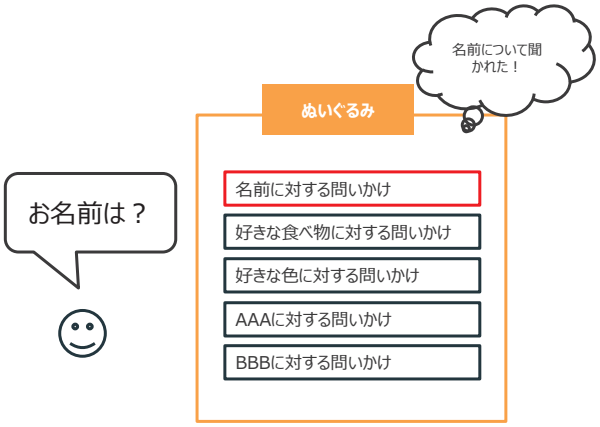
前回の振り返り

仕事の依頼

- 株式会社Lは金融工学が得意なベンチャー企業で、数学や工学が得意なメンバーが在籍しています。株式会社Lは、機械学習を金融分野に適用する仕事の実績があります。
- そんな株式会社Lに、玩具メーカーから自然言語処理分野の仕事の依頼がありました。玩具メーカーの人気商品に、人間と話す新機能を追加したいという依頼です。
- 「人間と話す」という先方の要望がどの程度の難易度の話なのか、ひとまず依頼先の会社を訪問して詳しい話を聞くことにしました。

ヒアリング

- 仕事を依頼してきたのは玩具メーカの株式会社Bでした。株式会社Bは、子供からの問いかけに対して答えることができるぬいぐるみを作りたいとのことです。
- 例えば「お名前は？」という子供の問いかけに対して、「私の名前はXXXだよ。」というように、簡単な会話が楽しめるようにしたいとのことです。
- 名前を聞かれる/好きな色を聞かれる/好きな食べ物を聞かれるなど、いくつかの問いかけに対して、まずは質問の分類ができるモデルを作成したいとのことでした。



コーパスの事前調査

- 株式会社L、株式会社Bともに、名前を聞かれる/好きな色を聞かれる/好きな食べ物を聞かれる、などを想定したコーパスは持っていませんでした。またライセンスの制約がないコーパスの調査をしましたが、適切なものが見つかりませんでした。
- そこで株式会社Lは、独自にコーパスを作成することにしました。

モデルの要件定義

- 株式会社Bは音声をテキストに変換するプロダクトを持っていたため、子供の発話をテキストに変換する機能はそのプロダクトを使用することにしました。したがってモデルのインプットは「テキスト」と決定しました。
- 分類する発話内容は下記の5つと決まりました。
  1. 名前を聞かれる
  2. 好きな色を聞かれる
  3. 好きな食べ物を聞かれる
  4. 年齢を聞かれる
  5. 挨拶される

SVMによる分類結果

- SVMでモデルを作成し、分類結果を考察しました。
- 「Class_2：好きな食べ物を聞かれる」の誤認識が多いことがわかります。
- 次のステップとして、モデル精度を向上させるための施策を実行することにしました。

テストデータに対する正解率：0.75

predict	0	1	2	3	4
class					
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	0	0	1
3	0	0	1	2	0
4	0	0	0	0	2

「Class_4：挨拶される」だとモデルは予測したが、実際は「Class_2：好きな食べ物を聞かれる」であった誤認識のケース

モデル精度向上の施策

コーパスの充実

- 学習データに表記揺れが含まれた方が、頑健なモデルが作成できる場合があります（※逆の発想で、ベクトル化の際にこのような表記揺れを落としてしまう手法もあります）。

Step1：解析対象の文章群を準備します。

```
['天気 を 教え て ください 。',
 '明日 の 天気 は どう です か ?',
 '今日 の 天気 を 教え て よ 。',
 '新宿 の 天気 は どう なっ て いる ?',
 '横浜 の 明日 の 天気 は どう か の かな ?',
 '気温 を 教え て よー 。',
 '明日 の 気温 は どう な の ?',
 '今日 の 気温 は 低い ね !',
 '横浜 の 気温 は ?',
 '新宿 の 昨日 の 気温 を 教え て ください 。']
```

