LECTURE 1: 统计学习概论

1. 统计学习

统计学习（statistical learning）是基于数据构建概率统计模型并运用模型对数据进行预测及分析的一门学科。

对象：数据（data）。什么是数据？所有可以被记录的都是数据，例如数字、文字、图像、音频等。

目的：基于数据构建概率统计模型，实现对数据进行预测及分析。

统计学习方法： (1) 有监督学习 (supervised learning)； (2) 无监督学习 (unsupervised learning); (3) 强化学习 (reinforcement learning) 等组成。

学习步骤（概括）： 从训练数据（training data）出发，假设要学习的模型属于某个函数的集合，称为假设空间（hypothesis space）； 应用某个评价准则（evaluation criterion），从假设空间中选择一个最优模型，使它对已知的数据及未知的测试数据（test data）在给定的准则下具有最好的预测性能。

举例：线性回归模型

三要素：模型（model）、策略（strategy）和算法（algorithm）。

2. 统计学习的分类

2.1. 基本分类

1. 监督学习

从标注数据中学习预测模型的机器学习问题。

(1) 输入空间、特征空间和输出空间

将输入与输出所有可能取值的集合分别称为输入空间（input space）与输出空间（output space）。【一般来说输出空间一般小于输入空间】

每个具体的输入是一个实例（instance），通常由特征向量（feature vector）表示。输入实例x的特征向量写作：

$$x = (x^{(1)}, x^{(2)}, \cdots, x^{(p)})^\top \in \mathbb{R}^p. \tag{2.1}$$

【一般用小写表示】

在监督学习中，将输入与输出看做是定义在输入（或输出）空间上的随机变量的取值。 输入与输出变量用大写字母表示，习惯上输入变量写作X，输出变量写作Y。 输入与输出变量的取值一般写作x与y。 一般将第i个输入实例记作x_i.

监督学习从训练数据(training data)集合中学习模型，对测试数据进行预测。 一般训练数据表示为： $\{(x_1, y_1), \cdots, (x_N, y_N)\}$ 输入与输出对又称为样本（sample）或样本点。

输入变量X与输出变量Y可以有不同的类型： 输出变量为连续型变量时称为回归问题； 输出变量为离散变量时称为分类问题。

(2) 联合概率分布

假设输入与输出的随机变量X与Y 遵循联合概率分布 $P(X,Y)$, $P(X,Y)$ 表示分布函数或者分布密度函数。 训练数据与测试数据一般可以看做是依据概率分布$P(X,Y)$独立同分布产生的。

(3) 假设空间

监督学习目的是寻找一个由输入到输出的映射，这一映射由模型来表示。 模型属于由输入空间到输出空间映射的集合，这个集合称为假设空间（hypothesis space）。

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

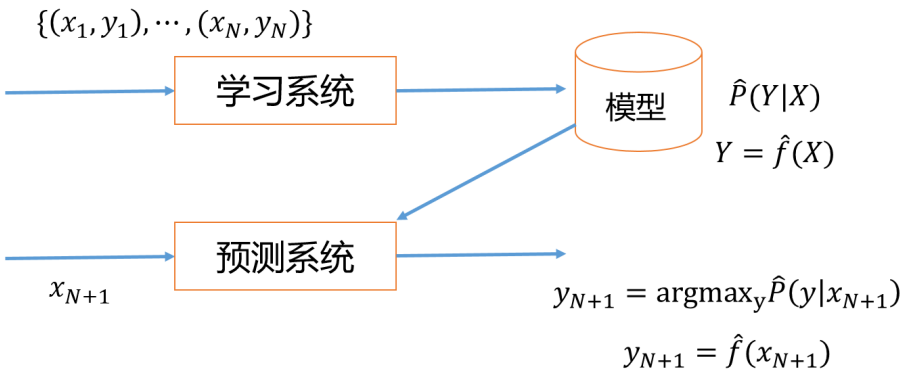

Figure 1: 学习系统与预测系统

2. 无监督学习

无监督学习是指从无标注数据中学习模型的机器学习问题。 其本质是学习数据中的统计规律与潜在结构。 模型可以实现对数据的聚类、降维或者概率估计。

假设$\mathbb{X}$是输入空间，$\mathbb{Z}$是隐式结构空间。要学习的模型可以表示为函数$z = g(x)$条件概率分布$P(z|x)$等形式。 其中$x \in \mathbb{X}$是输入， $z \in \mathbb{Z}$是输出。 无监督学习旨在从假设空间中选出给定评价标准下的最优模型。 在预测过程中，对于给定的x，输出$z = \hat{g}(x)$ 或者 $z = \arg\max_z \hat{P}(z|x)$.

