统计学习第六章

贾金柱

Nov 21, 2017

讲课老师

- 主讲人： 贾金柱
  - Email: jzjia@math.pku.edu.cn
- 助教：肖一君
  - Email: xiaoyijun1994@126.com

第六章 SVM

- 课程目标
  - linear SVM
  - Nonlinear SVM
  - SVM Regression

Linear SVM

- training data $(x_i, y_i), i = 1, 2, \dots, n$, $y_i \in \{-1, +1\}, x_i \in \mathbb{R}^p$.
- 定义一个 Classifier:

$$G(x) = sign(x^T\beta + \beta_0)$$

- 首先假设数据是完全线性可分的，即所有的数据点满足

$$y_i(x_i^T\beta + \beta_0) \geq 1.$$

- 那么最好的分割线应该使得两组数据尽可能地分开，即

$$\max \frac{2}{\|\beta\|_2}$$
$$s.t.\ y_i(x_i^T\beta + \beta_0) \geq 1.$$

Linear SVM

$$
\begin{aligned}
\min \frac{1}{2}\|\beta\|_2^2 \\
s.t. \quad y_i(x_i^T\beta + \beta_0) \geq 1.
\end{aligned}
$$

- 通常数据不会完美分开。因此， 需要引入松弛因子：

$$
\begin{aligned}
\min \frac{1}{2}\|\beta\|_2^2 \\
s.t. \quad y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \\
\xi_i \geq 0 \\
\sum_{i=1}^{n} \xi_i \leq constant
\end{aligned}
$$

Linear SVM

$$
\begin{aligned}
\min \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^{n}\xi_i \\
s.t.\quad y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \\
\xi_i \geq 0
\end{aligned}
$$

- 这是一个有线性约束的凸二次规划
- 可以通过QP（quadratic programming）解决
- 也可以通过Lagrange 对偶问题求解

SVM

- SVM 示意图

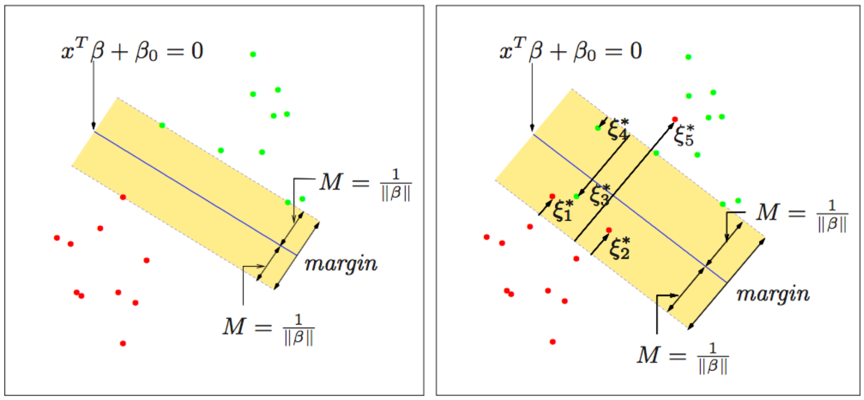

图: SVM 示意图

Lagrange Dual Problem

- Lagrange function:

$$L(\beta, \beta_0, \xi, \alpha, \mu) = \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^{n}\mu_i\xi_i$$

- 记SVM的最优值是p^*.
- 易知

$$\min_{\beta,\beta_0,\xi}\ \max_{\alpha>0,\mu>0} L(\beta, \beta_0, \alpha, \mu) = p^*$$

Lagrange Dual Problem

- 对偶问题 (Dual Problem)

$$\max_{\alpha>0,\mu>0} \min_{\beta,\beta_0,\xi} L(\beta, \beta_0, \xi, \alpha, \mu) := d^*$$

- 性质：

$$d^* \le p^*$$

求解对偶问题

分两步，

- 对于任意 α, μ， 求 L 关于 β, β_0, ξ 的最小值， 得到 $\theta(\alpha, \mu)$
- 然后求 $\theta(\alpha, \mu)$ 的最大值

$$\frac{\partial L}{\partial \beta} = 0 \Rightarrow \beta = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - \mu_i = 0$$

求解对偶问题

将 β 带入 $L(\beta, \ldots.)$ 并化简，得到

$$\theta(\alpha, \mu) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

最后，对偶问题可以正式描述如下

$$\begin{aligned} \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ s.t. \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \\ 0 \leq \alpha_i \leq C, i = 1, 2, \ldots, n \end{aligned}$$