Step2：重複しない形態素のリストを作成し、次元を決定します。

```
{'いる': 0,
 'かな': 1,
 'ください': 2,
 'です': 3,
 'どう': 4,
 'なっ': 5,
 'よー': 6,
 '今日': 7,
 '低い': 8,
 '天気': 9,
 '教え': 10,
 '新宿': 11,
 '明日': 12,
 '昨日': 13,
 '横浜': 14,
 '気温': 15}
```

Step3：形態素リストを元に、解析対象の文書群をベクトルに変換します。

```
array([[0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0],
       [0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0],
       [0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0],
       [1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0],
       [0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0],
       [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1],
       [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1],
       [0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1],
       [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1],
       [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1]], dtype=int64)
```

演習4：独自データの拡充

- データ数を増やして、モデルの精度の変化を確認してください。

```
sentences = [
    "名前は何",
    "名前は何ですか",
    "名前を何ていうの",
    "何ていうの名前は",
    "何ですか名前は",
    "名前を何ていうの",
    "あなたのお名前は",
    "お名前はあなたの",
    "お名前教えて",
    "お名前教えてよ",
```

表現の揺れを足してみる

SVMによる分類結果

- SVMでモデルを作成し、分類結果を考察しました。
- 全体の精度は向上し「Class_2：好きな食べ物を聞かれる」の誤認識は改善しましたが、「Class_4：挨拶される」の誤認識が増えてしまいました。

テストデータに対する正解率: 0.75

predict	0	1	2	3	4
class					
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	0	0	1
3	0	0	1	2	0
4	0	0	0	0	2

トレーニングデータに対する正解率: 0.87

predict	0	1	2	3	4
class					
0	2	0	0	0	0
1	0	3	0	0	0
2	0	0	2	0	2
3	0	0	0	5	0
4	0	0	0	0	1

文章のベクトル化：TF-IDF

- 文章をベクトル化する際に、単語に重みを付けて評価する手法です。単語の出現頻度であるTF（Term Frequency）と、IDF（Inverse Document Frequency）という2つの指標を使用します。
- TF（Term Frequency）は、「各文書においてその単語がどのくらい出現したのか」を意味します。よく出現する単語は、その文章の特徴を捉えるのに有用だろうという考え方です。
- IDF（Inverse Document Frequency）は、単語が稀にしか出現しないなら高い値を、「色々な文書によく出現する単語」なら低い値を示すものです。稀少な単語は、その文書の特徴を捉えるのに有用だろうという考え方です。

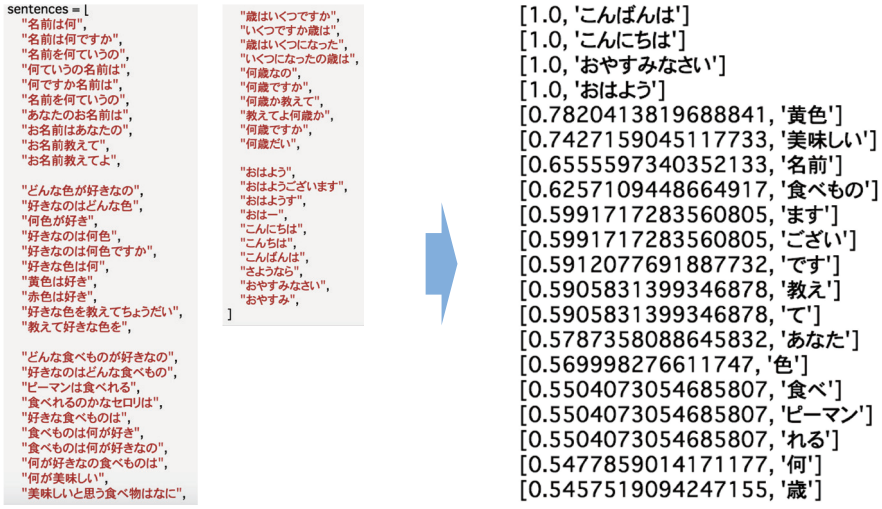
$$tf = \frac{\text{文書}A\text{における単語}X\text{の出現頻度}}{\text{文書}A\text{における全単語の出現頻度の和}} \qquad idf = log(\frac{\text{全文書数}}{\text{単語}X\text{を含む文書数}}) \qquad tfidf = tf * idf$$

