统计学习方法概论

1.2监督学习

- **统计学习**：监督学习、非监督学习、本监督学习和强化学习；
- **监督学习**：学习一个模型，使得模型能够对任意输入对其对应的输出做一个好的预测；

1.2.1 基本概念

- **输入空间与输出空间**：将输入与输出的所有可能取值的集合叫做输入空间（input space）和输出空间（output space）；可以离散可以连续；可以是同一个空间也可以是不同空间，但是通常输出空间远小于输入空间；
- **特征空间**：每一个具体的输入是一个实例（instance），通常由特征向量（feature vector）表示，所有特征向量存在的空间叫做特征空间（feature space）；
- 输入输出变量用大写字母表示X，Y；输入输出标量取值用小写字母表示x，y；输入实例的特征向量记作

$$x = (x^{(1)}, x^{(2)}, \dots, x^{(n)})^T \tag{1}$$

$x^{(i)}$表示x的第i个特征；$x^{(i)}$与x_i不同，x_i表示多个输入实例的第i个；

- 训练数据由输入与输出组成，训练集表示为：

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\} \tag{2}$$

- **回归问题**：输入变量和输出变量均为连续变量的预测问题；
- **分类问题**：输出变量为有限个离散变量的预测问题；
- **标注问题**：输入变量和输出变量均为变量序列的预测问题；
- **联合概率分布**：监督学习假设输入和输出遵循联合概率分布$P(X, Y)$，学习过程中假定其存在，训练数据与测试数据被视为依联合概率分布$P(X, Y)$独立同分布产生的；

- **假设空间**：模型属于输入空间到输出空间的映射集合，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者非概率模型，由条件概率分布$P(Y|X)$或者决策函数$Y=f(X)$表示;

1.3 统计学习的三要素

> 方法=模型+策略+算法

1.3.1 模型

- **假设空间**：用$\mathcal{F}$表示，是指决策函数的集合，且通常是由一个参数向量决定的函数族

$$\mathcal{F}=\{f|Y=f_\theta(X),\theta\in\mathbb{R}^n\} \tag{3}$$

参数向量θ取值于n维欧式空间$\mathbb{R}^n$，成为参数空间（parameter space）；概率模型类似

1.3.2 策略

有了模型假设空间，需要考虑按照什么准则学习或者选择最优的模型。定义损失函数和风险函数的概念，*损失函数度量一次预测的好坏*，*风险函数度量平均意义下的好坏*:

- **损失函数**（loss function）或者叫**代价函数**（cost function），用来度量预测错误的程度，损失函数是$f(X)$和Y的*非负实值函数*，$L(Y,f(X))$
  - 0-1损失函数:

$$L(Y,f(X))=\begin{cases}1,Y\neq f(X)\\0,Y=f(X)\end{cases}$$

  - 平方损失函数:

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

  - 绝对损失函数:

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

  - 对数损失函数，或对数似然损失函数:

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

- 损失值越小，模型越好，由于输入输出符合联合概率分布$P(X,Y)$，所以损失函数的期望是：

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

这是理论上模型f在联合概率分布$P(X,Y)$平均意义下的损失，称为**风险函数**（risk function）或者**期望损失**（expected loss），学习的目标就是选择期望风险最小模型。

- **经验风险**（empirical risk）或**经验损失**（empirical loss）：给定训练集T，模型$f(X)$关于训练数据的平均损失：

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

- 期望风险是模型关于联合分布的期望损失，经验风险是模型关于训练样本集的平均损失，根据**大数定律**，当样本容量N区域无穷时，$R_{emp}(f)$趋于$R_{exp}(f)$，可以使用经验风险估计期望风险；
- 由于现实中训练样本有限，使用经验风险不理想，需要进行矫正，通常有两个策略：**经验风险最小化**和**结构风险最小化**；
- **经验风险最小化**（empirical risk minimization，ERM）策略认为经验风险最小的模型是最优模型：

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

极大似然估计（maximum likelihood estimation）就是经验风险最小化例子，当模型是条件概率分布，损失函数是对数损失函数时，ERM等价于MLE；但是当样本很小时，会产生**过拟合现象**，学习效果未必很好；

- **结构风险最小化**（structural risk minimization，SRM）为了防止过拟合而提出，等价于正则化（regularization），在经验风险上加上*表示模型复杂度*的正则化项（regulaizer）或惩罚项（penalty term），其策略是最小化结构风险：

