贝叶斯方法

背后的深刻原因在于，现实世界本身就是不确定的，人类的观察能力有限，看到的只是世界的表象

> 具体参见**统计机器学习： 贝叶斯决策理论**的笔记

神经网络

> 具体参见**机器学习：神经网络**的笔记

SVM (支撑向量机)

Intuition

- 如何取寻找一个最优的判别面
- 引入Margin的概念，然后最大化这个Margin，记为M

SVM提取最优化问题方式一

hype-Plane

$$
\begin{aligned}
\boldsymbol{Plus\ Plane}: w^T x + b = +1 \\
\boldsymbol{Minus\ Plane}: w^T x + b = -1
\end{aligned}
$$

Maxinize Margin

Let $\boldsymbol{x}^+$ be an point on the plus-plane-point

Let $\boldsymbol{x}^-$ be an point on the minus-plane-point

$$
\begin{aligned}
M = |x^+ - x^-| \\
w^T x^+ + b = 1 \\
w^T x^- + b = -1
\end{aligned}
$$

Then We can get

$$
x^+ = x^- + \lambda w
$$

Then

$$
\lambda = \frac{2}{\|w\|^2}
$$

So we can get

$$M = |x^+ - x^-| = |\lambda w| = \lambda |w| = \frac{2}{||w||}$$

In order to maximize M we can Minimize

$$\frac{1}{2} \parallel w \parallel^2$$
$$s.t.\ \ w^T x_+ + b \geq 1, w^T x_- + b \leq -1$$

SVM提取最优化问题方式二

超平面方程

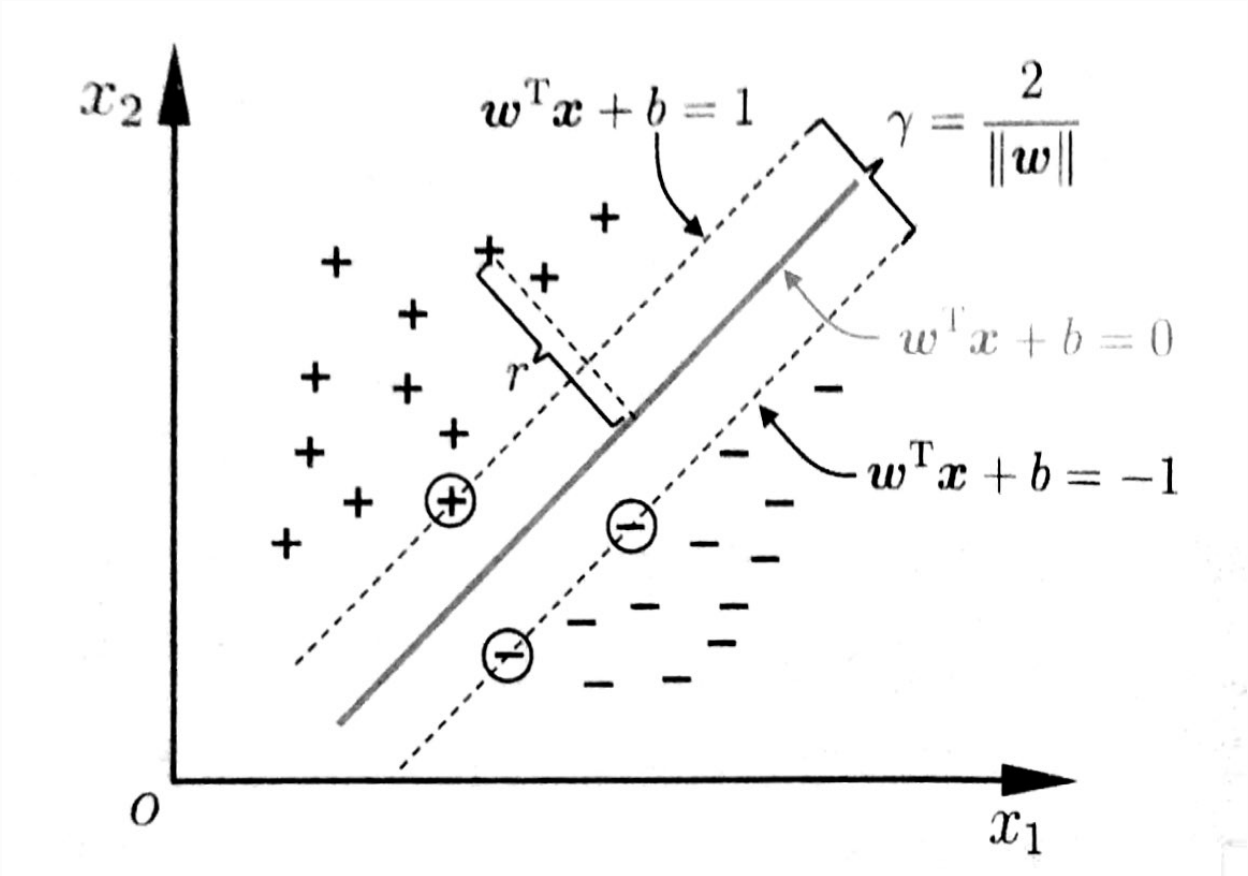
$$w^T x + b = 0$$

我们知道点到超平面对距离可按如下公式计算

$$r = \frac{|w^T x + b|}{||w||}$$

假设超平面能将数据分为两类，对于(x_i, y_i)有

$$\begin{cases} w^T x_i + b \geq +1 & y_i = +1 \\ w^T x_i + b \leq -1 & y_i = -1 \end{cases}$$



