

『言語処理のための機械学習入門』要点まとめ

ばろすけ

2013 年 6 月 11 日

0 この文書について

専ら自身の学習のために趣味で作成したものです。これさえ読み返せば内容が思い返せる（必要に応じて参照できる）ことを目指します。

1 必要な数学的知識

1.1 準備と本書における約束事

2 文書および単語の数学的表現

2.1 タイプ、トークン

単語トークンとは、ひとつひとつの単語の出現を指し、同じ単語が複数回出現しても別のものとして数える。単語タイプとは単語の種類を指す。

2.4 文書に対する前処理とデータスパースネス問題

2.4.1 文書に対する前処理

”the”などの話題の種類と関連を持たないと考えられる単語をストップワードとよぶ。派生語なども含めて同一の素性とみなす作業をステミングと呼ぶ。ポーターのステマーなどが知られる。単語を基本形に戻す作業を見出し語化という。

2.5 単語のベクトル表現

2.5.1 単語トークンの文脈ベクトル表現

単語トークンをベクトル化するとき、その単語自身ではなく、その前後にどのような単語が出現しているかでベクトル化したものを文脈ベクトルという。

3 クラスタリング

3.2 凝縮型クラスタリング

凝縮型クラスタリングではクラスタ同士の類似度の測り方が複数考えられる。単連結法では2つのクラスタ内でもっとも近い事例対の類似度をクラスタの類似度とする。完全連結法ではもっとも遠い事例対の類似度をクラスタの類似度とする。重心法では各クラスタの重心同士の類似度を用いる。完全連結法では長く伸びたクラスタができにくい。

3.5 EM アルゴリズム

EM アルゴリズムは、クラスタリングの文脈でよく登場するが、一般に多変数確率分布において変数の一部が観測できない場合にパラメータを推定する手法である。E ステップでは仮パラメータを用いて隠れ変数を求める。M ステップでは求められた隠れ変数の値を用いてパラメータを更新する。クラスタリングにおいては仮パラメータが各事例の所属するクラスタに該当する。

3.7 この章のまとめ

スペクトラルクラスタリングは次元圧縮を行った後に別の方法を用いてクラスタリングを行う。自己組織化マップはデータの視覚化に重点が置かれているので注意が必要である（詳細の記述なし）。

4 分類

5 系列ラベリング

6 実験の仕方など

6.3 評価指標

6.3.3 精度と再現率の統合

F 値とは、精度と再現率の調和平均である。

精度と再現率が一致する点における精度を再現率/精度 break-even ポイントと呼び、評価指標としてよく用いられる。

6.3.5 評価指標の平均

多クラスの分類問題を考える。各クラスの F 値の平均をとったものをマクロ平均という。一方でそれぞれのクラスの分割表を統合した表から求めたものをマイクロ平均という。マクロ平均は各テストデータセットの大きさを無視して平等に扱うため注意が必要である。

6.4 検定

検定の種類と用途だけが列挙される。既存手法と提案手法について、対応ある二つの結果の組に差が定義されない場合は符号検定を用いる。差が定義できる場合はウィルコクソンの符号付順位和検定を用いる。結果が正規分布に従っていると考えられる場合は t -検定を用いるが、まず正規分布に従っているかを検定するためにコルモゴロフ・スミルノフ検定などを用いる。