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

其中$J(X)$表示模型复杂度，是定义在假设空间上的泛函，模型越复杂，复杂度越大；反之，模型越简单，复杂度越小；$\lambda \geq 0$是系数，用于权衡经验风险和模型复杂度；结构风险小需要经验风险和模型复杂度同时小；

贝叶斯估计中的**最大后验概率估计**（maximum posterior probability estimation，MAP）就是结构风险最小化的例子，当模型是条件概率分布时，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，SRM等价于MAP；

1.3.3 算法

算法是指学习模型的具体计算方法。统计学习的问题归结为最优化问题，统计学习的算法成为求解最优化问题的算法。

1.4 模型评估与模型选择

1.4.1 训练误差与测试误差

- **训练误差**：假设学习到的模型为$\hat{f}\,(X)$，训练误差是模型关于训练集的平均损失：

$$R_{emp}(\hat{f}\,) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}\,(x_i)) \tag{11}$$

- **测试误差**：是模型关于测试集的平均损失：

$$e_{test} = \frac{1}{N'}\sum_{i=1}^{N'} L(y_i, \hat{f}\,(x_i)) \tag{12}$$

1.4.2 过拟合与模型选择

- **模型选择**（model selection）：当假设空间中含有不同复杂度（不同参数个数）模型时，希望选择或学习一个合适模型，如果空间中存在“真”模型，那么选择的模型应该逼近真模型（参数个数相同，参数向量相近）；
- **过拟合**（over-fitting）：如果一味追求对训练数据的预测能力，所选模型的复杂度往往比真模型更高，出现过拟合现象（所选模型参数过多，对已知数据预测得很好，对未知数据预测很差）；
- 两种模型选择方法：正则化与交叉验证

1.5 正则化与交叉验证

1.5.1 正则化

- 正则化项一般是**模型复杂度的单调递增函数**，模型越复杂，正则化值就越大；正则化项可以是模型参数向量的范数；
- 在回归问题中，损失函数是平方损失，正则化项可以是参数向量的L_2范数：

$$L(w)=\frac{1}{N}\sum_{i=1}^{N}(y_i,\hat{f}\left(x_i;w\right))^2+\frac{\lambda}{2}\|w\|_2^2 \tag{13}$$

其中$\|w\|_2$表示参数向量w的L_2范数；

- 正则化项也可以是参数向量的L_1范数：

$$L(w)=\frac{1}{N}\sum_{i=1}^{N}(y_i,\hat{f}\left(x_i;w\right))^2+\frac{\lambda}{2}\|w\|_1^2 \tag{14}$$

其中$\|w\|_1$表示参数向量w的L_1范数；

- 正则化符合**奥卡姆剃刀**（Occam's razor）原理：在所有可能选择的模型中，能够很好地*解释已知的数据*并且*十分简单*才是最好的模型；
- 从贝叶斯角度来看，*正则化项对应于模型的先验概率*，可以假设复杂的模型有较小的先验概率，简单的有较大的先验概率；

1.5.2 交叉验证

模型选择最简单方法是把数据集切分成3部分，训练集（training set）、验证集（validation set）和测试集（test set）。并在学习中选择对验证集有最小预测误差的模型。

- 测试集用于训练模型；
- 验证集用于模型选择；
- 测试集用于最终对学习方法的评估；

当数据不充足时，可以采用**交叉验证**选择更好的模型：

- **简单交叉验证**：首先把数据随机切成两部分，一个作为训练集，一个作为测试集；然后用测试集在各种条件下训练模型，在测试集上评价各个模型的测试误

差，选出最小的；

- **S折交叉验证**（S-fold cross validation）：首先随机将数据集切成S个互不相交且大小相同的子集；利用S-1个子集的数据训练模型，利用余下的测试模型；将该过程的S种可能重复，选择平均测试误差最小的；
- **留一交叉验证**（leave-one-out cross validation）：即S折交叉验证的特殊情形$S=N$，往往在数据缺乏的情况下使用。

1.6 泛化能力

1.6.1 泛化误差

- **泛化误差**（generalization ability）：指该方法学习到模型对未知数据的预测能力，现实中采用最多的方法是通过*测试误差*来评价方法的泛化能力；由于测试集是有限的，得到的评价结果不一定可靠；
- **定义**：如果学习到的模型为$\hat{f}$，用这个模型对未知数据预测的误差即为泛化误差：

$$R_{exp}(\hat{f})=\mathbb{E}_P[L(Y,\hat{f}(X))]=\int_{\mathcal{X}\times\mathcal{Y}}L(y,\hat{f}(x))P(x,y)dxdy \tag{15}$$