2.2. 按模型分类

1. 概率模型与非概率模型

监督学习中，概率模型取条件概率分布形式 $P(y|x)$（例如：朴素贝叶斯模型），非概率模型取函数形式 $y = f(x)$（例如：神经网络模型）。

2. 线性模型与非线性模型

如果函数$y = f(x)$是线性函数，则称模型是线性模型，否则是非线性模型。 例如，线性回归模型是线性模型，深度学习则是复杂的非线性模型。

3. 参数化模型与非参数化模型

参数化模型(parametric model)假设模型的参数维度固定，模型可以由有限维参数完全刻画。 非参数化模型(non-parametric model)假设模型的参数维度不固定或者无穷大，随训练数据的增加而不断增大。

3. 统计学习方法的三要素

1. 模型

模型的假设空间包含所有可能的条件概率分布或者决策函数。假设空间中的模型一般无穷个。

假设空间用$\mathcal{F}$ 表示，假设空间可以定义为决策函数的集合： $\mathcal{F} = \{f|Y = f(X)\}$. 这时$\mathcal{F}$通常是由一个参数向量决定的函数族： $\mathcal{F} = \{f|Y = f_\theta(X) : \theta \in \mathbb{R}^p\}$.

2. 策略

统计学习的策略指的是按照何种准则学习或者选择最优模型。

(1) 损失函数和风险函数

损失函数：度量一次预测的好坏。
风险函数：度量平均意义下模型预测的好坏。

监督学习问题是从假设空间$\mathcal{F}$中选择模型f作为决策函数。 对于给定的输入变量X, $f(X)$ 为模型的预测值，Y为对应的真实值。 这里希望预测值与真实值尽量相似，则用一个损失函数（loss function）$L(Y, f(X))$来度量预测错误的程度。 常见的损失函数有：

（a）0-1损失函数（0-1 loss function）

$$L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases} \tag{3.1}$$

（b）平方损失函数（quadratic loss function）

$$L(Y, f(X)) = (Y - f(X))^2 \tag{3.2}$$

（c）绝对损失函数（absolute loss function）

$$L(Y, f(X)) = |Y - f(X)| \tag{3.3}$$

（d）对数损失函数（logarithmic loss function）或对数似然损失函数（log-likelihood loss function）

$$L(Y, P(Y|X)) = -\log P(Y|X) \tag{3.4}$$

风险函数（risk function）：即损失函数的期望 $E_{exp}(f) = E_P\{L(Y, f(X))\} = \int L(y, f(x))P(x, y)dxdy$. 但事实上，由于$P(X, Y)$未知，风险函数无法求得。

给定一个训练集 $T = \{(x_1, y_1), \cdots, (x_N, y_N)\}$. 模型$f(X)$关于训练数据集的平均损失称为经验风险（empirical risk）【或者经验损失（empirical loss）】:

$$R_{emp} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) \tag{3.5}$$

当$N \to \infty$时，经验风险$R_{emp}(f)$趋于期望风险$R_{exp}(f)$.

(2) 经验风险最小化与结构风险最小化

经验风险最小化(empirical risk minimization, ERM)即按照经验风险最小的准则求解最优模型：

$$\min_{f \in \mathcal{F}} \frac{1}{N}\sum_{i} L(y_i, f(x_i)) \tag{3.6}$$

当样本量足够大时，经验风险最小化能够保证有比较好的学习效果。 极大似然估计是经验风险最小的话的例子： 其中，模型是条件概率分布，损失函数是对数损失函数，此时经验风险最小化等价于极大似然估计。【请写出极大似然估计的目标函数】

但是，当样本量比较小时，经验风险最小往往不能得到好的学习效果，容易产生过拟合（over-fitting）的现象。

结构风险最小化是为了防止过拟合而提出的策略。 结构风险在经验风险上加上表示模型复杂度的正则化项（regularizer）或者是罚项（penalty term），其定义为：

