第1章 统计学习方法概论

1.1 统计学习

1.统计学习的特点

- 定义：计算机系统运用数据和统计方法提高性能的机器学习，又称统计机器学习。
- 平台：计算机和网络。
- 对象：数据。
- 目的：对数据进行预测与分析。
- 方法：基于数据构建概率统计模型。
- 理论：概率论、统计学、信息论、计算理论、最优化理论、计算机科学。

2.统计学习的对象

- 对象：数据。
- 表示：变量或者变量组。
- 分类：连续变量和离散变量。
- 基本假设：同类数据具有一定的统计规律性，所以可用随机变量描述数据中的特征，用概率分布描述数据的统计规律。
- 连续变量：不可数的变量。
- 离散变量：可数的变量。

3.统计学习的目的

- 如何选择模型和学习模型，使模型对数据进行准确的预测和分析，同时提高学习效率。

2.统计学习的方法

- 组成：监督学习、无监督学习、半监督学习、强化学习。
- 假设空间：模型所属函数的集合。
- 要素：模型(模型的假设空间)、策略(模型选择的准则)、算法(模型学习的算法)。
- 实现步骤：
    - ①获取一个有限的训练数据集合。
    - ②确定包含所有可能的模型的假设空间。
    - ③确定模型选择的准则。
    - ④实现求解最优模型的算法。
    - ⑤通过学习方法选择最优模型。
    - ⑥利用学习的最优模型对新数据进行预测或分析。

1.2 监督学习

1.监督学习

- 监督学习任务：学习一个模型，使模型能对任意给定的输入，对其相应的输出做出一个好的预测。
- 输入空间：所有可能取值的输入集合。
- 输出空间：所有可能取值的输出集合。
- 线性空间：定义在某个数域上且满足加法和数乘规则的集合。
- 内积空间：带有内积运算的线性空间。
- 欧式空间：定义在实数域上的内积空间。
- 实例：一个具体的输入，用特征向量x表示。
- 特征空间：所有特征向量存在的空间，每一维对应一个特征，通过输入空间的映射得到。
- 输入：定义在输入(特征)空间上随机变量的取值，输入变量用X表示，输入变量取值用x表示。
- 输出：定义在输出空间上随机变量的取值，输出变量用Y表示，输出变量取值用y表示。

- 输入实例的特征向量：记作$x=(x^{(1)},x^{(2)},\dots,x^{(i)},\dots x^{(n)})^T$，其中$x^{(i)}$表示实例$x$的第$i$个特征。
- 第i个输入实例的特征向量：记作$x_i=(x_i^{(1)},x_i^{(2)},\dots,x_i^{(i)},\dots x_i^{(n)})^T$，其中$x_i$表示第$i$个实例。
- 训练集：由输入与输出对组成，记作$T=\{(x_1,y_1),(x_2,y_2),\dots,(x_N,y_N)\}$。
- 样本：输入与输出对，又称样本点。
- 回归问题：输入变量与输出变量均为连续变量的预测问题。
- 分类问题：输出变量维有限个离散变量的预测问题。
- 标注问题：输入变量与输出变量均为变量序列的预测问题。

2.联合概率分布

- 联合概率：事件AB同时发生的概率，记作$P(AB)$。
- 条件概率：事件A发生的条件下，事件B发生的概率$P(B|A)$。
- 联合概率与条件概率的关系：$P(AB)=P(B|A)P(A)$。
- 概率分布函数：随机变量小于特定值的概率，记作$F(x)=P\{X\leq x\}$，又称概率分布或者分布函数或者分布，离散型和连续型的概率分布函数都满足这个定义，其中离散型又满足下面的定义。
- 离散型随机变量的概率分布：离散型随机变量取特定值的概率，记作$F(x)=P\{X=x\}=p$。
- 联合概率分布：记作$P(X,Y)$，其中X为输入，Y为输出，又记作$F(x,y)=P\{X\leq x,Y\leq y\}$。
- 独立：样本之间相互独立，满足$P(AB)=P(A)P(B)$。
- 同分布：样本都服从相同的概率分布。
- 监督学习的基本假设：输入X和输出Y符合联合概率分布P，且样本数据要求独立同分布。
  - ①假设任意划分样本数据为训练集A和测试集B，$A=\{(ax_1,ay_1),\dots,(ax_n,ay_n)\}$，$B=\{(bx_1,by_1),\dots,(bx_n,by_n)\}$，因为样本之间相互独立，所以$P(AB)=P(A)P(B)$。
  - ②因为样本之间同分布，所以可以用训练集训练一个联合概率分布模型，然后对测试集进行预测。
- 注意：$P(X,Y)$有的时候表示概率分布函数，有的时候表示概率密度函数，要看语境分析。

