统计学习方法概论

1.监督学习

- **输入空间，特征空间，输出空间**

  对输入空间的数据进行处理，添加修补，以及构造得到特征空间。

- **联合概率分布**

  监督学习假设输入输出满足一个联合概率分布，然而学习系统对这个联合概率分布未知。并且我们假设他存在，也就是说统计学习假设数据存在一定的统计规律，我们利用学习系统来逼近这个统计规律。

  - 联合概率

    > 联合概率指的是包含多个条件且**所有条件同时成立**的概率，记作P(X=a,Y=b)或P(a,b)，有的书上也习惯记作P(ab).

  - 边缘概率

    > 边缘概率是与联合概率对应的，P(X=a)或P(Y=b)，这类仅与单个随机变量有关的概率称为边缘概率

    - 联合概率和边缘概率的关系

$$P(X=a)=\sum^{b}P(X=a,Y=b)$$
$$P(Y=b)=\sum^{a}P(X=a,Y=b)$$

  - 条件概率

    > 条件概率表示在条件Y=b成立的情况下，X=a的概率，记作P(X=a|Y=b)或P(a|b)，它具有如下性质：

> $$\sum^{a}P(X=a|Y=b)=1$$

  - 联合概率和边缘概率与条件概率之间的关系（利用面积求解）

$$P(X=a|Y=b)=\frac{p(X=a,Y=b)}{P(Y=b)}$$

- **假设空间**

  输入空间到输出空间的所有映射的集合，这个集合我们将他称作假设空间。一般模型为概率模型或非概率模型，由条件概率分布$P(Y|X)$ 或决策函数$Y=f(X)$ 表示。

2.统计学习三要素

2.1 模型

	假设空间 $\mathcal{F}$	输入空间 $\mathcal{X}$	输出空间 $\mathcal{Y}$	参数空间
决策函数	$\mathcal{F}=\{f\|Y=f_\theta(X),\theta\in\mathbf{R^n}\}$	变量	变量	$\mathbf{R^n}$
条件概率分布	$\mathcal{F}=\{P\|P_\theta(Y\|X),\theta\in\mathbf{R^n}\}$	随机变量	随机变量	$\mathbf{R^n}$

2.2 策略

- **期望风险**

无论是对于概率模型或者是非概率模型，都存在一个联合概率分布$P(X,Y)$，于是我们可以得到损失函数的期望，他是作为整体的期望:

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

我们将其称为风险函数或期望损失。由于无法求出一个过程的联合分布，我们也无法算出期望损失，

- **经验风险**

  在整体的一批样本中:

  $$T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$$

  模型关于训练数据集的平均损失称为经验风险或经验损失：

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

根大数定律可知，当样本容量足够大时，样本均值等于总体期望。此时，经验风险趋近于期望风险。

- **经验风险最小化和结构风险最小化**

  一般情况下，我们认为经验风险最小化能够保证模型学习的效果。然而，当样本容量较小时，经验风险最小化会产生过拟合的现象。结构风险最小化是为了防止过拟合而提出的策略，结构风险最小化等价于正则化。

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

  其中$J(f)$代表模型的复杂度，模型越复杂，$J(f)$越大，反之越小。λ代表系数。

  > 1.极大似然估计是经验风险最小化的一个例子 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于极大似然估计
  >
  > 2.贝叶斯估计中的最大后验概率估计是结构风险最小化的一个例子 当模型是条件概率分布，损失函数是对数损失函数，模型复杂度由模型的先验概率表示时，结构风险最小化等价于最大后验概率估计

  - **正则化**

    结构风险最小化的策略的实现:

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

    其中第一项为经验风险，第二项为正则化项。正则化项可以取不同的形式。

    - 正则化项为L_2范数,此时损失函数是平方损失:

$$L(W)=\frac{1}{N}\sum_{i=1}^{N}(f(x_i;w)-y_i)^2+\frac{\lambda}{2}||w||^2$$

$||w||$表示参数向量的L_2范数。$||X||_2=\sqrt{\sum_{i=1}^{N}x_i^2}$。

- 正则化项为L_1范数：

$$L(W)=\frac{1}{N}\sum_{i=1}^{N}(f(x_i;w)-y_i)^2+\lambda||w||_1$$

其中，$||X||_1=\sum_{i=1}^{N}|x_i|$。

正则化符合**奥卡姆剃刀原理**，我们需要选择结构更简单的模型。

2.3 算法

学习模型的具体方法

3.训练误差和测试误差

假设我们学习到的模型$Y=\hat{f}(X)$.

- **训练误差**

  模型Y关于训练数据集的平均损失：

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

- **测试误差**

  模型Y关于测试数据集的平均损失：

$$e_{test}=\frac{1}{N'}\sum_{i=1}^{N'}L(y_i,\hat{f}(x_i))$$

4.泛化能力

4.1 泛化误差

泛化误差的定义，如果学到的模型为$\hat{f}$，那门使用这个模型对未知数据的误差则为泛化误差：

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

观察上式发现泛化误差为数据集的期望风险。

4.2 泛化误差上界

一般我们通过分析学习方法的泛化误差上界来比较他们之间的优劣。已知一个二分类任务，假设空间是有限集合$\mathcal{F}=\{f_1,f_2,f_3,...,f_d\}$,d为函数个数，则$f$的期望风险和经验风险分别为:

$$R(f) = E[L(Y, f(X))]$$
$$\hat{R}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i))$$

经验风险最小化:

$$f_N = arg \min_{f \in \mathcal{F}} \hat{R}(f)$$

于是我们得出f_N的泛化能力:

$$R(f_N) = E[L(Y, f_N(X))]$$

下面我们讨论函数的泛化误差上界。

泛 化 误 差 上 界 :在二分类问题中，假设空间中的函数集合$\mathcal{F} = \{f_1, f_2, f_3, \ldots, f_N\}$,对于任意的$f \in \mathcal{F}$,至少存在概率$1-\delta$,使得:

$$R(f) \leq \hat{R}(f) + \varepsilon(d, N, \delta)$$

其中:

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

结论:

- 当样本容量N增加，泛化误差上界越小。
- 假设空间d的容量越大，模型难以学习，泛化误差上界越大。
- 训练误差越小，泛化误差也越小。

5.分类指标

- TP：将正类预测为正类的个数，在预测结果中分正确的个数。
- FN：将正类预测为负类的个数，在真实数据中出现，但是未在预测结果中出现的个数。
- FP：将负类预测为正类的个数，在预测结果中出现，但是未在真实数据中出现的个数。
- TN：将负类预测为负类的个数。

  精确率(precision):

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

  召回率(recall):

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

  F_1:

$$F_1 = \frac{2TP}{2TP + FP + FN}$$