Support Vector Machines

422 группа

Tue 21 Mar 2017

Содержание

1 Maximium margin classifier

Пусть задана выборка $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$.

- Гиперплоскость
$$\mathcal{H} = \left\{\mathbf{x} : f(\mathbf{x}) = \boldsymbol{\beta}^\mathsf{T}\mathbf{x} + \beta_0 = 0\right\}, \quad \boldsymbol{\beta} = \begin{pmatrix}\beta_1\\ \vdots \\ \beta_p\end{pmatrix}.$$

- Нормаль к $\mathcal{H}$
$$\nabla f(\mathbf{x}) = \boldsymbol{\beta}.$$

- Расстояние со знаком от $\mathbf{x}$ до проекции $\mathbf{x}_0 \in \mathcal{H}$:
  - Пусть $\mathbf{x} \parallel \mathbf{x}'$; тогда $\langle \mathbf{x}, \mathbf{x}' \rangle = \pm \|\mathbf{x}\| \|\mathbf{x}'\|$
  - Поскольку $\nabla f(\mathbf{x}_0) \parallel (\mathbf{x} - \mathbf{x}_0)$ и $\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_0 = -\beta_0$,
  $$\|(\mathbf{x} - \mathbf{x}_0)\| = \frac{\langle \mathbf{x} - \mathbf{x}_0, \nabla f(\mathbf{x}_0)\rangle}{\|\nabla f(\mathbf{x}_0)\|} = \frac{1}{\|\boldsymbol{\beta}\|}\boldsymbol{\beta}^\mathsf{T}(\mathbf{x} - \mathbf{x}_0) = \frac{1}{\|\boldsymbol{\beta}\|}(\boldsymbol{\beta}^\mathsf{T}\mathbf{x} + \beta_0) = \frac{f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}.$$
  - Если $\|\boldsymbol{\beta}\| = 1$, то расстояние от $\mathcal{H}$ до точки $\mathbf{x}$ есть
  $$f(\mathbf{x}).$$

- Оптимизационная задача
$$\max_{\boldsymbol{\beta}, \beta_0} M, \quad \text{subject to } \|\boldsymbol{\beta}\| = 1,\ f(\mathbf{x}_i)y_i \geq M,\ i \in 1:N,$$
что то же
$$\max_{\boldsymbol{\beta}, \beta_0} M, \quad \text{subject to } \frac{f(\mathbf{x}_i)}{\|\boldsymbol{\beta}\|}y_i \geq M \iff f(\mathbf{x}_i)y_i \geq M\|\boldsymbol{\beta}\|,\ i \in 1:N,$$
и, т.к. $\forall \boldsymbol{\beta}, \beta_0$ удовлетворяющих неравенств $c \cdot \boldsymbol{\beta}, c \cdot \beta$ тоже удовлетворяет, положим $\|\boldsymbol{\beta}\| = 1/M$, тогда
$$\min_{\boldsymbol{\beta}, \beta_0} \|\boldsymbol{\beta}\|, \quad \text{subject to } f(\mathbf{x}_i)y_i \geq 1,\ i \in 1:N.$$

- Лагранжиан

$$L = \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_{i=1}^{N} \lambda_i (f(\mathbf{x}_i) y_i - 1).$$

- Решение

$$\begin{aligned} \boldsymbol{\beta} &= \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i && (1) \\ 0 &= \sum_{i=1}^{N} \lambda_i y_i. && (2) \end{aligned}$$

- При подстановке в L, двойственная

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \lambda_i \lambda_k y_i y_k \mathbf{x}_i^{\mathsf{T}} \mathbf{x}_k, \quad \text{subject to } \lambda_i \geq 0,\ \sum_{i=1}^{N} \lambda_i y_i = 0 \qquad (3)$$

  и еще условии

$$\lambda_i (y_i f(\mathbf{x}_i) - 1) = 0, \quad i \in 1:N. \qquad (4)$$

  Максимизируя её получают решение, которое должно удовлетворять ККТ условиям (1)–(4).

- Опорные вектора: из (4) видно, что если $\lambda_i \neq 0$, то $y_i f(\mathbf{x}_i) = 1$ и $\mathbf{x}_i$ лежит на границе; если $y_i f(\mathbf{x}_i) > 1$, то $\lambda_i = 0$. Поэтому в (1) участвуют только вектора на границе — *опорные*.