3.假设空间

- 定义：由输入空间到输出空间的映射的集合。
- 监督学习的模型
  - ①概率模型，即条件概率分布，记作$P(Y|X)$或者$P(y|x)$。
  - ②非概率模型，即决策函数，记作$Y=f(X)$或者$y=f(x)$。

4.监督学习问题的形式化

- 学习过程：学习系统根据给定的训练数据集，通过学习(训练)得到概率模型$P(y|x)$或者非概率模型$f(x)$。
- 预测过程：预测系统根据给定的测试数据集中的输入x_{N+1}，由模型$y_{N+1}=\arg\max\limits_{y_{N+1}}P(y_{N+1}|x_{N+1})$或者$y_{N+1}=f(x_{N+1})$得到输出$y_{N+1}$。

1.3 统计学习三要素

1.模型——模型的假设空间

- 非概率模型的假设空间：记作$F=\{f|Y=f(X)\}$或者$F=\{f|Y=f_\theta(X),\theta\in R^n\}$。
- 概率模型的假设空间：记作$F=\{P|P(Y|X)\}$或者$F=\{P|P_\theta(Y|X),\theta\in R^n\}$。
- 参数空间：假设空间中参数向量所在的欧式空间R^n。

2.策略——模型的选择准则

- MLE最大似然估计：通过样本估计概率分布的参数θ，属于频率学派。

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg\max P(X;\theta) \\ &= \arg\max P\left(x_1;\theta\right) P\left(x_2;\theta\right) \cdots P\left(x_n;\theta\right) \\ &= \arg\max \log \textstyle\prod_{i=1}^{n} P\left(x_i;\theta\right) \\ &= \arg\max \textstyle\sum_{i=1}^{n} \log P\left(x_i;\theta\right) \\ &= \arg\min -\textstyle\sum_{i=1}^{n} \log P\left(x_i;\theta\right)\end{aligned}$$

- MAP最大后验估计：通过样本和先验概率估计后验概率分布的参数θ，属于贝叶斯学派。

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max P(\theta|X) \\ &= \arg\min -\log P(\theta|X) \\ &= \arg\min -\log P(X|\theta) - \log P(\theta) + \log P(X) \\ &= \arg\min -\log P(X|\theta) - \log P(\theta)\end{aligned}$$

- 损失函数：(Loss)，度量模型一次预测的好坏，又称为代价函数，记作$L(Y, f(X))$或$L(Y, P(Y|X))$。
  - 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：$L(Y, f(X)) = (Y - f(X))^2$
  - 绝对损失函数：$L(Y, f(X)) = |Y - f(X)|$
  - 对数损失函数：$L(Y, P(Y|X)) = \log \frac{1}{P(Y|X)} = -\log P(Y|X)$
- 期望风险：(expect Risk)，度量模型平均预测的好坏，模型关于联合分布的期望损失，或损失函数的期望，又称期望损失/风险函数，记作$R_{exp}(f)$。
  - 符号说明：
    - 损失函数为L。
    - 期望为E。
    - 风险函数为R。
    - 输入变量为X，实例为x。
    - 输出变量为Y，实例为y。
    - 关于X, Y的联合概率密度函数为$P(X, Y)$
  - 公式：

