统计学习方法 （李航）

第一章 统计学习方法概论

1.2 监督学习

1.2.1 基本概念

1. 输入空间、特征空间

每个具体的输入是一个实列（instance）， 通常由特征向量，feature vector表示。特征向量存在的空间称为特征空间，feature space。有时假设输入空间与特征空间为不同的空间，将实例从输入空间映射到特征空间。模型实际上都是定义在特征空间上。

2. 联合概率分布

监督学习假设输入与输出的随机变量X和Y遵循联合分布，$P(X, Y)$. $P(X, Y)$表示分布函数或分布密度函数。

3. 假设空间

有输入到输出的映射，是监督学习的目的，这一映射由模型来表示。模型属于由输入空间到输出空间的映射的集合，这个集合就是假设空间，hypothesis space。

监督学习的模型可以是概率模型或非概率模型，有条件概率分布$P(Y|X)$或决策函数，decision function，$Y = f(X)$表示。

1.3 统计学习的三要素

1.3.1 模型

假设空间用$\mathcal{F}$表示

对于决策函数，

$$\mathcal{F}=\{f|Y=f_\theta(X),\theta\in\mathbb{R}^n\} \tag{1}$$

$\mathcal{F}$是参数向量θ决定的函数族

对于条件概率

$$\mathcal{F}=\{P|P_\theta(Y|X),\theta\in\mathbb{R}^n\} \tag{2}$$

1.3.2 策略

有了模型的假设空间，统计学习需要考虑的是按照什么样的准则学习或者选择最优的摸行。

1. 损失函数和风险函数

loss function $L(Y,f(X))$度量预测值$f(X)$与真实值Y的不一致程度，是$f(x)$和Y的非负实值函数。有以下几种常见的：

(1) 0-1 loss function

$$L(Y,f(x))=\begin{cases}1, & Y\neq f(x)\\ 0, & Y=f(x)\end{cases} \tag{3}$$

(2) quadratic loss function

$$L(Y,f(x))=(Y-f(x))^2 \tag{4}$$

(3) absolute loss function

$$L(Y,f(x))=|Y-f(x)| \tag{5}$$

(4) logarithmic loss function

$$L(Y,P(Y|X))=-logP(Y|X) \tag{6}$$

损失函数越小越好，由于摸行的输入、输出(X,Y)是遵循联合分布，所以损失函数的期望是

$$R_{exp}(f)=E_p[L(Y,f(X))]=\int_{\mathcal{X}\times\mathcal{Y}}L(y,f(x))P(x,y)dxdy \tag{7}$$

称为risk function 或者期望损失，expected loss。

$$R_{emp}(f)=\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i)) \tag{8}$$

称为经验风险，empirical risk，或经验损失，empirical loss。

2. 经验风险最小化与结构风险最小化

以经验风险最小化作为策略

$$\min_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i)) \tag{9}$$

比如极大似然估计，当模型是条件概率分布，损失函数是对数损失函数时，经验风险最小化等价与MLE

以结构风险最小化，是为了防止过拟合。结构风险最小化等价于正则化，regularization。结构风险在经验风险上加上表示复杂的的正则化项，regularizer或惩罚项，penalty term。

$$R_{srm}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \tag{10}$$

比如贝叶斯估计中的最大后验概率估计，maximum postexrior probability estimation，MAP就是结构风险最小化的例子。其中模型复杂度有模型的先验概率表示。

正则化项可以是参数向量的L_2范数：$\|w\|^2$，或者L_1范数：$\|w\|_1$

1.3.3 算法

1.6 泛化能力

1.6.1 泛化误差

generalization error, 假设学到的模型是$\hat{f}$，那么可表示为：

$$R_{exp}(\hat{f}) = E_p[L(T, \hat{f}(x))] = \int_{\mathcal{X}\times\mathcal{Y}} L(y, \hat{f}(x))P(x, y)dxdy \tag{11}$$

1.6.2 泛化误差上界

定理 1.1 泛化误差上界

对二类分类问题，当假设空间是有限个函数的集合$\mathcal{F} = \{f_1, f_2, \cdots, f_d\}$时，对任意一个函数$f \in \mathcal{F}$，至少以概率$1-\delta$, 以下不等式成立：

$$R(f) \le \hat{R}(f) + \varepsilon(d, N, \delta) \tag{12}$$

其中

$$\varepsilon(d, N, \delta) = \sqrt{\frac{1}{2N}(logd + log\frac{1}{\delta})} \tag{13}$$

该项是N的单调递减函数，是d的递增函数，表明$\mathcal{F}$中函数越多，其值越大

证明 用到**Hoeffding**不等式：

设$S_n = \sum_{i=1}^{n} X_i$是独立随机变量$X_1, X_2, \cdots, X_n$之和，$X_i \in [a_i, b_i]$，则对任意$t > 0$，以下不等式成立

$$P(S_n - ES_n \ge t) \le exp(\frac{-2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}) \tag{14}$$

$$P(ES_n - S_n \ge t) \le exp(\frac{-2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}) \tag{15}$$