$$R_{srm}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \tag{3.7}$$

其中， $J(f)$是模型的复杂度。 $J(f)$越大，则模型越复杂，否则模型越简单。 结构风险小的模型往往对训练数据和未知的测试数据都有较好的预测效果。

3. 算法

算法指学习模型的具体计算方法，可以归结为最优化问题的求解。

4. 模型评估与模型选择

1. 训练误差与测试误差

假设学习到的模型是$Y = \hat{f}(X)$，训练误差是模型关于训练数据集的平均损失：

$$R_{emp}(\hat{f}) = \frac{1}{N}\sum_{i} L(y_i, \hat{f}(x_i)) \tag{4.1}$$

测试误差是模型关于测试数据集的平均损失：

$$e_{test} = \frac{1}{N'}\sum_{i=1}^{N'} L(y_i, \hat{f}(x_i)) \tag{4.2}$$

当损失函数是0-1损失时，测试误差就变成了误差率（error rate）：

$$e_{test} = \frac{1}{N'}\sum_{i=1}^{N'} I(y_i \neq \hat{f}(x_i)) \tag{4.3}$$

2. 过拟合与模型选择

过拟合指学习时选择的模型所包含的参数过多，以至于模型对训练数据拟合较好，但对未知数据预测很差的现象。

【例1】假设给定一个训练数据集：$T = (x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$

其中，$x_i \in \mathbb{R}$是输入x的观测值，$y_i \in \mathbb{R}$是相应的输出y的观测值，$i = 1, 2, ..., N$。多项式函数拟合的任务是假设给定数据由M次多项式函数生成，选择最有可能产生这些数据的M次多项式函数，即在M次多项式函数中选择一个对已知数据以及未知数据都有很好预测能力的函数。

假设给定图2所示的10个数据点，用0 9次多项式函数对数据进行拟合。图中画出了需要用多项式函数曲线拟合的数据。

设M次多项式为

$$f_M(x, w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

式中x是单变量输入，$w_0, w_1, ..., w_M$是M+1个参数。

解决这一问题的方法可以是这样的。首先确定模型的复杂度，即确定多项式的次数；然后再给定的模型复杂度下，按照经验风险最小化的策略，求解参数，即多项式的系数，具体地，求以下经验风险最小化：

$$L(w) = \frac{1}{2} \sum_{i=1}^{N} \left(f\left(x_i, w\right) - y_i \right)^2 \tag{4.4}$$

这时，损失函数为平方损失，系数1/2是为了计算方便。

这是一个简单的最优化问题。将模型与训练数据代入式(4.4)中，有

$$L(w) = \frac{1}{2} \sum_{i=1}^{N} \left(f\left(x_i, w\right) - y_i \right)^2$$

这个问题可用最小二乘法求得拟合多项式系数的唯一解，记作$w_0^*, w_1^*, ..., w_M^*$，求解过程这里不予叙述，读者可参阅相关资料。

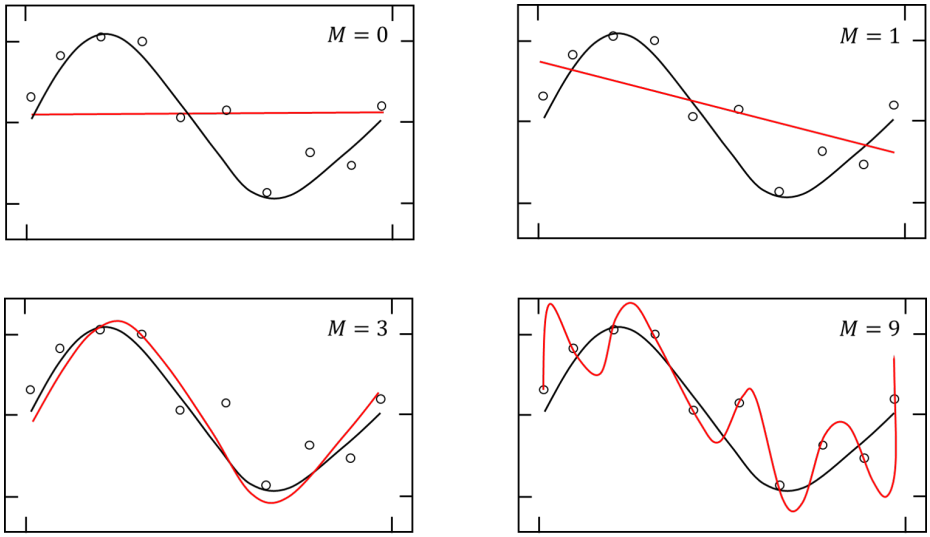

Figure 2: M次多项式函数拟合问题的例子

图2给出了M=0,M-1,M=3和M=9时多项式函数拟合的情况。如果M=0，多项式曲线是一个常数，数据拟合效果很差。如果M=1，多项式曲线是一条直线，数据拟合效果也很差。相反，如果M=9，多项式曲线通过每个数据点，训练误差为0。从对给定训练数据拟合的角度来说，效果是最好的。但是，因为训练数据本身存在噪声，这种拟合曲线对未知数据的预测能力往往并不是最好的，在实际学习中并不可取。这是过拟合现象就会发生。这就是说，模型选择时，不仅要考虑对已知数据的预测能力，还要考虑对未知数据的预测能力。当M=3时，多项式曲线对训练数据拟合效果足够好，模型也比较简单，是一个较好的选择。