演習5：TF-IDFの実装

- 作成したベクトルにTF-IDFを適用し、新たなベクトルを作成してください。
- TF-IDFで作成したベクトルの、スコアの高い形態素の上位N件を表示するプログラムを作成してください。

TF-IDFによる重要単語の判定

- TF-IDFで作成したベクトルの、スコアの高い形態素の上位N件を取得した例です。



```
sentences = [
    "名前は何",
    "名前は何ですか",
    "名前を何ていうの",
    "何ていうの名前は",
    "何ですか名前は",
    "名前を何ていうの",
    "あなたのお名前は",
    "お名前はあなたの",
    "お名前教えて",
    "お名前教えてよ",

    "どんな色が好きなの",
    "好きなのはどんな色",
    "何色が好き",
    "好きなのは何色",
    "好きなのは何色ですか",
    "好きな色は何",
    "黄色は好き",
    "赤色は好き",
    "好きな色を教えてちょうだい",
    "教えて好きな色を",

    "どんな食べものが好きなの",
    "好きなのはどんな食べもの",
    "ピーマンは食べれる",
    "食べれるのかなセロリは",
    "好きな食べものは",
    "食べものは何が好き",
    "食べものは何が好きなの",
    "何が好きなの食べものは",
    "何が美味しい",
    "美味しいと思う食べ物はなに",
    "歳はいくつですか",
    "いくつですか歳は",
    "歳はいくつになった",
    "いくつになったの歳は",
    "何歳なの",
    "何歳ですか",
    "何歳か教えて",
    "教えてよ何歳か",
    "何歳ですか",
    "何歳だい",

    "おはよう",
    "おはようございます",
    "おはようす",
    "おはー",
    "こんにちは",
    "こんちは",
    "こんばんは",
    "さようなら",
    "おやすみなさい",
    "おやすみ",
]
```

```
[1.0, 'こんばんは']
[1.0, 'こんにちは']
[1.0, 'おやすみなさい']
[1.0, 'おはよう']
[0.7820413819688841, '黄色']
[0.7427159045117733, '美味しい']
[0.6555597340352133, '名前']
[0.6257109448664917, '食べもの']
[0.5991717283560805, 'ます']
[0.5991717283560805, 'ござい']
[0.5912077691887732, 'です']
[0.5905831399346878, '教え']
[0.5905831399346878, 'て']
[0.5787358088645832, 'あなた']
[0.569998276611747, '色']
[0.5504073054685807, '食べ']
[0.5504073054685807, 'ピーマン']
[0.5504073054685807, 'れる']
[0.5477859014171177, '何']
[0.5457519094247155, '歳']
```

演習6：TF-IDFでベクトル化したデータによるモデル作成

- TF-IDFで作成したベクトルをSVM（サポート・ベクター・マシン）に投入してモデルを作成し、モデルの精度を確認してください。

SVMによる分類結果

- TF-IDFにより分類に寄与しない形態素を除去した結果、モデルの分類精度が向上したことが確認できました。

テストデータに対する正解率: 0.75

predict	0	1	2	3	4
class					
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	0	0	1
3	0	0	1	2	0
4	0	0	0	0	2

トレーニングデータに対する正解率: 0.87

predict	0	1	2	3	4
class					
0	2	0	0	0	0
1	0	3	0	0	0
2	0	0	2	0	2
3	0	0	0	5	0
4	0	0	0	0	1

テストデータに対する正解率: 1.00

predict	0	1	2	3	4
class					
0	2	0	0	0	0
1	0	3	0	0	0
2	0	0	4	0	0
3	0	0	0	5	0
4	0	0	0	0	1

