『言語処理のための機械学習入門』要点まとめ

ぱろすけ

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

4.2 ナイーブベイズ分類器

ナイーブベイズ分類器では事例に対して $P(c|d)$ が最大になるクラスを出力する。ベイズの定理を用いると

$$c_{\max} = \underset{c}{\operatorname{argmax}} \frac{P(c)P(d|c)}{P(d)} = \underset{c}{\operatorname{argmax}}\, P(c)P(d|c)$$

であるが、一般に $P(d|c)$ を計算することでは容易ではない。そのため簡易化したモデルが利用される。

4.2.1 多変数ベルヌーイモデル

多変数ベルヌーイモデルではクラス対する単語の含まれやすさでモデル化する。具体的には以下である。

$$P(d|c) = \prod_{w \in d} p_{w,c} \prod_{d \notin d} (1 - p_{w,c})$$

このパラメータを最尤推定で求めると、等式制約付き凸計画問題として解け、

$$p_{w,c} = \frac{\text{あ}}{\text{い}}$$

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。