在多项式函数拟合中可以看到，随着多项式次数（模型复杂度）的增加，训练误差会减小，直至趋向于0，但是测试误差却不如此，它会随着多项式次数（模型复杂度）的增加先减小而后增大。而最终的目的是使测试误差达到最小。这样，在多项式函数拟合中，就要选择合适的多项式次数，以达到这一目的。这个结论对一般的模型选择也是成立的。

图3描述了训练误差和测试误差与模型的复杂度之间的关系。当模型的复杂度增大时，训练误差会逐渐减小并趋向于0；而测试误差会先减小，达到最小值后又增大。当选择的模型复杂度过大时，过拟合现象就会发生。

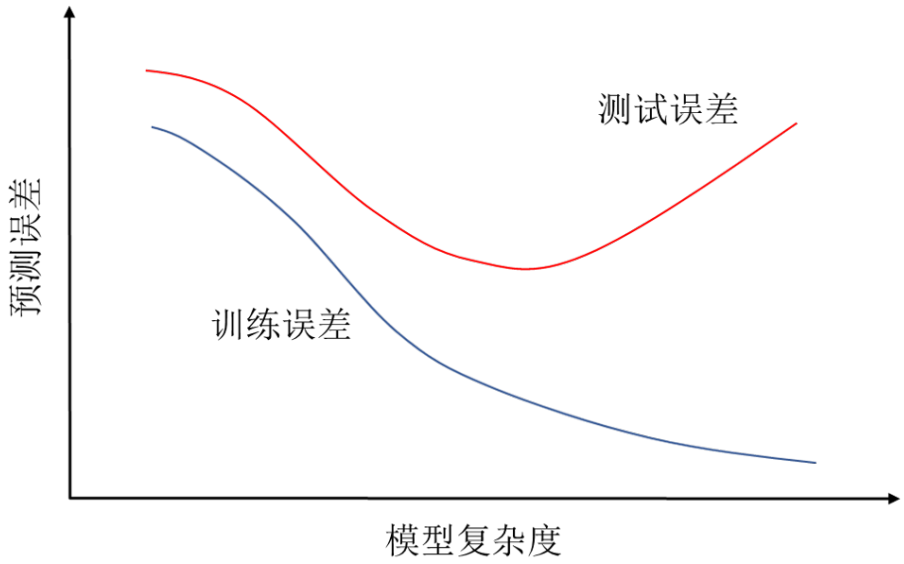

Figure 3: 训练误差和测试误差与模型复杂度的关系

因此，在学习时应防止过拟合，应选择复杂度适当的模型，达到测试误差最小的学习目的。

实践中常用两种模型选择方法：正则化与交叉验证。

5. 正则化与交叉验证

1. 正则化

正则化形式如下：

$$\min \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \tag{5.1}$$

其中，$\lambda \geq 0$是调整经验风险和正则化项之间的系数。 例如，在回归模型中，可以取以下形式：

$$\min \frac{1}{N}\sum_{i=1}^{N} \{y_i - f(x_i; w)\}^2 + \lambda \|w\|_1 \tag{5.2}$$

正则化的作用是选择经验风险与模型复杂度同时较小的模型。 它在模型选择中基于的想法是：选择能够很好的解释已知数据并且简单的模型。

2. 交叉验证

如果样本数据充足，可以将数据分成三部分：训练集（training set）、验证集（validation set）以及测试集（test set）。 训练集用于训练模型，验证集用于模型选择，测试集用于模型评估。

实际中往往数据不充足，这时往往采用交叉验证（cross validation）的方式，通过重复使用数据，对数据进行切分，反复实现模型的训练、测试和评估。

(1) 简单交叉验证 将数据随机分成两部分，一部分做训练集、一部分做测试集。在不同的参数下训练模型，进行模型评估。

(2) S折交叉验证

随机将数据切分成S折互不相交的子集，利用其中的S-1折子集进行数据训练，用余下的一折进行模型测试。这种评测可以进行S次，选择S次平均误差最小的模型。

(3) 留一交叉验证（leave-one-out cross validation）

设置(2)中的S = N，则称为留一交叉验证。在数据缺乏时使用。

注：以上交叉验证都需要重复试验R次，对模型误差求平均，以获得稳定的模型评估结果。

6. 泛化能力

1. 泛化误差

泛化误差：如果学到的模型是$\hat{f}$，那么用这个模型对未知数据进行预测误差就是泛化误差（generalization error）:

