第4章
凸計画問題を用いたカーネル多変量解析(1)

佐々木研究室
06T4056N　三沢 博章

はじめに

第4章では、
サポートベクトルマシン（SVM)に代表されるような
比較的最近になって研究された手法について述べる。

計算法：
・線形計画法
・凸二次計画法
これらは、固有値問題を使った手法より難しい。
なぜ、このような計算法を使用するのか？

続き

- 損失関数のロバスト性

  前章は2次式の損失関数が前提。

  外れ値がある場合に深刻な問題となる。

  ここでの手法は、外れ値に対しても

  頑健(ロバスト)なデータ解析が可能であるというメリットがある。

- 結果のスパース性

  本章で述べる手法から得られる解は、

  サンプルの中でもほんの一部分だけしか使われないので、

  計算量やメモリ量の節約ができるというメリットがある。

SVMの損失関数

誤識別関数・・・クラス識別ではよく使われている。

正しい出力と学習した出力値との誤差

凸関数ではないので、局所最適解を複数持つため最適化が難しいという欠点がある。

区分線形関数・・・凸関数の中で誤識別関数に近い関数

カーネル最小2乗クラス識別において、区分線形関数 γ_{hinge} を損失として用いているのがSVMである。

誤差の増え方が緩やかであるため、外れ値に対するロバスト性をもつ。

SVMの定義

SVMでは損失関数としてγ_{hinge}を取る以外カーネル回帰と同じ

よって、

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x^{(i)}, x) \qquad (1)$$

と書ける。そこで求める値は、

$$f(x) = \sum_{i=1}^{n} \gamma_{hinge}(y^{(i)} f(x^{(i)})) + \lambda \alpha^T K \alpha \qquad (2)$$

という最小化問題を解くことで求められる。

（凸関数なので最適解は一つしかない）

凸二次計画問題

式(2)では、実際に関数値を計算するためには、
yf(x)の値での場合分けが必要なため最適化が困難
そこで、式(2)を凸二次計画問題に変形する。
式(2)の変形:
サンプルx^(i)、y^(i)に対するγ_{hinge}関数の値をξ_iとすると

$$\xi_i \geq 0, \quad \xi_i \geq 1 - y^{(i)} f(x^{(i)}) = 1 - y^{(i)} \sum_{j=1}^{n} \alpha_j K_{ij} \qquad (3)$$

上式を同時に満たすξ_iの値のうちで最小となる値である。
そこで式(2)を式(3)とξ_iを使って変形すると、

$$\min_{\xi_i \alpha} \sum_{i=1}^{n} \xi_i + \frac{\lambda}{2} \alpha^T K \alpha \qquad (4)$$

続き

式(4)のように変形することにより、

$\xi = (\xi_1, \ldots, \xi_n)^T$ と α に関する2次関数の最小化問題

↓

凸二次計画問題を解くことに帰着できる。

メリット:

凸二次計画問題には、これを解くための数理計算パッケージが存在するので、計算時間を大幅に減少できる。

ラグランジュの未定乗数法

制約付きの最適化問題を解くには、
「ラグランジュの未定乗数法」を使用する。
式(4)にラグランジュの未定乗数β_i、γ_i(i=1,…,n)を導入すると、

$$L(\xi,\alpha,\beta,\gamma)=\sum_{i=1}^{n}\xi_i+\frac{\lambda}{2}\alpha^T K\alpha-\sum_{i=1}^{n}\beta_i\xi_i-\sum_{i=1}^{n}\gamma_i(\xi_i-1+y^{(i)}\sum_{j=1}^{n}\alpha_j K_{ij}) \qquad (5)$$

という極値問題となる。不等式制約より、
$\beta_i \geqq 0$、$\gamma_i \geqq 0$を満たす。
一般に制約付きの凸最適化問題の解が満たす条件を
ラグランジュ関数によって述べたのが、
「カルーシュ-キューン-タッカー定理(KKT定理)」