SVM的解

一旦通过对偶问题，求得 α, 即可通过下式求得β,

$$\beta = \sum_{i=1}^{n} \alpha_i y_i x_i$$

KKT condition 告诉我们：

$$\alpha_i[y_i(x_i^T\beta + \beta_0)] = 0$$

由此，可以求得 $\beta_0 = y_k x_k^T \beta$, for some k with $\alpha_k \neq 0$

SVM的解

最后， classifier 就是 $sign(f(x))$,

$$\begin{aligned} f(x) &= x^T\hat{\beta} + \hat{\beta}_0 \\ &= x^T\hat{\beta} - y_k x_k^T\hat{\beta} \\ &= x^T\sum_{i=1}^{n}\alpha_i y_i x_i - y_k x_k^T\sum_{i=1}^{n}\alpha_i y_i x_i \\ &= \sum_{i=1}^{n}\alpha_i y_i\langle x, x_i\rangle - y_k\sum_{i=1}^{n}\alpha_i y_i\langle x_k, x_i\rangle \end{aligned}$$

SVM 总结

- 它是一个带有线性约束的二次规划问题
- 转到对偶问题，更容易求解
- 特别的，转到对偶问题，我们发现:
    - 参数 $\alpha_i, i = 1, 2, \ldots, n$ 仅依赖于样本点的内积 $\langle x_i, x_k \rangle$
    - classifier 也仅依赖于样本点的内积 $\langle x, x_i \rangle$, x 可以是测试样本或者观测样本

Kernel SVM

- 将数据映射到高维空间（比如泛函空间）

$$x_i :\to \phi(x_i, \cdot)$$

- 定义内积

$$\langle \phi(x_i, \cdot), \phi(x_k, \cdot) \rangle = K(x_i, x_k)$$

- 求解 SVM

$$\begin{aligned} \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ s.t. \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \\ 0 \le \alpha_i \le C, i = 1, 2, \dots, n \end{aligned}$$

Kernel SVM的解

最后， classifier 就是 $sign(f(x))$,

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \langle x, x_i \rangle - y_k \sum_{i=1}^{n} \alpha_i y_i \langle x_k, x_i \rangle$$

SVM 的另一种表达

$$\min_{\beta,\beta_0} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2} \|\beta\|_2^2$$

证明： 上式可以重新表达

$$\begin{aligned} \min_{\beta,\beta_0} \sum_{i=1}^{N} \xi_i + \frac{\lambda}{2} \|\beta\|_2^2 \\ s.t.\ 1 - y_i f(x_i) \le \xi_i \\ \xi_i \ \ge 0 \end{aligned}$$

Generalize SVM to Regression

One regression problem:

$$\min \frac{1}{2} \|\beta\|_2^2$$
$$s.t. \quad y_i - x_i^T \beta - \beta_0 \leq \epsilon$$
$$y_i - x_i^T \beta - \beta_0 \geq -\epsilon$$

SVM Regression

引入松弛因子：

$$
\begin{aligned}
&\min \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\
&s.t. \quad y_i - x_i^T\beta - \beta_0 \le \epsilon + \xi_i \\
&x_i^T\beta + \beta_0 - y_i \le \epsilon + \xi_i^* \\
&\xi_i, \xi_i^* \ge 0
\end{aligned}
$$

Dual Problem

定义Lagrangian

$$
\begin{aligned}
L := \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\
- \sum_{i=1}^{n}\alpha_i(\epsilon + \xi_i - y_i + x_i^T\beta + \beta_0) \\
- \sum_{i=1}^{n}\alpha_i^*(\epsilon + \xi_i^* + y_i - x_i^T\beta - \beta_0) \\
+ \sum_{i=1}^{n}(\eta_i\xi_i + \eta_i^*\xi_i^*)
\end{aligned}
$$

其中 $\alpha_i, \eta_i, \alpha_i^*, \eta_i^* \geq 0$

Dual Problem

$$\frac{\partial L}{\partial \beta} = 0 \Rightarrow \beta = \sum_{i=1}^{n} \left(\alpha_i^* - \alpha_i \right) x_i$$

$$\frac{\partial L}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{n} \left(\alpha_i^* - \alpha_i \right) = 0$$

$$\frac{\partial L}{\partial \xi^{(*)}} = 0 \Rightarrow C - \alpha_i^{(*)} - \eta_i^{(*)} = 0$$

Dual Problem

$$
\begin{aligned}
\max -\frac{1}{2} \sum_{i,j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j \rangle \\
- \epsilon \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) \\
s.t. \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0 \\
\alpha_i, \alpha_i^* \in [0, C].
\end{aligned}
$$

SVM Regression

$$\beta = \sum_{i=1}^{n} \left(\alpha_i^* - \alpha_i \right) x_i,$$

So,

$$f(x) = x^T \beta = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) \langle x, x_i \rangle + \beta_0$$

- Easily extend to non-linear SVM regression!

code

我推荐 http://svmlight.joachims.org