$$R_{\exp}(f) = E_P[L(Y, f(X))] = \int_{x \times y} L(y, f(x)) P(x, y) dx dy$$

  - 说明1：
    - ①计算EX：如果随机变量为离散型随机变量，那么随机变量的数学期望等于随机变量值与对应概率的乘积和。如果随机变量为连续性随机变量，那么随机变量的数学期望等于随机变量与概率密度函数乘积的积分。
    - ②计算$E(f(X))$：f是连续函数，如果X是离散型随机变量，它的分布律$P\{X = x_k\} = p_k$，且$\sum_{k=1}^{\infty} g\left(x_k\right) p_k$收敛，那么$E[g(X)] = \sum_{k=1}^{\infty} g\left(x_k\right) p_k$；如果$X$是连续型随机变量，它的概率密度函数为$f(x)$，且$\int_{-\infty}^{\infty} g(x) f(x) dx$收敛，那么$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$。
  - 说明2：
    - 学习目标：选择期望风险最小的模型。
    - 无法学习：通过最小期望风险进行学习需要已知联合分布，但是联合分布是未知的。
    - 解决方法：用经验风险代替期望风险。
- 经验风险：(empirical risk)，训练集的平均损失。
  - 符号说明
    - 训练集为$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_n)\}$。
    - 模型或者决策函数为$f(X)$
    - 经验风险R_{emp}
  - 公式：

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

- 说明：
  - 期望风险是关于联合分布的期望损失。
  - 经验风险是模型关于训练集的平均损失。
  - 辛钦大数定律：设随机变量$X_1,X_2,\ldots X_n,\ldots$独立同分布，且有数学期望$E(X_i)=\mu,i=1,2,\ldots$，则对任意$\epsilon>0$，有

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|<\varepsilon\right\}=1$$

- 根据大数定律，当样本容量N趋于无穷，经验风险趋于期望风险，所以可以用经验风险估计期望风险。但是现实种训练样本数目有限，所以要对经验风险进行矫正。
- 经验风险最小化：求最优模型的策略，即求解下面的最优化问题，即找到使经验风险最小的函数f，F是假设空间。

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

  - 说明1：当训练集样本量足够大效果好，当训练样本量很小可能产生过拟合现象。
  - 说明2：当模型是条件概率分布，损失函数是对数损失函数时，经验风险最小化就等价于最大似然估计。

$$\text{即}\min_{f\in F}\frac{1}{N}\sum_{i=1}^{N}(-\log P(y_i|x_i))\text{等价于}\arg\min-\sum_{i=1}^{n}\log P\left(x_i;\theta\right)$$

- 结构风险最小化：(structural risk)，为了防止过拟合，结构风险=经验风险+表示模型复杂度的正则化项。
  - 符号说明
    - 表示模型复杂度的正则化项为$J(f)$，是定义在假设空间上的泛函。
  - 公式：

$$R_{srm}(f)=\frac{1}{N}\sum_{i=1}^{N}L\left(y_i,f\left(x_i\right)\right)+\lambda J(f),\text{其中}\lambda\geq0$$

  - 说明1：结构风险最小需经验风险和模型复杂度都小，对训练集和测试集都有较好的预测。
  - 说明2：泛函，定义域是函数集，值域是实数或者复数。输入是函数，输出是标量。
  - 说明3：先验概率，事件A发生前的预估概率$P(A)$。后验概率，事件A发生条件下，事件B发生的概率为$P(B|A)$，后验概率属于条件概率。
  - 说明4：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。先验概率越小，模型越复杂。

$$\text{即}\min_{f\in F}\frac{1}{N}\sum_{i=1}^{N}(-\log P(y_i|x_i))+\lambda J(P(x_i))\text{等价于}\arg\min-\log P(X|\theta)-\log P(\theta)$$

3.算法——模型学习的算法

- 统计学习问题归结为最优化问题。
- 统计学习算法归结为最优化算法。

1.4 模型评估与模型选择

1.训练误差与测试误差

- 误差率：预测错误的样本数/总样本数。

- 准确率：预测正确的样本数/总样本数。
- 泛化能力：学习方法对未知数据的预测能力。

2.过拟合与模型选择

- 过拟合：所选择的模型复杂度比真模型的复杂度高，所选择的模型包含的参数过多。
- 训练误差和测试误差与模型复杂度的关系：模型复杂度增加，训练误差下降，测试误差先下降后增加。

1.5 正则化与交叉验证

1.正则化

- 正则化：结构风险=经验风险+正则化项。正则化项是模型复杂度的单调递增函数，模型越复杂。正则化值越大。
- 正则化项的形式：
  - 回归问题，损失函数是平方损失，正则化项可以是参数向量w的L_p范数。