如上图，我们可以算出，两个超平面的距离为

$$r = \frac{2}{||w||}$$

为了最大化这个距离，我们的最优化问题可定义如下

$$min_{w,b} \ \frac{2}{||w||}$$
$$s.t. \ y_i(w^T x_i + b) \geq 1$$

转换后可变为

$$min_{w,b} \ \frac{1}{2}||w||^2$$
$$s.t. \ y_i(w^T x_i + b) \geq 1$$

SVM二次规划问题求解

问题：

$$min_{w,b} \ \frac{1}{2}||w||^2$$
$$s.t. \ 1 - y_i(w^T x_i + b) \leq 0$$

首先写出其拉格朗日函数

$$L(w,b,\alpha) = \frac{1}{2}||w||^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(w^T x_i + b))$$

对w_i, b求偏导使其为0，得到

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i$$
$$0 = \sum_{i=1}^{m} a_i y_i$$

带入整理可得对偶问题：

$$max_{\alpha} \ \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$$
$$s.t. \ \sum_{i=1}^{m} \alpha_i y_i = 0$$
$$\alpha_i \geq 0, i = 1, 2, \dots m$$

最终可得

$$f(x) = w^T x + b = \sum_{i=1}^{m} a_i y_i x_i^T x + b$$

KKT定理补充知识

> Optimization problem:
>
> $$\begin{gathered} Minimize \ \ f(x) \\ s.t. \ \ g_i(x) \leq 0, h_j(x) = 0 \\ i = 1, 2, \ldots, m \ ; \ j = 1, 2, \ldots l \end{gathered}$$
>
> Stationarity
>
> $$For\ maximizing\ f(x): \ \ \nabla f(x^*) = \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) + \sum_{j=1}^{l} \lambda_j \nabla h_j(x^*)$$
>
> $$For\ minimizing\ f(x): \ \ -\nabla f(x^*) = \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) + \sum_{j=1}^{l} \lambda_j \nabla h_j(x^*)$$
>
> Primal feasibility
>
> $$\begin{gathered} g_i(x^*) \leq 0, \ \text{for all} \ i = 1, \ldots, m \\ h_j(x^*) = 0, \ \text{for all} \ j = 1, \ldots, l \end{gathered}$$
>
> Dual feasibility
>
> $$\mu_i \geq 0, \ \text{for all} \ i = 1, \ldots, m$$
>
> Complementary slackness
>
> $$\mu_i g_i(x^*) = 0, \text{for all} \ i = 1, \ldots, m.$$

SVM二次规划问题求解需满足KKT，所以我们可以得到

$$\begin{cases} \alpha_i \geq 0 \\ y_i f(x_i) - 1 \geq 0 \\ \alpha_i (y_i f(x_i) - 1) = 0 \ (*) \end{cases}$$

由$(*)$可知,对于样本(x_i,y_i),总有 $\alpha_i=0\ or\ y_if(x_i)=1$，当 $\alpha_i=0$时，该样本不会在我们最终得到的函数$f(x)=\sum_{i=1}^m a_iy_ix_i^Tx+b$ 中出现，若$\alpha_i>0$，则必有$y_if(x_i)=1$，这些点恰好在两个最大间隔边界上，这些样本点称为支撑向量

Kernel

如果原始空间是有限维，即样本属性数有限，那么肯定存在一个高维特征空间使样本可分

常见的核函数

名称	表达式	参数
线性核	$k(x_i,x_j)=x_i^Tx_j$	
多项式核	$k(x_i,x_j)=(x_i^Tx_j)^d$	$d\geq1$为多项式次数
高斯核	$k(x_i,x_j)=exp^{-\frac{\|x_i-x_j\|^2}{2\sigma^2}}$	σ为高斯带的带宽
拉普拉斯核	$k(x_i,x_j)=exp^{-\frac{\|x_i-x_j\|}{\sigma}}$	$\sigma>0$
Sigmoid核	$k(x_i,x_j)=tanh(\beta x_i^Tx_j+\theta)$	双曲正切函数

Clustering

Partition method

Construct a partition of a database D of n objects into a set of k clusters s.t. minimize sum of squared distance

- K-mean
    - 迭代计算中心
    - 如何初始中心是个关键问题
        - 随机选择
        - 基于其他聚类算法(效果不一定好，但是效率高)的结果估算中心
    - 优势
        - 可扩展性强
        - 效率较高
        - 可实现局部最优 (退火算法和遗传算法等可以用找全局最优)
    - 缺点
        - 类数必须事先确定
        - 对噪音数据处理不好

        - 某些特殊分布无法划分（如凹型的）
- K-medoids (PAM)

有噪音和奇异点时，PAM比 k-means 鲁棒

Density based method