《统计学习方法》

第一章讲解

作者：李航

演讲者：王宗威

统计学习的特点

- （1）以计算机及网络为平台，是建立在计算机及网络之上的。
- （2）统计学习以数据为研究对象
- （3）目的是对数据进行预测与分析
- （4）统计学习以方法为中心，统计学习方法构建模型并运用模型进行预测与分析
- （5）是概率论、统计学等多个领域的交叉学科

统计学习的对象是数据

- 从数据出发
- 提取数据特征
- 抽象出数据模型
- 发现数据中的知识
- 回到对数据的分析与预测中

统计学习的目的

- 对数据进行预测与分析

统计学习的方法

- 监督学习
- 非监督学习
- 半监督学习
- 强化学习

统计学习的三要素

- 模型的假设空间，简称模型
- 模型选择的准则，简称策略
- 模型学习的算法，简称算法

统计学习方法的步骤

- （1）得到训练数据集合
- （2）确定假设空间
- （3）确定学习的策略
- （4）确定学习的算法
- （5）选择最优模型
- （6）预测与分析

监督学习的基本概念

- 输入空间
- 输出空间
- 特征空间：所有特征向量存在的空间称为特征空间

模型实际上是定义在特征空间上的。

欧式空间：设V是实数域R上的线性空间（或称为向量空间），若V上定义着正定对称双线性型g（g称为内积），则V称为（对于g的）内积空间或欧几里德空间（有时仅当V是有限维时，才称为欧几里德空间）。

联合概率分布

- 两个及以上随机变量组成的随机向量的概率分布
- 监督学习假设输入与输出的随机变量X和Y遵循联合概率分布P（X, Y），P（X,Y）表示分布函数或者分布密度函数

统计学习三要素

- 模型
- 策略
- 方法

模型

- 假设空间为决策函数的集合
- 假设空间为条件概率的集合

策略

- 1.损失函数
- 0-1损失函数
- 平方损失函数
- 绝对损失函数
- 对数损失函数
- 风险函数（期望损失）：即损失函数的期望

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

策略

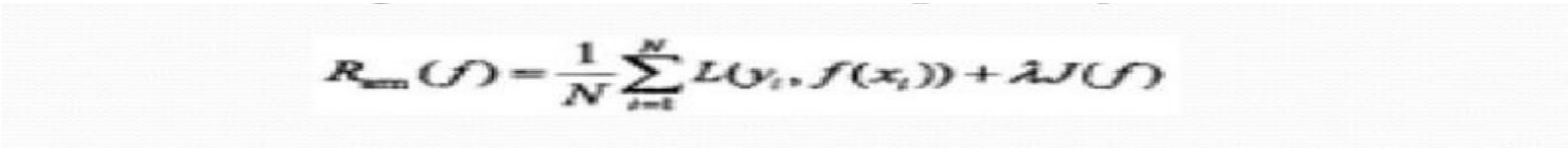
- 经验风险（经验损失）：模型f（X）关于训练数据集的平均损失

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

- 经验风险最小化（ERM）：
- P9第一段第二行
- 极大似然估计，习题里讲

策略

- 结构风险最小化
- 等价于正则化
- 正则化：代数几何中的一个概念，就是对最小化经验误差函数上加约束，这样的约束可以解释为先验知识(正则化参数等价于对参数引入先验分布)。约束有引导作用，在优化误差函数的时候倾向于选择满足约束的梯度减少的方向，使最终的解倾向于符合先验知识。
- 结构风险在经验风险上加上表示模型复杂度的正则化项或罚项，定义为：



策略

- 先验概率：是指根据以往经验和分析得到的概率
- 后验概率：后验概率是指在得到“结果”的信息后重新修正的概率

算法

- 学习模型的具体计算方法

模型评估

- 训练误差：训练数据集的平均损失
- 测试误差：测试数据集的平均损失

- 训练误差本质上不重要，对判定问题是否容易有意义
- 测试误差小的方法具有更好的预测能力

过拟合和模型选择

- 过拟合：指学习时选择的模型所包含的参数过多，以至于出现这一模型对已知数据预测的很好。

- 例1.1说明过拟合与模型选择的关系

- 最小二乘法：最小二乘法（又称最小平方法）是一种数学优化技术。它通过最小化误差的平方和寻找数据的最佳函数匹配。利用最小二乘法可以简便地求得未知的数据，并使得这些求得的数据与实际数据之间误差的平方和为最小。

模型选择方法

- 正则化
- 交叉验证

模型选择方法

- 正则化

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

- L2范数：

L2范数即欧氏距离：

$$||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

L2范数越小，可以使得w的每个元素都很小，接近于0，但L1范数不同的是他不会让它等于0而是接近于0.

模型选择方法

- L1范数：

L1范数表示向量中每个元素绝对值的和：

$$||x||_1 = \sum_{i=1}^{n} |x_i|$$

- 奥卡姆剃刀原理：
- 《机器学习》P7
- 即若有多个假设与观察一致，则选择最简单的那个

模型选择方法

- 交叉验证：分为训练集，验证集和测试集
- 简单交叉验证
- S折交叉验证：？？？
- 会不会再次选到之前选到过得？
- 回答：不会，第一次迭代中留存第1份，第二次留存第2份，其余依此类推，第i次留存第i份。
- 留一交叉验证

泛化误差

- 泛化能力：指由该方法学习到的模型对未知数据的预测能力
- 泛化误差：

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

- 和损失函数的期望很相似？
- 事实上，泛化误差就是所学习到的模型的期望风险。

泛化误差上界

定理：泛化误差上界，二分类问题，当假设空间是有限个函数的结合 $\mathcal{F}=\{f_1,f_2,\cdots,f_d\}$，对任意一个函数f，至少以概率1-δ，以下不等式成立：

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

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

泛化误差上界

- 证明：
- Hoeffding不等式
- 过程

生成模型和判别模型

- 监督学习方法分为生成方法和判别方法：
- 所学到的模型为生成模型和判别模型
- 生成模型：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）作为预测的模型
- 判别模型：由数据直接学习决策函数f（X）或条件概率分布P（X|Y）
- 生成方法的特点
- 判别方法的特点

分类问题

- 输入变量X可以是离散的，也可以是连续的
- 输出变量Y取有限个离散值
- 分别学习和分类两个过程
- 分类器
- 评价分类器性能的指标一般是分类准确度
- 精确率：判别正类中真正是正类的概率
- 召回率：正类真正判别出来的概率
- 调和均值

标注问题

- 输入是一个观测序列
- 输出是一个标记序列或状态序列
- 分为学习和标注两个过程
- 指标常用的有标注准确率、精准率和召回率
- 统计学习方法有：隐马尔科夫模型、条件随机场。

回归问题

- 映射输入变量和输出变量之间的关系
- 按输入变量分：分为一元回归和多元回归
- 按输入变量和输出变量之间的类型即模型的类型分：分为线性回归和非线性回归
- 回归学习最常用的损失函数是平方损失函数

习题

- 打开文档

第一章结束

谢谢大家！~