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

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

- 奥卡姆剃刀原理：若无必要，勿增实体。在所有可选择模型种，最好的模型是对已知数据预测效果好且模型复杂度低的模型。

2.交叉验证

- 直接验证：如果样本充足，随机将数据集分为三部分，训练集，验证集和测试集，训练集用来训练模型，验证集用于模型选择，测试集用于对模型评价。在不同复杂度的模型中，选择对验证集有最小预测误差的模型。
- 交叉验证：如果样本不充足。将数据集分为两部分，训练集和测试集，然后进行反复训练和测试。
- 简单交叉验证
  - ①随机的将数据分为两部分，一部分为训练集，另一部分为训练集。
  - ②用训练集在不同超参数下训练模型，得到不同的模型。
  - ③在测试集上评价各个模型的测试误差，选出测试误差最小的模型。
- S折交叉验证
  - ①随机地将数据分为S个互不相交地大小相同的子集。
  - ②选择用$S-1$个子集的数据训练模型。
  - ③用余下的1个子集测试模型。
  - ④重复②③过程，一共需要S轮。
  - ⑤选择S轮后平均测试误差最小的模型。
- 留一交叉验证：样本数量严重不足时，S折交叉验证的S取数据集容量，即$S=N$。

1.6 泛化能力

1.泛化误差

- 泛化能力：学习方法的泛化能力是由该方法学习到的模型对未知数据的预测能力。
- 训练误差：模型在训练集上的误差，对应的是经验风险。
- 测试误差：模型在测试集上的误差，计算方法和训练误差相同。
- 泛化误差：模型对未知数据预测的误差，对应的是期望风险。
  - 公式：

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

- 说明：泛化误差无法计算，但是泛化误差的上界可以计算。
- 泛化能力可以用测试误差或者泛化误差来评价。

2.泛化误差上界

- 泛化误差上界的作用：用来比较学习方法的优劣。
- 泛化误差上界的性质：
  - ①泛化误差上界是样本容量的函数，当样本容量增加时，泛化上界趋于0。
  - ②泛化误差上界是假设空间容量的函数，假设空间容量越大，模型越难学，泛化误差上界越大。
- 二分类问题的泛化误差上界举例：
  - 符号说明：
    - 训练集$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$
    - 联合概率分布为$P(X, Y)$，训练集根据P独立同分布产生
    - 输入为$X \in \mathbb{R}^n$
    - 输出为$Y \in \{-1, +1\}$
    - 假设空间为函数的有限集合$F = \{f_1, f_2, \dots, f_d\}$
    - 从F中选择的函数为f
    - 损失函数为0-1损失
  - 期望风险：$R(f) = E[L(Y, f(X))]$
  - 经验风险：$\hat{R}(f) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, f\left(x_i\right)\right)$
  - 通过经验风险最小化选择函数f：$f_N = \arg\min_{f \in F} \hat{R}(f)$
  - f_N的泛化误差：$R\left(f_N\right) = E\left[L\left(Y, f_N(X)\right)\right]$
- **从有限集合$F = \{f_1, f_2, \dots, f_d\}$中任意选出函数$f$，计算$f$的泛化误差上界。**
  - 符号说明：
    - 假设空间，有限函数集合为F
    - 函数f的泛化误差为$R(f)$
    - 函数f的训练误差为$\hat{R}(f)$
    - 有限函数集合F的函数数量为d
    - 训练集样本数量为N
    - 给定一个概率为$1-\delta$
  - 公式

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

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

- 说明：训练误差越小，泛化误差越小；训练样本越多，泛化误差越小；假设空间越小，泛化误差越小。

1.7 生成模型与判别模型

- 监督学习方法分类：生成方法，判别方法。
- 监督学习模型分类：生成模型，判别模型。
- 生成方法：通过数据学习联合概率分布$P(X, Y)$，求出条件概率分布作为$P(Y|X)$作为预测的模型即生成模型：$P(Y|X) = \frac{P(X,Y)}{P(X)}$。
- 典型的生成模型：朴素贝叶斯，隐马尔可夫。