カルーシュ-キューン-タッカー定理

m個の不等式制約をもつ最適化問題

$$\min_x f(x), \quad g_i(x) \le 0, \quad i = 1,...,m \qquad (6)$$

において、f,g_iは微分可能な凸関数であるとすると、

$$L(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \qquad (7)$$

とラグランジュ関数をおけ、ある正則条件を満たすと仮定する。

$$\nabla L(x^*,\lambda) = \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) \qquad (8)$$

$$\lambda_i^* \ge 0, \quad g_i(x^*) \le 0, \quad \lambda_i^* g_i(x^*) = 0, \quad i = 1,...,m \qquad (9)$$

式(8)は大域的最適解がラグランジュ関数の極値として与えられることを意味し、式(9)は$\lambda i^*=0$または不等式制約が等式で満たされることを示す。 → 相補性条件

スパース性

相補性条件はSVMのスパース性と関連している。

↓

式(5)のラグランジュ関数γ_iに関する項を見ると、

相補性条件よりξ i=1-y^(i)f(x^(i))が成り立つ。

α i≠0となるサンプルのことをサポートベクトルと呼ぶ。

f(x)はSVの集合だけを使って、

$$f(x) = \sum_{x^{(i)} \in SV} \alpha_i k(x^{(i)}, x) \qquad (10)$$

と書くことが出来る。

続き

＊P89の図4.1(c)を参照

A点、もしくはそれより左側のサンプル

・・・サポートベクトル

A点の右にあり、なおかつ損失の値が0のサンプル

・・・カーネルには現れない

クラス識別が簡単な問題

↓

損失の値が正になることが少ない

↓

少ない計算量で計算できる。（スパース性がある）

サポートベクトルマシンの双対問題

凸二次計画問題の双対問題を考えると、
変数の数を減らした、より単純な凸二次計画問題となるので、
さらに計算が単純化される。
アルゴリズム:
[1]サンプルx^(i),…,x^(n)からのグラム行列K_{ij}を計算する。
[2]L_{dual}を$0 \leqq \gamma_i \leqq 1$の制約下で最適化する凸二次計画問題を解き、$\gamma=(\gamma_1,\ldots,\gamma_n)^T$を求める。
[3]識別関数

$$f(x)=\frac{1}{\lambda}\sum_{x^{(i)}\in SV}\gamma_i y^{(i)}k(x^{(i)},x) \qquad (11)$$

を求める。

＊ $$L_{dual}(\beta,\gamma)=\sum_{i=1}^{n}\gamma_i-\frac{1}{2\lambda}\sum_{i=1}^{n}\sum_{j=1}^{n}y^{(i)}y^{(j)}\gamma_i\gamma_j K_{ij} \qquad (12)$$

マージン最大化

マージン・・・

識別面とそれぞれのクラスのサンプル集合との最小距離

SVMは特徴空間において、マージンを最大にする識別関数として説明されることが多い。

マージン最大化：

f(x)=ω^Tφ(x)=0とサンプルの特徴ベクトルφ(x^(i))との距離は、

$$d_i = \frac{\left|\omega^T \phi(x^{(i)})\right|}{\|\omega\|} \qquad (13)$$

続き

ここでは、全てのサンプルが正しく分類されると仮定すると、

$$y^{(i)}\omega^{T}\phi(x^{(i)}) \geq 1 \qquad (14)$$

という条件が成り立つ。

つまりd_iの最小値は$1/\|\omega\|$で与えられるので、

マージンの最大化は

$$\min_{\omega} \|\omega\|^2 \qquad (15)$$

という最小化問題に帰着される。

ただし、全てのサンプルを正しく分類してしまうと、

過学習を起こす可能性がある。

制約(14)を破ることを許すが,その破った分を最小化関数のほうに加える　→　ソフトマージン

SVMの汎化能力

汎化能力・・・

習時に与えられた訓練データだけに対してだけでなく，未知の新たなデータに対するクラスラベルや関数値も正しく予測できる能力。

SVMの汎化能力：

・得られたマージン1/||ω||の値が大きいほど汎化能力が高い

・サポートベクトルの数が少ないほど汎化能力が高い