演習7：モデルとテストデータの保存

- モデルとテストデータを保存してください。

第14回：簡易Web-API

サービス利用環境と提供環境の切り離し

アジェンダ

- 前回の振り返り
- Flaskを利用したWeb-APIの構築

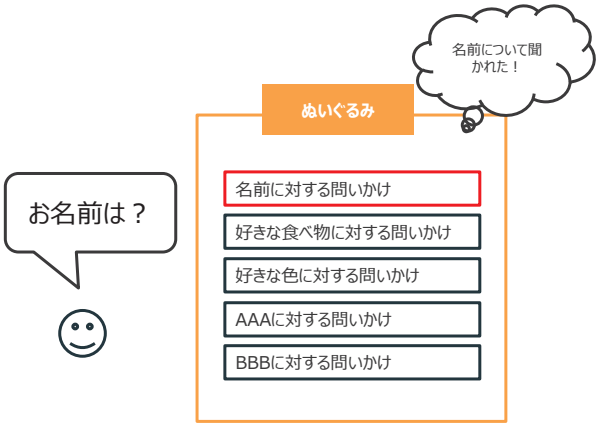
前回の振り返り

仕事の依頼

- 株式会社Lは金融工学が得意なベンチャー企業で、数学や工学が得意なメンバーが在籍しています。株式会社Lは、機械学習を金融分野に適用する仕事の実績があります。
- そんな株式会社Lに、玩具メーカーから自然言語処理分野の仕事の依頼がありました。玩具メーカーの人気商品に、人間と話す新機能を追加したいという依頼です。
- 「人間と話す」という先方の要望がどの程度の難易度の話なのか、ひとまず依頼先の会社を訪問して詳しい話を聞くことにしました。

ヒアリング

- 仕事を依頼してきたのは玩具メーカの株式会社Bでした。株式会社Bは、子供からの問いかけに対して答えることができるぬいぐるみを作りたいとのことです。
- 例えば「お名前は？」という子供の問いかけに対して、「私の名前はXXXだよ。」というように、簡単な会話が楽しめるようにしたいとのことです。
- 名前を聞かれる/好きな色を聞かれる/好きな食べ物を聞かれるなど、いくつかの問いかけに対して、まずは質問の分類ができるモデルを作成したいとのことでした。



SVMによる分類結果

- TF-IDFにより分類に寄与しない形態素を除去したデータとSVMを使用し、分類モデルを作成しました。

テストデータに対する正解率：0.75

predict	0	1	2	3	4
class					
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	0	0	1
3	0	0	1	2	0
4	0	0	0	0	2

BoWによる初回モデル

トレーニングデータに対する正解率：0.87

predict	0	1	2	3	4
class					
0	2	0	0	0	0
1	0	3	0	0	0
2	0	0	2	0	2
3	0	0	0	5	0
4	0	0	0	0	1

コーパスを拡充したBoWによるモデル

テストデータに対する正解率：1.00

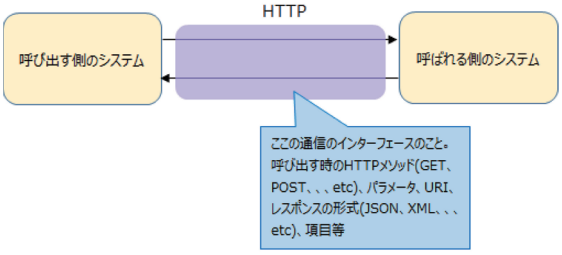
predict	0	1	2	3	4
class					
0	2	0	0	0	0
1	0	3	0	0	0
2	0	0	4	0	0
3	0	0	0	5	0
4	0	0	0	0	1

コーパス拡充後、TF-IDFを実施したモデル

FlaskをWeb-APIの構築

Web APIとは

- HTTPプロトコルを用いてネットワーク越しに呼び出すアプリケーション間、システム間のインターフェースのことです。
- システムの中身の動作は詳細に把握しなくても、機能の塊を外部から呼び出すことができます。特に右側の「呼ばれる側のシステム」のことをWeb APIと呼びます。



