SVM and Nearest Neighbor Classifiers

21307035 邓栩瀛

K-Nearest Neighbor Classifier

- 对于一个新的点，从训练数据中找出 k 个最接近的点
- 对 k 个点的标签进行分类
- 数据量大
- 合适的距离函数

距离函数

L1 distance

$$D(h_1, h_2) = \sum_{i=1}^{N} |h_1(i) - h_2(i)| \tag{1}$$

χ^2 distance

$$D(h_1, h_2) = \sum_{i=1}^{N} \frac{(h_1(i) - h_2(i))^2}{h_1(i) + h_2(i)} \tag{2}$$

Quadratic distance

$$D(h_1, h_2) = \sum_{i,j} A_{ij}(h_1(i) - h_2(i))^2 \tag{3}$$

Earth Mover's Distance

$$EMD(S_1, S_2) = \sum_{i,j} \frac{f_{ij}d(m_{1i}, m_{2j})}{f_{ij}} \tag{4}$$

SVM

Linear SVM

找到 w 和 b 使得 $\Phi(w) = ||w||^2 = w^T w$ 最小，对于 $\forall (x_i, y_i), i = 1..n, y_i(w^T x_i + b) \geq 1$

二次优化问题：构造对偶问题

寻找 $\alpha_1 \ldots \alpha_n$ 使得

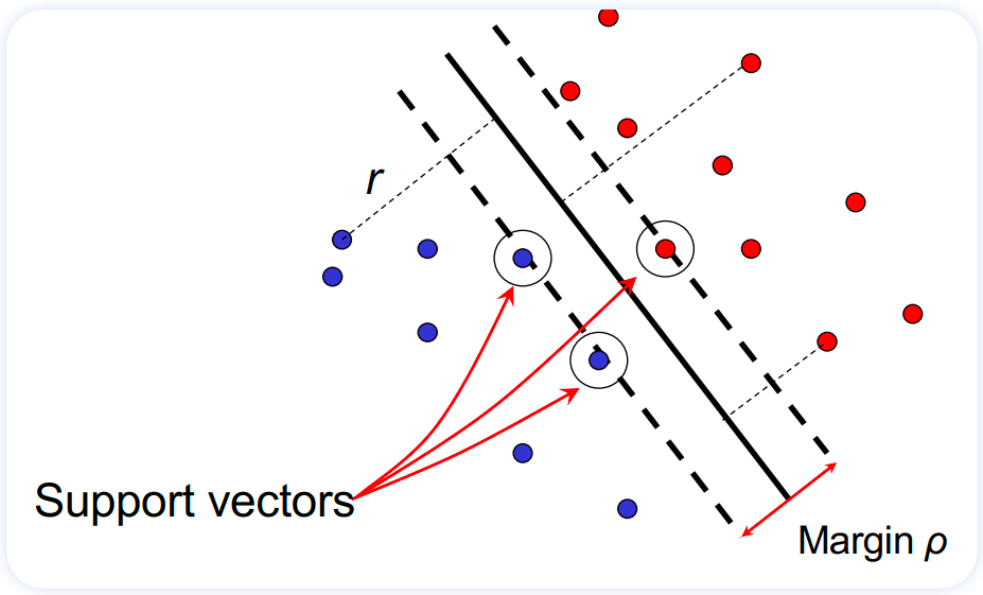$$Q(\alpha) = \sum \alpha_i - \frac{\sum \sum \alpha_i \alpha_j y_i y_j x_i^T x_j}{2} \tag{5}$$

最大化

其中

$$\sum \alpha_i y_i = 0$$
$$\alpha_i \geq 0, \forall \alpha_i \tag{6}$$

解得

$$w = \sum_i \alpha_i y_i x_i$$
$$b = y_i - w x_i \tag{7}$$

分类函数的决策边界

$$f(x) = wx + b = \sum_i \alpha_i y_i x_i x + b \tag{8}$$

Nonlinear SVMs

寻找 $\alpha_1 \dots \alpha_n$ 使得

$$Q(\alpha) = \sum \alpha_i - \frac{\sum \sum \alpha_i \alpha_j y_i y_j K(x_i, x_j)}{2} \tag{9}$$

最大化

其中

$$\sum \alpha_i y_i = 0$$
$$\alpha_i \geq 0, \forall \alpha_i \tag{10}$$

解得

$$f(x) = \sum_i \alpha_i y_i K(x_i, x_j) + b \tag{11}$$

Kernel Functions

Linear:

$$K(x_i, x_j) = x_i^T x_j \tag{12}$$

Polynomial of power p:

$$K(x_i, x_j) = (1 + x_i^T x_j)^p \tag{13}$$

Gaussian:

$$K(x_i, x_j) = e^{-\frac{||x_i - x_j||}{2\sigma^2}} \tag{14}$$

Histogram intersection kernel:

$$I(h_1, h_2) = \sum_{i=1}^{N} min(h_1(i), h_2(i)) \tag{15}$$

Generalized Gaussian kernel:

$$K(h_1, h_2) = exp(-\frac{D(h_1, h_2)^2}{A}) \tag{16}$$