- 判别方法：通过数据直接学习决策函数$f(X)$或者条件概率分布$P(Y|X)$作为预测的模型即判别模型。
- 典型的判别模型：k近邻、感知机、决策树、逻辑斯蒂回归，最大熵，支持向量机，提升方法，条件随机场。
- 生成方法相比判别方法的优点：
  - ①可以还原出联合概率分布$P(X,Y)$。
  - ②学习收敛速度快，当样本容量增加，学习的模型更快的收敛于真实模型。
  - ③存在隐变量时仍然可以使用生成方法。
- 判别方法相比生成方法的优点：
  - ①直接学习条件概率$P(Y|X)$或者决策函数$f(X)$。
  - ②学习准确率更高。
  - ③直接学习条件概率或者决策函数，可对数据抽象、定义、使用特征来简化问题。

1.8 分类问题

- 监督学习的核心问题：分类问题。
- 分类问题：输出变量Y取有限个离散值，输入变量X既可以离散，也可以连续。
- 分类器：监督学习从数据中学习一个分类模型或者分类决策函数。
- 分类：分类器对新的输入进行输出的预测。
- 学习：根据已知的训练数据集通过学习方法学习一个分类器。
- 准确率：对于给定测试集，分类准确率=正确分类样本数/总样本数=将正类预测为正类数+将负类预测为负类数/样本总数。
- 正类：关注的类。
- 负类：除了正类其他的类。
- TP：将正类预测为正类数。
- TN：将负类预测为负类数。
- FP：将负类预测为正类数。
- FN：将正类预测为负类数。
- 精确率P：将正类预测为正类数/(讲正类预测为正类数+将负类预测为正类数)
- 召回率R：将正类预测为正类数/(将正类预测为正类数+将正类预测为负类数)
- F_1值：精确率和召回率的调和均值，$\frac{2}{F_1}=\frac{1}{P}+\frac{1}{R}$，$F_1=\frac{2TP}{2TP+FP+FN}$.
- 应用举例:①银行根据客户信息和贷款风险大小进行分类。②网络安全使用日志数据的分类对非法入侵进行检测。③图像处理，分类可以进行目标检测，图像识别。④网页分类可以进行网页抓取索引和排序。
- ROC曲线：
  - 横轴=将负类预测为正类数/(将负类预测为正类数+将负类预测为负类数)；
  - 纵轴=将正类预测为正类数/(将正类预测为正类数+将正类预测为负类数)。
- 最优曲线：将正类预测为负类数=0时对应的曲线。
- AUC值：ROC曲线下面的面积。
- AUC最大值：ROC曲线下面面积最大时。

1.9 标注问题

- 标注问题：分类问题的推广。输入一个观测序列$x_i=\left(x_i^{(i)},x_i^{(2)},\cdots,x_i^{(n)}\right)^T,i=1,2,\ldots,N$

输出一个标记序列$y_i = \left(y_i^{(n)}, y_i^{(2)}, \cdots, y_i^{(n)}\right)^T$,其中$n$是一个序列的长度，一般$n << N$。

- 标注器：通过训练数据集学习一个模型表示为条件概率分布：
  $P\left(Y^{(1)}, Y^{(2)}, \cdots, Y^{(n)} | X^{(1)}, X^{(2)}, \cdots, X^{(n)}\right)$
- 标注：对于一个观测序列$x_{N+1} = \left(x_{N+1}^{(1)}, x_{N+1}^{(2)}, \cdots, x_{N+1}^{(n)}\right)^T$找到使条件概率$P(y_{N+1}|x_{N+1})$最大的标记序列$y_{N+l} = \left(y_{N+1}^{(1)}, y_{N+1}^{(2)}, \cdots, y_{N+1}^{(n)}\right)^T$。
- 评价指标：标注准确率、精确率、召回率。
- 常用方法：隐马尔可夫、条件随机场。
- 应用举例：NLP中词性标注，信息抽取。

1.10 回归问题

- 回归问题：预测输入变量(自变量)和输出变量(因变量)之间的关系，回归问题的学习等价于函数拟合。
- 回归模型：从输入变量到输出变量之间映射的函数。
- 一元/多元回归：输入变量个数。
- 线性/非线性回归：输入变量和输变量出之间的关系的类型即关系模型。
- 最小二乘法：在回归学习中使用的平方损失函数进行函数拟合。
- 应用举例：股价预测。