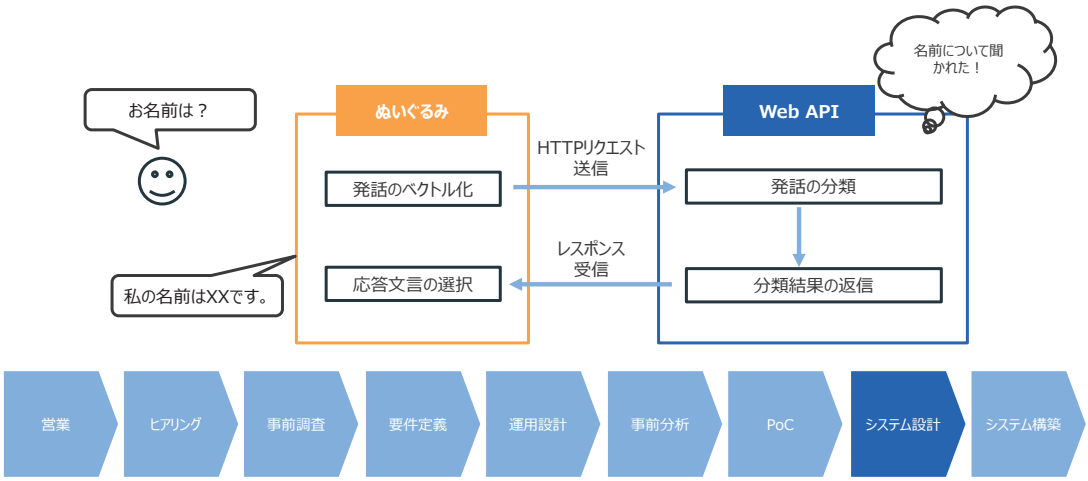
出展：https://qiita.com/NagaokaKenichi/items/df4c8455ab527aeacf02

Flaskとは

- Pythonのための軽量なウェブアプリケーションフレームワークです。
- Pythonで用いられるフレームワークとしてはDjangoなども人気がありますが、簡易なAPIを作成する場合はFlaskを採用したほうが速やかにAPIを構築することができます。
- 公式サイト（http://flask.pocoo.org/）

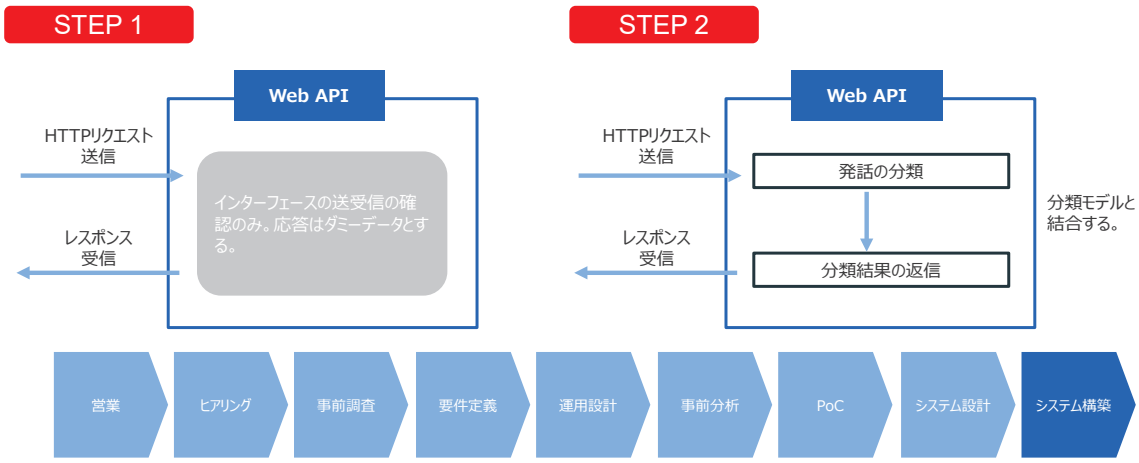
分類モデルの切り出し

- 株式会社Bの意向として、分類モデルはWeb APIとして切り出したいそうです。そうすることによってWeb APIに搭載されている分類モデルを更新すれば、全てのぬいぐるみに対して即座に最新の分類機能を提供することができます。



Web API構築ステップ

- Web APIの構築は2ステップに分けることになりました。STEP 1ではインターフェースの構築のみを行い、STEP 2で分類モデルと結合することになりました。



