Semi-Supervised Learning and Domain Adaptation in Natural Language Processing

2.2 NEAREST NEIGHBOR

2.3 NAIVE BAYES

小野寺喜行

Nearest neighbor

ノンパラメトリックな手法

最近傍のラベル付きデータをもとに分類

↓

距離測定法

例)ユークリッド距離、マンハッタン距離、ハミング距離

わずかに異なるラベル付きデータ

↓

非常に異なる結果を導く可能性あり

↓

k-nearest neighbor

k-nearest neighbor

k個の近傍のラベル付きデータから各ラベルの数を調べ、
ラベル無しデータが各ラベルである確率を推定

xがクラスy_iである確率 $\hat{p}\left(y_i \mid \mathbf{x}\right) = \frac{k_i}{k}$

k_i : k近傍のクラスy_iの個数

kが小さいとき過学習する傾向がある

効率やメモリの面から、
大規模なデータでは用いる事が出来ない

k-nearest neighbor

Figure:

(a) Performance of k-nearet neighbor with varying k on HOCKEY-BASEBALL

(b) Learning curvy for 1-nearest neighbor on HOCKEY-BASEBALL

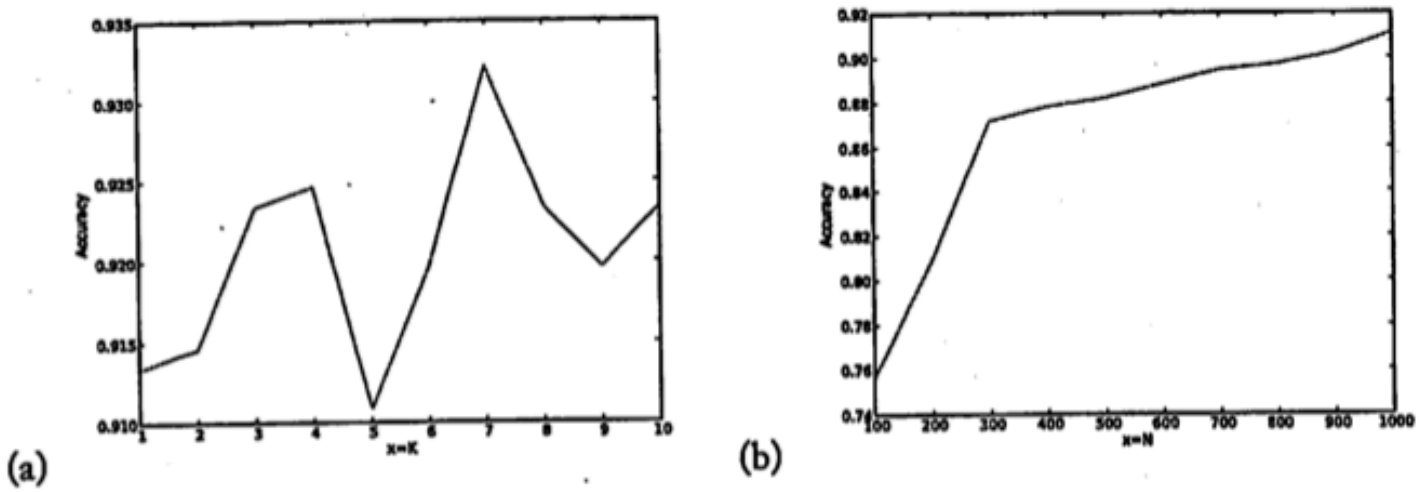

Figure 2.4: (a) Performance of k-nearest neighbor with varying k on HOCKEY-BASEBALL. (b) Learning curve for 1-nearest neighbor on HOCKEY-BASEBALL.

Naive Bayes

連鎖率 $P(A,B) = P(A)P(B \mid A) = P(B)P(A \mid B)$

お互いに独立なとき $P(A,B) = P(A)P(B)$

新しいデータ$\mathbf{x}$がクラスyとなる確率

$$P(y \mid \mathbf{x}) = \frac{P(y)P(\mathbf{x} \mid y)}{P(\mathbf{x})} = \frac{P(y)\prod_i P(x_i \mid y)}{P(\mathbf{x})}$$

最も近いクラスを見つけるために以下を計算

$$\hat{y} = \arg\max_{y \in Y} P(y)\prod_i P(x_i \mid y)$$

Naive Bayes

線形決定境界を学習することが可能

xの分類決定が下記の式にするために重みベクトル**w**として、線形分類するものが作成できる

$$\mathrm{sign}\left(\mathbf{w}\cdot\mathbf{x}+b\right)=\mathrm{sign}\left(\sum_i w_i x_i+b\right)$$

$$\mathbf{w}=\left(\log\theta_1-\log\theta_0\right)^{\mathrm{T}}$$

$$b=\left(\log P\left(y=1\right)-\log P\left(y=0\right)\right)$$

θ_yはクラスyでの多項分布

Naive Bayes

Figure: Learning curve for Naive Bayes on Gweb-Email-Transitions

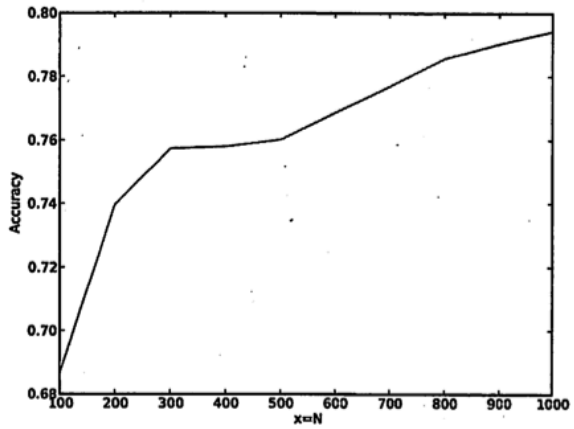

Figure 2.5: Learning curve for Naive Bayes on GWEB-EMAIL-TRANSITIONS.

Naive Bayes

学習の開始点では一般的に良い結果となる

最上の精度に早く到達する

訓練データが少ない場合に向いている