ビジネス理解 → データ理解 → データ準備

2017年の千葉市の
インフルエンザの
発生予測を行う

インフルエンザ発生
は気候と関係がありそう

過去3年分の
気象データが必要

過去3年分の
インフルエンザ報告数
データが必要

気象データの確認と加工

①csvファイルからDataFrameを生成

②不要なデータを削除

③欠損値処理

④特徴量を選定

⑤日付をインデックスで設定

データ準備

インフルエンザ報告データの確認と加工

①csvファイルからDataFrameを生成

②必要なデータを取得

③日付をインデックで設定
※１週間ごとであることに注意

④日付をインデックスで設定

⑤2014年、2015年、2016年ごとで処理

気象データとインフルデータをマージ（その１）

①日付をキーにしてマージ

②インフルデータは１週間ごとなので欠損が発生

③欠損値処理（前週の値をセット）

データ準備

インフルエンザ報告者数から流行と増加を定義

①報告者から10以上なら流行

②１週前より５より大きければ増加

③インフルデータに流行と増加の列を追加

④データ保存

気象データとインフルデータをマージ（その②）

①日付をキーにしてマージ

②インフルデータは１週間ごとなので欠損が発生

③欠損値処理（前週の値をセット）

④相関関係を見て特徴量を選択

⑤データ保存

モデリング

ロジスティック回帰で予測モデル構築

①データ読み込み

②説明変数(X)と目的変数(y)を抽出

③ホールドアウト法でデータ分割

④ロジスティック回帰で学習/予測

⑤評価を可視化

複数モデルを構築し、どれが汎化能力が高いかを検証

・サポートベクターマシン SVC
・カーネルSVM
・決定木 DecisionTreeClassifier
・ランダムフォレスト RandomForestClassifier
・k近傍

※流行の予測が高いモデルが良いモデルと判断できる

評価

2017年のインフルエンザを予実を確認する

①2017年1月以降のデータ取得

②データの加工

③学習済みデータを読み込み

④学習済みデータで予測

⑤2017年のインフル報告データと予測結果で予実確認