演習1：Web APIによるGETリクエストの処理

- GETリクエストを処理できるAPIを作成してください。

演習2：Web APIによるGETリクエストの処理

- パラメータ付きのGETリクエストを処理できるAPIを作成してください。

演習3：Web APIによるPOSTリクエストの処理

- パラメータ付きのPOSTリクエストを処理できるAPIを作成してください。

第15回：Web-APIと分類モデルの連携

サービス利用環境と提供環境の切り離し

アジェンダ

- 前回の振り返り
- Web-APIからのSVMモデルの呼び出し

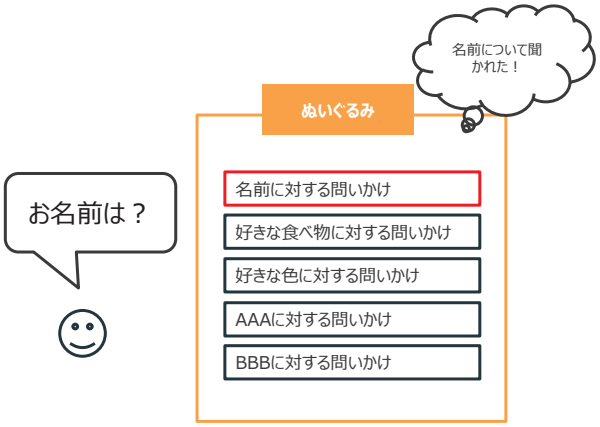
前回の振り返り

仕事の依頼

- 株式会社Lは金融工学が得意なベンチャー企業で、数学や工学が得意なメンバーが在籍しています。株式会社Lは、機械学習を金融分野に適用する仕事の実績があります。
- そんな株式会社Lに、玩具メーカーから自然言語処理分野の仕事の依頼がありました。玩具メーカーの人気商品に、人間と話す新機能を追加したいという依頼です。
- 「人間と話す」という先方の要望がどの程度の難易度の話なのか、ひとまず依頼先の会社を訪問して詳しい話を聞くことにしました。

ヒアリング

- 仕事を依頼してきたのは玩具メーカの株式会社Bでした。株式会社Bは、子供からの問いかけに対して答えることができるぬいぐるみを作りたいとのことです。
- 例えば「お名前は？」という子供の問いかけに対して、「私の名前はXXXだよ。」というように、簡単な会話が楽しめるようにしたいとのことです。
- 名前を聞かれる/好きな色を聞かれる/好きな食べ物を聞かれるなど、いくつかの問いかけに対して、まずは質問の分類ができるモデルを作成したいとのことでした。



SVMによる分類結果

- TF-IDFにより分類に寄与しない形態素を除去したデータとSVMを使用し、分類モデルを作成しました。

テストデータに対する正解率：0.75

predict	0	1	2	3	4
class					
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	0	0	1
3	0	0	1	2	0
4	0	0	0	0	2

BoWによる初回モデル

トレーニングデータに対する正解率：0.87

predict	0	1	2	3	4
class					
0	2	0	0	0	0
1	0	3	0	0	0
2	0	0	2	0	2
3	0	0	0	5	0
4	0	0	0	0	1

コーパスを拡充したBoWによるモデル

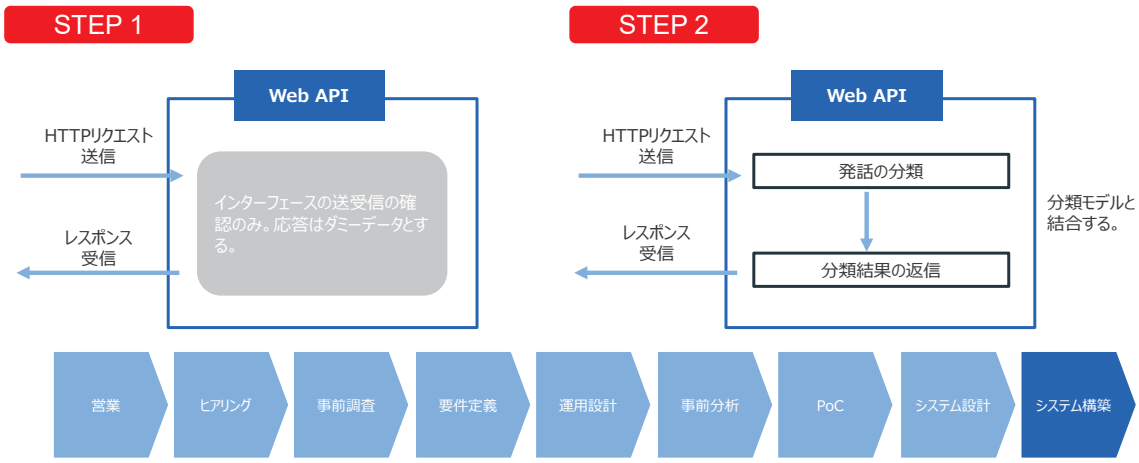
テストデータに対する正解率：1.00

predict	0	1	2	3	4
class					
0	2	0	0	0	0
1	0	3	0	0	0
2	0	0	4	0	0
3	0	0	0	5	0
4	0	0	0	0	1

