


HATENA SUMMER INTERNSHIP 2018

機械学習 実践編

目次

- ・ **機械学習のワークフロー** 
- ・ 特徴量エンジニアリング
- ・ テキストデータの処理

機械学習のワークフロー

1. 問題設定
2. 前処理・特徴量作成
3. 学習・チューニング
4. サービスへの組み込み

ステップ1：問題設定

- 解決したい課題は何か
 - 改善したいビジネス指標と、改善のためのアクションが具体的だとよい
- 典型的な機械学習タスクに帰着させる

機械学習を使わないという選択

- ・ 機械学習アプリケーションは技術的負債になりがち
 - ・ テストしづらい
 - ・ コードが理解しづらい
 - ・ 長期の運用で挙動が変化

ステップ2：前処理・特徴量作成

- ・ データを集める
- ・ 生のデータを機械学習モデルの入力形式（特徴量）に変換する
- ・ 詳しくはこの講義で後ほど説明


ステップ3：学習・チューニング

- ・ 基礎編を参照
- ・ チューニングしても十分な性能が出ないときは特徴量の作り方、もしくは問題設定から見直す
- ・ 学習アルゴリズムの工夫より、こちらが重要なことも多い

ステップ4：サービスへの組み込み

- ・ モニタリング
- ・ モデルの性能
- ・ ステップ1で考えたビジネス指標

目次

- ・ 機械学習のワークフロー
- ・ **特徴量エンジニアリング** 
- ・ テキストデータの処理

特徴量作成

- ・ 機械学習モデルの入力：実数値のベクトル（もしくは行列）
- ・ 生データは必ずしもモデルの入力として理想的な形ではない

カテゴリ変数

- ・ 例：天気
 - ・ 値は「曇り」「晴れ」「雨」の3通りとする

ワンホットエンコーディング

天気

データ1

曇り

データ2

晴れ

ワンホットエンコーディング

| | 天気_晴れ | 天気_曇り | 天気_雨 |
|------|-------|-------|------|
| データ1 | 0 | 1 | 0 |
| データ2 | 1 | 0 | 0 |

ビンニング (binning)

- 区間 (bin) で分割して離散化
- 例：年齢
 - 「10代」 「20代」 のようなカテゴリ変数に変換

スケーリング

- ・ 標準化：平均0、分散1に変換
- ・ Min-Max スケーリング：最小値0、最大値1に変換
- ・ 特に非線形のモデルで大きく性能が向上することもある

組み合わせ特徴量 (interaction features)

- 単語 a, b が出現するかを表す特徴量 $x_a, x_b \in \{0, 1\}$
- 新しい特徴量 $x_{ab} = x_a x_b$
 - 単語 a, b が両方出現することを表す特徴量
- 線形なモデルでも元々の特徴量に対して非線形になる


特徴選択 (feature selection)

- 重要な特徴量だけに絞った方が性能が上がることもある
- やり方はさまざま
 - 貪欲法で特徴量を落としていく
 - 決定木やL1正則化のような教師あり学習の手法
 - 統計量などの指標を用いる

前処理方法の評価・選択

- ・ 前処理もモデルの一部と考えて（交差検定などで）評価する
- ・ 前処理の内容が学習（訓練データ）と予測（テストデータ）で異ならないよう注意

目次

- ・ 機械学習のワークフロー
- ・ 特徴量エンジニアリング
- ・ **テキストデータの処理** 

テキスト特徴量

- テキストデータ特有の前処理を行って特徴量にする
 - BoW表現
 - 単語埋め込み (word2vecなど)

Bag of Words (BoW) 表現

- 単語ごとの出現頻度をベクトルにまとめたもの
- ベクトル中のインデックスが暗黙に単語と対応
- 単語が文書の中で現れる順番は考慮できない
- 「私はラーメンは好きだが、つけ麺は嫌いだ」
- 「私はつけ麺は好きだが、ラーメンは嫌いだ」

テキストデータの前処理

- まず生のテキストを単語の列にする
 1. 文字の正規化
 2. 単語分割
 3. 単語の正規化
 4. ストップワード除去

文字の正規化

- Unicode正規化
 - 例：「U S （全角）」→「US （半角）」
- アクセント記号の削除

単語分割

- 英語の場合：ホワイトスペースで区切る
- 日本語の場合：形態素解析器を用いる
- MeCab、Kuromojiなどが有名

単語の正規化

- 単語の活用形を見出し語に正規化 (lemmatization)
- 形態素解析器が見出し語や品詞などの情報をくれる
- 辞書にない語 (typoなど) を編集距離の近い単語に正規化

ストップワード除去

- ・ あまり意味のない頻出語を取り除く
- ・ ストップワード辞書
- ・ 品詞（冠詞、助詞など）を見る

BoW表現への変換

- 辞書の作成
 - 単語にIDを振る
 - IDが特徴ベクトル中のインデックスを表す
- 単語出現数のカウント

単語出現数のカウント

- 単語頻度 (term frequency) : その文書に単語が現れた数
- 文書頻度 (document frequency) : データセット中にその単語が現れた文書の数

BoW表現のバリエーション

- 単語が出現したかどうかの2値
- 単語頻度
- 単語頻度を対数変換した値
- tf-idf：単語頻度と文書頻度の逆数（+対数変換）をかけた値

高次元でスパースなデータの扱い

- ・ 疎ベクトル（疎行列）表現を使う
- ・ 次元を減らす
 - ・ 単語の正規化
 - ・ 文書頻度が高すぎる、もしくは低すぎる単語を足切り
 - ・ 特徴選択や次元削減の手法を適用