$$R_{exp}(\hat{f}) = \int L(y, \hat{f}(x))P(x, y)dxdy \tag{6.1}$$

泛化误差就是所学习到的模型的期望风险，表示学习方法的泛化能力。

2. 泛化误差上界*

泛化误差的上界具有以下性质：它是样本容量的函数，样本容量增加时，泛化上界趋于0； 它是假设空间容量的函数，假设空间容量越大，则模型学习难度越高，

泛化误差上界越大。

【例2】（二分类问题的泛化误差上界）

已知训练集数据 $T = \{(x_1, y_1), \cdots, (x_N, y_N)\}$由联合概率分布$P(X, Y)$独立同分布产生， $X \in \mathbb{R}^n$, $Y \in \{-1, 1\}$. 假设空间的函数是有限集合 $\mathcal{F} = \{f_1, \cdots, f_d\}$. 损失函数为0-1损失函数。 设$f$ 的期望风险和经验风险分别表示为：

$$R(f) = E\big\{L(Y, f(X))\big\}, \quad \hat{R}(f) = \frac{1}{N}\sum_i L(y_i, f(x_i)).$$

经验风险最小化的函数为：$f_N = \arg\min_{f\in\mathcal{F}} \hat{R}(f)$. 则$f_N$ 的泛化误差可以表示为：$R(f_N) = E\{L(Y, f_N(X))\}$.

Theorem 1. *(泛化误差上界)* 对于以上二分类问题，当假设空间为有限个函数的集合$\mathcal{F} = \{f_1, \cdots, f_d\}$时，至少以概率$1 - \delta$， $0 < \delta < 1$,以下不等式成立：

$$R(f) \le \widehat{R}(f) + \varepsilon(d, N, \delta), \tag{6.2}$$

其中，

$$\varepsilon(d, N, \delta) = \Big\{\frac{1}{2N}\Big(\log d + \log\frac{1}{\delta}\Big)\Big\}^{1/2}. \tag{6.3}$$

注：(6.2) 的左边是泛化误差，右边是泛化误差上界。 其中，$\widehat{R}(f)$ 代表训练误差。 给定δ时， $\varepsilon(d, N, \delta)$ 是 d的递增函数，N的递减函数。

Proof:

在证明中要用到Hoeffding不等式，先叙述如下：

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

$$
\begin{aligned}
P\left(S_n - ES_n \geqslant t\right) &\leqslant \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}\left(b_i - a_i\right)^2}\right) \\
P\left(ES_n - S_n \geqslant t\right) &\leqslant \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}\left(b_i - a_i\right)^2}\right)
\end{aligned}
\tag{6.4}
$$

对任意函数$f \in \mathcal{F}$，$\hat{R}(f)$是N个独立随机变量$L(Y, f(X)$的样本均值，$R(f)$是随机变量$L(Y, f(X))$的期望值。如果损失函数取值于区间$[0, 1]$，即对所有i，$[a_i, b_i] = [0, 1]$，那么由Hoeffding不等式可知，对$\epsilon > 0$，以下不等式成立：

$$
P(R(f) - \hat{R}(f) \geqslant \varepsilon) \leqslant \exp\left(-2N\varepsilon^2\right)
$$

由于$\mathcal{F} = f_1, f_2, ..., f_d$是一个有限集合，故

$$
\begin{aligned}
P(\exists f \in \mathcal{F} : R(f) - \hat{R}(f) \geqslant \varepsilon) &= P\left(\bigcup_{f \in \mathcal{F}} \{R(f) - \hat{R}(f) \geqslant \varepsilon\}\right) \\
&\leqslant \sum_{f \in \mathcal{F}} P(R(f) - \hat{R}(f) \geqslant \varepsilon) \\
&\leqslant d \exp\left(-2N\varepsilon^2\right)
\end{aligned}
$$

或者等价的，对任意$f \in \mathcal{F}$，有

$$
P(R(f) - \hat{R}(f) < \varepsilon) \geqslant 1 - d \exp\left(-2N\varepsilon^2\right)
$$

令

$$
\delta = d \exp\left(-2N\varepsilon^2\right) \tag{6.5}
$$

则

$$
P(R(f) < \hat{R}(f) + \varepsilon) \geqslant 1 - \delta
$$

即至少以概率$1 - \delta$有$R(f) < \hat{R}(f) + \epsilon$，其中$\epsilon$由式(6.5)得到，即为式(6.3)。