如果该模型的泛化误差更小，说明该方法更有效；泛化误差就是模型的期望风险；

1.6.2 泛化误差上界

- **泛化误差上界**（generalization error bound）：学习方法的*泛化能力分析*往往通过研究泛化误差的概率上界进行（即通过比较泛化误差上界比较方法的优劣）；
- **性质**：是样本容量的函数，当N增加时，泛化上界趋于0；是假设空间容量的函数，假设空间容量越大，模型就越难学，泛化误差上界就越大；

1.7 生成模型与判别模型

- 监督学习方法又可以分为**生成方法**（generative approach）和**判别方法**（discriminative approach），学习出的模型叫**生成模型**（generative model）和**判别模型**（discriminative model）；

- 生成方法由数据学习联合概率分布$P(X,Y)$，然后求出条件概率分布$P(Y|X)$作为预测模型，即生成模型：

$$P(Y|X)=\frac{P(X,Y)}{P(X)} \tag{16}$$

之所以叫生成方法，是因为模型*表示了给定输入*X*产生输出*Y*的生成关系*（如朴素贝叶斯和隐马尔科夫模型）；

- 判别方法由数据直接学习决策函数$f(X)$或者条件概率分布$P(Y|X)$作为预测模型（k近邻、感知机、决策树、逻辑回归、最大熵、支持向量机、提升方法和条件随机场模型）；
- **生成方法特点**：可以*还原出联合概率分布*$P(X,Y)$；学习*收敛速度快*，当样本容量增加时，学习的模型可以更快收敛与真实模型；当*存在隐变量时*，仍可以使用，但判别方法不能；
- **判别方法特点**：学习结果直接面对预测，往往学习*准确率更高*；由于直接学习$f(X)$或$P(Y|X)$，可以对数据进行各种程度*抽象、定义特征并使用特征*，即简化学习问题；

1.8 分类问题

- **分类问题**：输出变量Y可以取有限个离散值的预测问题，输入变量X可以是离散的，也可以是连续的；学习出的分类模型成为分类器（classifier）；
- 性能评价指标一般是**分类准确率**（accuracy）：对于给定的测试数据集，分类器正确分类样本数与总样本数之比；
- **二分类问题**常用评价指标是**精确率**（precision）和**召回率**（recall），把关注的类作为正类，其他作为负类：
  - 真阳性（True Positive，TP）：将正类预测为正类数；
  - 假阴性（False Negative，FN）：将正类预测为负类数；
  - 假阳性（False Positive，FP）：将负类预测为正类数；
  - 真阴性（True Negative，TN）：将负类预测为负类数；
  - 精确率：

$$\text{precision}=\frac{TP}{TP+FP} \tag{17}$$

  - 召回率：

$$\text{recall} = \frac{TP}{TP + FN} \tag{18}$$

- F_1值，是精确率与召回率的调和均值：

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{19}$$

精确率和召回率都高时，F_1值也高；

1.9 标注问题

- **标注问题**（tagging）：可以认为标注问题是分类问题的一个推广，是更复杂的结构预测（structure prediction）问题的简单形式；其输入是一个观测序列，*输出是一个标记序列或者状态序列*；
- 给定一个训练数据集：

$$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \tag{20}$$

$x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)})^T, i = 1, 2, \ldots, N$是输入观测序列，而$y_i = (y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(n)})^T$是相应的输出标记序列，希望学习出一个模型，表示为条件概率分布：

$$P(Y^{(1)}, Y^{(2)}, \ldots, Y^{(n)} | X^{(1)}, X^{(2)}, \ldots, X^{(n)}) \tag{21}$$

- 评价标注模型的指标与评价分类模型的指标一样；

1.10 回归问题

- **回归问题**（regression）：当输入标量的值发生变化时，输出变量的值随之发生改变；回归问题的学习等价于函数拟合（选择一条函数曲线使其很好地拟合已知数据和预测未知数据）；
- 变量个数：一元回归和多元回归；输入输出变量关系：线性回归和非线性回归；
- 回归学习常用损失函数是*平方损失函数*，此时可用最小二乘法求解；