コーパス拡充後、TF-IDFを実施したモデル

Web API構築ステップ

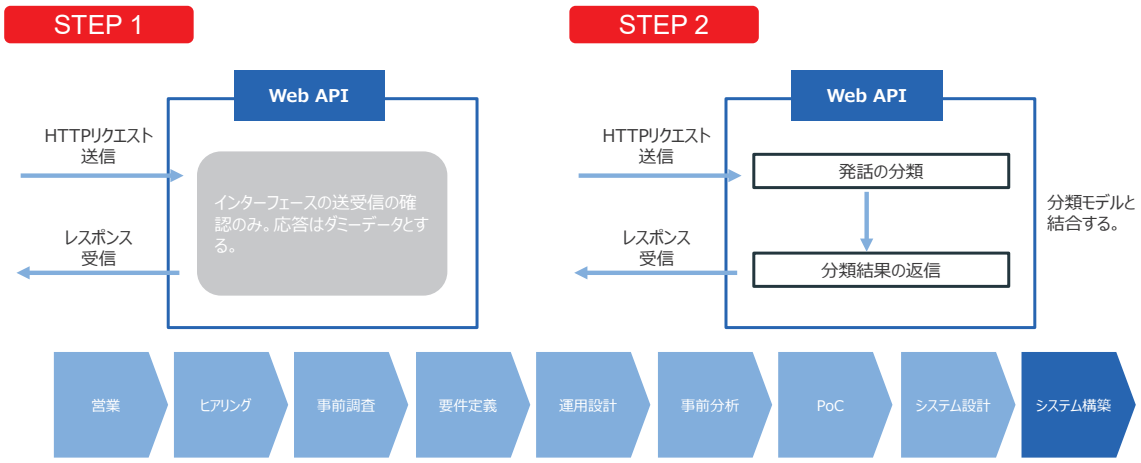
- Web APIをの構築は2ステップに分けることになりました。STEP 1ではインターフェースの構築のみを行い、STEP 2で分類モデルと結合することになりました。



