Semi-Supervised Learning and Domain Adaptation in Natural Language Processing

2.5 Comparisons of Classification Algorithms
2.6 Learning from Weighted Data
2.6.1 Weighted k-nearest Neighbor
2.6.2 Weighted Naive Bayes
2.6.3 Weighted Perceptron
2.6.4 Weighted Large-margin Learning

理工学研究科
國井慎也

Comparisons of Classification Algorithms

- 3つの識別アルゴリズムを比較する
  - ナイーブベイズ、パーセプトロン、最近傍法
  - 20のニュースグループからランダムに選ぶ
  - 25回実験を行う
  - accuracyとKL情報量との相関係数を求める

learner	acc	p(KL)
nb	0.753	−0.22
perc	0.709	−0.09
nn	0.614	−0.27

実験から分かったこと

- パーセプトロンについて
  - 正解率とKL情報量に相関関係はない
    - Sourceとtargetドメインが異なっていても大丈夫

- 学習する際の素性の重みについて
  - 全ての素性が同じ重み

    perceptron
  - 頻繁に出現する素性の重みがかかる

    Nearest nenighor

クラス内分散とacc

- クラス内分散とaccuracyとの相関関係

learner	acc	p(wcs)
nb	0.842	−0.34
perc	0.868	−0.03
nn	0.627	−0.01

- 実験から
  - Perceptronは、クラス内分散に敏感ではない
  - Naïve Bayesのaccはクラス内分散と相関がある
    決定境界を作るから

Weighted k-Nearest Neighbor

- 同じクラスのデータは近くにあると仮定
  - より近いデータに重みをつける

$$\arg max_c \, vote(x, c) = \sum_{(e_i, c_i) \in N_k} w(x, e_i) \delta(c, c_i)$$

$w(x, e_i)$:重み（距離関数）

Weighted Naive Bayse

- 事前分布βを導入、各データに対して重みを付ける

β	y	zebra	viagra	venus
β_1:0.6	spam	0	1	0
β_2:0.2	non−spam	1	0	0
β_2:0.3	non−spam	1	0	1

$$p(spam) = \frac{0.6}{1.1}$$

$$p(venus = 1 \mid non\text{-}spam) = 3/5$$

Weighted perceptron

- 学習時のパラメータwの更新時にβを導入

$$\boldsymbol{w}^{i+1} \leftarrow \boldsymbol{w}^{i} + \beta_n \alpha\left(y_n - sign\left(\boldsymbol{w}^{i} \cdot x_n\right)\right)\boldsymbol{x}_n$$

分類が失敗したときに更新

- Weighted large-margin leaning の更新式も$\beta\alpha$を導入する。