- Классификатор:

$$G(\mathbf{x}_0) = \operatorname{sign} \hat{f}(\mathbf{x}_0).$$

2 Support Vector Classifier (soft margin)

- Оптимизационная задача

$$\max_{\boldsymbol{\beta}, \beta_0, \boldsymbol{\epsilon}} M, \quad \text{subject to } \|\boldsymbol{\beta}\| = 1,\ f(\mathbf{x}_i) y_i \geq M(1 - \epsilon_i),\ \epsilon_i \geq 0\ i \in 1:N,\ \sum_{i=1}^{N} \epsilon_i \leq C_0.$$

  - C_0 — максимальное количество мисклассифицированных точек:
    - Если $\epsilon_i = 0$, то $\mathbf{x}_i$ лежит снаружи границы
    - Если $\epsilon_i \in (0, 1)$, то внутри с правильной стороны
    - Если $\epsilon_i > 1$, то с неправильной стороны
  - Как и прежде, эквивалентной задачей будет

$$\min_{\boldsymbol{\beta}, \beta_0} \|\boldsymbol{\beta}\|, \quad \text{subject to } f(\mathbf{x}_i) y_i \geq 1 - \epsilon_i,\ i \in 1:N,\ \sum_{i=1}^{N} \epsilon_i \leq C_0$$

  что то же,

$$\min_{\boldsymbol{\beta}, \beta_0} \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C \sum_{i=1}^{N} \epsilon_i, \quad \text{subject to } f(\mathbf{x}_i) y_i \geq 1 - \epsilon_i,\ i \in 1:N.$$

- Опорные вектора: теперь лежат как на границе, так и внутри.

3 Support Vector Machine (kernels)

Пусть $h_1, \ldots, h_M$ — базисные функции. Можно было бы спроецировать $\mathbf{h}(\mathbf{x}_i) = (h_1(\mathbf{x}_i), \ldots, h_M(\mathbf{x}_i))^\mathsf{T}$ и в результирующем пространстве построить классификатор. Расширение идеи — ядра. Ядро есть симметричная положительно определенная

$$K(\mathbf{x}, \mathbf{x}') = \left\langle \mathbf{h}(\mathbf{x}), \mathbf{h}(\mathbf{x}') \right\rangle.$$

Теорема (Мерцер)**.** *Пусть $K(\mathbf{x}, \mathbf{y})$ непрерывное, симметричное, положительно определенное ядро, $\mathbf{x}, \mathbf{y} \in \mathfrak{X}$ (полное, метрическое, сепарабельное). Тогда существует такое гильбертово пространство $\mathcal{H}$ и отображение $\mathbf{h} : \mathfrak{X} \to \mathcal{H}$ что*

$$K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{h}(\mathbf{x}), \mathbf{h}(\mathbf{y}) \rangle_{\mathcal{H}}.$$

Видно, что

$$\begin{aligned} \hat{f}(\mathbf{x}_0) &= \beta_0 + \mathbf{h}(\mathbf{x}_0)^\mathsf{T} \hat{\boldsymbol{\beta}} = \beta_0 + \mathbf{h}(\mathbf{x}_0)^\mathsf{T} \sum_{i=1}^N \lambda_i y_i \mathbf{h}(\mathbf{x}_i) = \beta_0 + \sum_{i=1}^N \lambda_i y_i \langle \mathbf{h}(\mathbf{x}_0), \mathbf{h}(\mathbf{x}_i) \rangle \\ &= \beta_0 + \sum_{i=1}^N \lambda_i y_i \langle \mathbf{h}(\mathbf{x}_0), \mathbf{h}(\mathbf{x}_i) \rangle = \beta_0 + \sum_{i=1}^N \lambda_i y_i K(\mathbf{x}_0, \mathbf{x}_i), \end{aligned}$$

так что в классификаторе $G(\mathbf{x}_0) = \operatorname{sign} \hat{f}(\mathbf{x}_0)$ участвуют только суммы ядер на опорных векторах (таких, что $0 < \lambda_i < C$).