サポートベクトルマシン 5.1 - 5.2節

参考文献

- 金森, 統計的学習理論, 講談社, 2015.

Table of contents

導入

- C-サポートベクトルマシン
  - データ数に応じて適切にCを調整することで，統計的一致性が達成される．
- ν-サポートベクトルマシン
  - ベイズ誤差を達成するためにはデータの分布の情報を用いてνを調整する必要がある．
    - 実用面 + 理論的な観点で交差確認法などによるパラメータ調整が必要．

C, νは正則化パラメータ．

$\rightsquigarrow$ これらを適切に調整すれば，これらの学習アルゴリズムは同じ学習データに対して同じ判別器を返す．

問題設定

Train data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{+1, -1\}$: test dataと同一分布，独立
目標
入力ベクトルと2値ラベルの間の関係をデータから学習し，予測精度の高い判別器を得ること.

- $\mathcal{G} = \{f + b \mid f \in \mathcal{H}, b \in \mathbb{R}\}$: 判別関数の集合
- $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$: $\mathcal{X}$上のRKHS
- $\| \cdot \|_{\mathcal{H}}$: $\mathcal{H}$のノルム
- $k(x, x')$: 対応するカーネル関数
- $\mathrm{sign} \circ \mathcal{G} = \{x \mapsto \mathrm{sign}(g(x)) \mid g \in \mathcal{G}\}$: 対応する判別器の集合 (仮説集合)

ヒンジ損失

$$\phi_{\mathrm{hinge}}(m) = \max\{1 - m, 0\}$$

から定義されるマージン損失

ヒンジ損失の特徴付け

学習データ(x_i, y_i)に対して，$f(x_i) + b$の符号がy_iと同じなら，$\mathrm{sign}(f(x_i) + b)$によって$y_i$を正しく判別できる
$\leadsto$ できるだけ多くの学習データに対して，$y_i(f(x_i) + b) > 0$が成立すれば，学習データに適合しているという意味で望ましい判別器が得られる
$\leadsto$ 0-1マージン損失$\phi_{\mathrm{err}}(m)$から定義される経験損失を最小化することで実現される

$\phi_{\mathrm{err}}(m) = \mathbf{1}[m \leq 0]$は凸関数ではないので，最小化は一般に困難
$\leadsto$ 0-1マージンの代わりに凸関数から定義されるマージン損失を用いることで，計算の困難を回避する

凸マージン損失$\phi(m)$が，$\phi'(0) < 0$を満たすとき，判別適合的な損失になる．
$\rightsquigarrow$ 損失を最小化して得られる判別器は，判別関数の統計モデルが十分大きいなら，ベイズ誤差に近い予測判別誤差を達成することが期待される．

Proposition 5.1

凸関数$\phi : \mathbb{R} \to \mathbb{R}$は原点で微分可能で，$\phi'(0) = -1$を満たすとする．また，$\rho > 0$として，任意の$m \in \mathbb{R}$に対して$\rho \cdot \mathbf{1}[m \leq 0] \leq \phi(m)$が成り立つとする．このとき，$\forall m \in \mathbb{R}$に対して

$$\rho \cdot \mathbf{1}[m \leq 0] \leq \max\{\rho - m, 0\} \leq \phi(m)$$

が成り立つ．

C-サポートベクトルマシンでは$\rho = 1$とした通常のヒンジ損失を用いる．ν-サポートベクトルマシンではρを可変パラメータとして扱う．