Week2 Report

Chengze Cai

caicz18@mails.tsinghua.edu.cn

1 Introduction

Support Vector Machines is a kind of linear classifier that try to maximize the margin (i.e. the max distance from data point to classicfication boundary).

2 Basis

In formal, let (x_i, y_i) to be the data points and their labels. $x_i \in \mathbb{R}^n,\ y_i \in \{-1, 1\}$.
We are going to
$minimize_{w,b}\ \frac{1}{2}\|w\|^2$ s.t. $y_i(\langle w, x_i\rangle + b) \geq 1$.
Using the Lagrange Multiplier Method, we can solve the primal problem by solving the dual problem:
$minimize_{\{a_i\}} \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle$ s.t. $\sum_i y_i\alpha_i = 0,\ \alpha_i \geq 0$.

3 Linearly non-separable case

Note that $y_i(\langle w, x_i\rangle + b) \geq 1$ has no solution when data are linearly non-separable, we can add hinge loss function to allow misclassification:
$minimize_{w,b,\{\epsilon_i\}}\ \frac{1}{2}\|w\|^2 + C\sum_i \epsilon_i$ s.t. $y_i(\langle w, x_i\rangle + b) \geq 1 - \epsilon_i,\ \epsilon_i \geq 0$.
Corresponding dual problem is:
$minimize_{\{a_i\}} \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle$ s.t. $\sum_i y_i\alpha_i = 0,\ C \geq \alpha_i \geq 0$.

4 Kernel trick

When data are linearly non-separable, we can select a mapping ϕ to map the data into another space (usually with higher dimensions, even infinite dimensions), and replace the inner-product $\langle w, x_i\rangle, \langle x_i, x_j\rangle$ with $\langle\phi(w), \phi(x_i)\rangle, \langle\phi(x_i), \phi(x_j)\rangle$.
Let $k(x_i, x_j) = \langle\phi(x_i), \phi(x_j)\rangle$, if k can be calculated efficiently, we don't need to do the mapping ϕ explicitly, and k is called the kernel function.
Mercer's theorem: There exists a mapping ϕ and an expansion
$k(x, y) = \langle\phi(x), \phi(y)\rangle$

if and only if, for any g(x) s.t. $\int g(x)^2 dx$ is finite, then $\int k(x,y)g(x)g(y)dxdy \geq 0$.
Commonly used kernels:
$k(x,y) = \langle x,y \rangle^d$
$k(x,y) = (\langle x,y \rangle + 1)^d$
$k(x,y) = exp(-\frac{\|x-y\|^2}{2\sigma^2})$
$k(x,y) = tanh(a\langle x,y \rangle + b)$