Web-APIからのSVMモデルの呼び出し

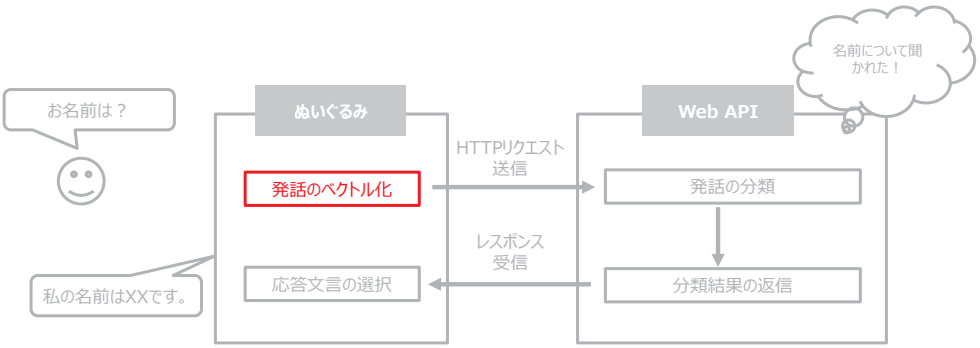
STEP 2：モデルとの結合

- 前回の講義ではSTEP 1（インターフェースの構築）を実施しました。
- 本講義ではSTEP 2（分類モデルと結合）を実施します。



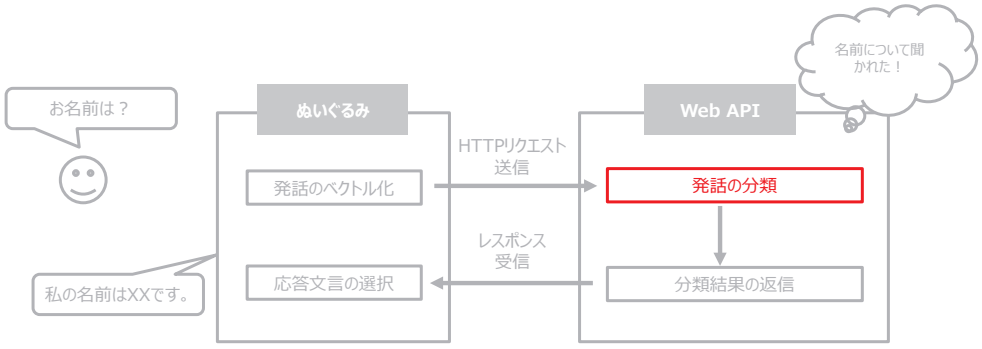
演習1：テストデータの読み込み

- SVMモデルに判定させたいベクトルを読み込んでください。



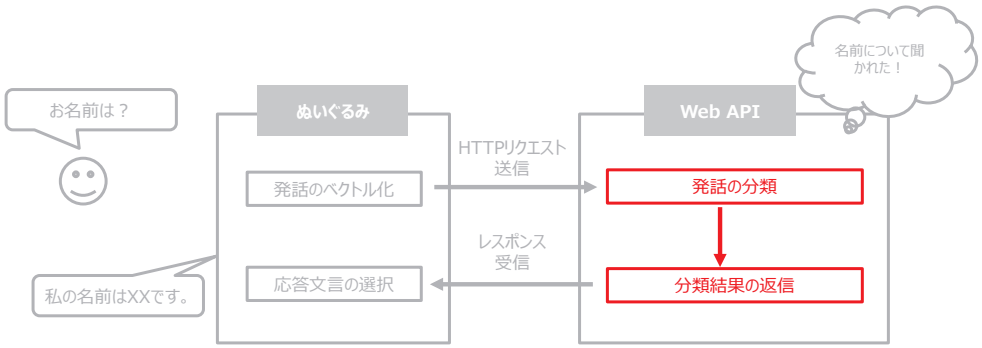
演習2：保存済み分類モデルの読み込み

- 作成済みのSVMモデルを読み込み、パラメータ入力されたベクトルを分類する関数を実装してください。



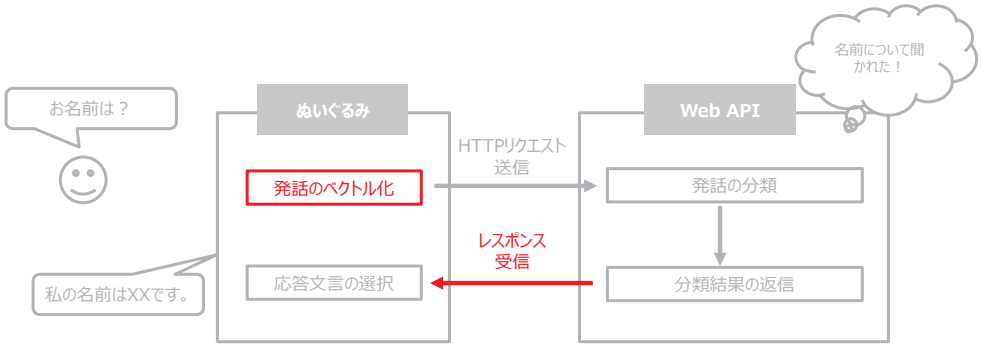
演習3：Web APIの拡張

- SVM分類器にアクセスするエンドポイントをAPIに追加実装してください。



演習4：Web APIへのアクセス

- SVMモデルに判定させたいベクトルをパラメータとし、HTTPリクエストを作成してください。
- 対話分類APIにリクエストを送信し、応答を観察してください。



2019 年度「専修学校による地域産業中核的人材養成事業」
Society5.0 実現のための IT 技術者養成モデルカリキュラム開発と実証事業

2019 年度「専修学校による地域産業中核的人材養成事業」
Society5.0 実現のための IT 技術者養成モデルカリキュラム開発と実証事業

AI システム開発教材

令和２年２月

学校法人電子学園（日本電子専門学校）
〒169-8522　東京都新宿区百人町 1-25-4
TEL 03-3369-9333　FAX 03-3363-7685