对任意函数$f \in \mathcal{F}$，$\hat{R}(f)$是N个独立的随机变量的$L(Y, f(X))$的样本均值，$R(f)$是期望值，对二类分类问题，$[a_i, b_i] = [0, 1]$，那么由不等式(15)，对$\varepsilon > 0$，以下不等式成立：

$$P(R(f) - \hat{R}(f) \ge \varepsilon) \le exp(-2n\varepsilon^2) \tag{16}$$

对于有限集合$\mathcal{F} = \{f_1, f_2, \cdots, f_d\}$，有

$$
\begin{aligned}
P(\exists f \in \mathcal{F}: R(f) - \hat{R}(f) \geq \varepsilon) \\
&= P(\bigcup_{f \in \mathcal{F}} \{R(f) - \hat{R}(f) \geq \varepsilon\}) \\
&\leq \sum_{f \in \mathcal{F}} P(R(f) - \hat{R}(f) \geq \varepsilon) \\
&\leq dexp(-2n\varepsilon^2)
\end{aligned}
$$

这等价为，$\forall f \in \mathcal{F}$

$$
P(R(f) - \hat{R}(f) < \varepsilon) \geq 1 - dexp(-2n\varepsilon^2) \tag{17}
$$

另$\delta = dexp(-2n\varepsilon^2)$，有

$$
P(R(f) \leq \hat{R}(f) + \varepsilon) \geq 1 - \delta \tag{18}
$$

1.7 生成模型与判别模型

监督学习方法可以分为生成方法，generative approach，和判别方法，discriminative approach。

生成方法由数据学习**联合概率分布**$P(X, Y)$，然后再求出条件概率分布$P(Y|X)$，作为预测模型的，即为生成模型。

$$
P(Y|X) = \frac{P(Y, X)}{P(X)} \tag{19}
$$

之所以称为生成方法，在于模型表示了给定输入X，产出Y的关系。***注意，学习对象是概率联合分布***

典型的生成模型有：**朴素贝叶斯，隐马尔可夫**

判别方法由数据直接学习决策函数$Y = f(X)$，或者条件概率分布$P(Y|X)$。判别方法关心的是给定输入，应该预测什么样的Y。

典型的判别模型有：**k邻近、感知机、决策树、logistic、最熵、SVM、boosting、条件随机场**等

生成方法的特点：

- 生成方法的学习收敛速度更快，即当样本容量增加时，学到的模型可以更快地收敛于真实模型；
- 生成方法可以还原出联合概率分布，而判别方法不能；
- 当存在隐变量时，可以生成方法，而其他方法则不行；

判别方法的特点：

- 直接学习条件概率或者决策函数，直接面对预测，往往学习的准确率更高
- 可以对数据进行各种程度上的抽象、定义特征并且使用，因此可以简化学习问题

1.8 分类问题

精确率，precision：

$$
P = \frac{TP}{TP + FP} \tag{20}
$$

召回率，recall：

$$
R = \frac{TP}{TP + FN} \tag{21}
$$

此外还有F_1值，是精确率和召回率的调和均值，即

$$
\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R} \tag{22}
$$

第二章 感知机

感知机，perceptron，是二类分类的线性模型，由Rosenblatt于1957年提出，是神经网络和SVM的基础

2.1 感知机模型

定义 $\mathcal{X} \subseteq \mathbf{R}^n$，是输入空间，$\mathcal{Y} - \{+1, -1\}$，是输出空间。由输入到输出如下函数

$$f(x) = sign(w \cdot x + b) \tag{23}$$

$w \in \mathbf{R}^n$是权重，weight，$b \in \mathbf{R}$叫偏置，bias。

几何解释：

$$w \cdot x + b = 0 \tag{24}$$

对应于特征空间中的一个超平面，w是法向量，b是对应截距。***参考平面截距式，已经平面参数方程***

2.2 感知机学习策略

存在超平面使得数据集的正、负实例点完全正确的划分到超平面两侧，则数据集线性可分。

损失函数的一个自然选择是误分类点的总数。但这样的损失函数不是参数的连续可导，不易优化。

感知机选择误分类点到超平面的总距离。$\mathbf{R}^n$中任一点x_0到超平面的距离：

$$\frac{1}{\|w\|}|w \cdot x_0 + b| \tag{25}$$

对于误分类点来说：

$$-y_i(w \cdot x_i + b) > 0 \tag{26}$$

因此，任意误分类点到超平面的距离为：

$$-\frac{1}{\|w\|}y_i|w \cdot x_i + b| \tag{27}$$

从而感知机的损失函数形式为：

$$L(w, b) = -\sum_{x_i \in M} y_i(w \cdot x_i + b) \tag{28}$$

其中M为误分类点集合。

2.3 感知机的算法

目标函数：

$$\min_{w,b} L(w, b) = -\sum_{x_i \in M} y_i(w \cdot x_i + b) \tag{30}$$