核函数

预备知识

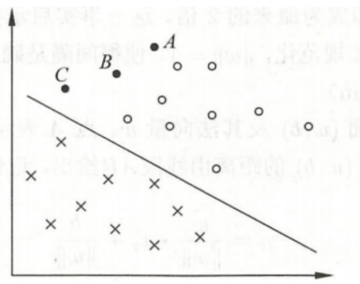
1. 线性可分支持向量机的分类超平面为：$\boldsymbol{w}^*\boldsymbol{x}+b^*=0$，其相应的分类决策函数为 $f(\boldsymbol{x})\ =sign(\boldsymbol{w}^*\boldsymbol{x}+b^*)$
2. 根据对偶问题可知：$\boldsymbol{w}^*=\sum_{i=1}^{N}\boldsymbol{\alpha}^*y_i\boldsymbol{x_i}, b^*=y_j-\sum_{i=1}^{N}y_i\boldsymbol{\alpha_i^*}(\boldsymbol{x_i}\boldsymbol{x_j})$,其中$y_j$对应的样本为支持向量。
3. 根据1,2可知，分类决策函数为：

$$f(\boldsymbol{x})=sign(\sum_{i=1}^{N}\boldsymbol{\alpha}^*y_i\langle\boldsymbol{x_i},\boldsymbol{x}\rangle+b) \tag{4.1}$$

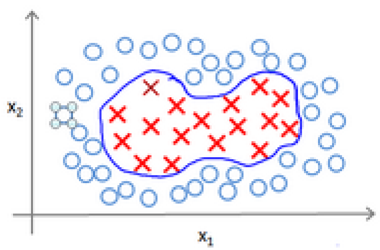
正文

线性可分与线性不可分：

以下图所示的二维特征空间中的分类问题为例，在这种情况下，找到能将两类数据正确划分并且间隔最大的直线即可，所以为线性可分：



但是，现实问题常常存入如下线性不可分的情况，如下图所示：



判定边界的模型可能是：$h_\theta(x)=\theta_0+\theta_1x_1+\theta_2x_2+\theta_3x_1x_2+\theta_4x_1^2+\theta_5x_2^2+...$的形式。用新的特征$f$来替换模型中的每一项，令：$f_1=x1,f2=x_2,f3=x_1x_2...$，也就是将低维特征映射到高维空间，得到：$h_\theta(x)=\theta_0+\theta_1f_1+...+\theta_nf_n$。

上面我们根据多项式经验，构造了新的特征$\boldsymbol{f}$。那么，在给定样本量m，原有特征数为k的情况下，新特征该怎么构造呢？ 在[课时97-核函数2](吴恩达机器学习 - 网易云课堂) 中，我们构造m+1个新特征。对于给定的样本$(x^{(i)}, y^{(i)})$（与《统计学习方法》的表达有些差异，第i个样本点表示为x_i, y_i）其新特征为：

$$f^{(i)}=\begin{bmatrix} f_0^{(i)}=1 \\ f_1^{(i)}=sim(x^{(i)}, l^{(1)}) \\ f_2^{(i)}=sim(x^{(i)}, l^{(2)}) \\ f_i^{(i)}=sim(x^{(i)}, l^{(i)})=e^0=1 \\ \vdots \\ f_m^{(i)}=sim(x^{(i)}, l^{(m)}) \end{bmatrix}$$

新特征为给定样本与训练集中所有样本的距离。

核技巧

由上可知，将低维特征映射到高维空间，则可以处理线性不可分的情况。但是将低维映射到高维$\boldsymbol{x}->\phi(\boldsymbol{x})$，存在一定计算量，如$f_1=x1, f2=x_2, f3=x_1x_2\ldots$。此外，$\langle\boldsymbol{x_i},\boldsymbol{x}\rangle->\langle\phi(\boldsymbol{x_i}),\phi(\boldsymbol{x})\rangle$计算量也会增大。根据式4.1可知，SVM决策函数最终是要得到$\langle\phi(\boldsymbol{x_i}),\phi(\boldsymbol{x})\rangle$，如果有函数$K(\boldsymbol{x_i},\boldsymbol{x})=\langle\phi(\boldsymbol{x_i}),\phi(\boldsymbol{x})\rangle$，这样就不用构造$\phi(\boldsymbol{x})$，也避免直接在高维空间中进行计算，而结果却是等价的！这时就可以用核函数，直接采用《统计学习方法》上的核函数定义：

定义 7.6（核函数） 设 $\mathcal{X}$ 是输入空间（欧氏空间 $\mathbf{R}^n$ 的子集或离散集合），又设 $\mathcal{H}$ 为特征空间（希尔伯特空间），如果存在一个从 $\mathcal{X}$ 到 $\mathcal{H}$ 的映射

$$\phi(x):\mathcal{X}\rightarrow\mathcal{H} \tag{7.65}$$

使得对所有 $x,z\in\mathcal{X}$，函数 $K(x,z)$ 满足条件

$$K(x,z)=\phi(x)\cdot\phi(z) \tag{7.66}$$

则称 $K(x,z)$ 为核函数，$\phi(x)$ 为映射函数，式中 $\phi(x)\cdot\phi(z)$ 为 $\phi(x)$ 和 $\phi(z)$ 的内积。

常见核函数

1. 多项式核函数（polynomial kernel function）

$$K(x,z)=(x\cdot z+1)^p \tag{7.88}$$

对应的支持向量机是一个 p 次多项式分类器。在此情形下，分类决策函数成为

$$f(x)=\operatorname{sign}\left(\sum_{i=1}^{N_s}a_i^*y_i(x_i\cdot x+1)^p+b^*\right) \tag{7.89}$$

2. 高斯核函数（Gaussian kernel function）

$$K(x,z)=\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right) \tag{7.90}$$

对应的支持向量机是高斯径向基函数（radial basis function）分类器。在此情形下，分类决策函数成为

$$f(x)=\operatorname{sign}\left(\sum_{i=1}^{N_s}a_i^*y_i\exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right)+b^*\right) \tag{7.91}$$

其中，高斯核函数通过泰勒展开可以将X映射到无穷维。如果σ很大，高次特征上的权重衰减得很快，所以实际上相当于一个低维的子空间；相反，如果σ很小，则可以将任意的数据映射为线性可分，但存在过拟合问题。不过，总的来说，通过调控参数，高斯核实际上具有相当高的灵活性，也是使用最广泛的核函数之一。下面总结下高斯核函数：

- σ较大时，可能会导致低方差，高偏差；
- σ较小时，可能会导致低偏差，高方差。