统计学习方法概论

> *统计学习方法 李航 笔记*

简介

统计学习方法三要素: model, startegy, algorithm

步骤如下

1. 得到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, 即学习模型的集合
3. 确定模型选择的基准, 即学习策略
4. 通过学习方法选择最优模型
5. 利用学习的最优模型对新数据进行预测或分析

监督学习

样本空间

输入空间: 输入的所有可能值的集合

输出空间: 输出的所有可能值的集合

特征空间: 每个输入通过特征向量表示, 特征向量存在的空间为特征空间, 有可能与输入空间相同. 模型都是作用于特征空间

基本假设

监督学习假设输入输出的随机变量X, Y 遵循联合概率分布 $P(X, Y)$

训练与测试数据被看做依据联合概率分布独立同分布产生的.

假设空间: hypothesis space 模型属于输入空间到输入空间映射的集合, 该集合为假设空间. 假设空间里的模型一般有无穷多个

预测的记过根据条件概率分布或者决策函数表示$P(Y|X)$ 或者 $Y = f(X)$

统计学习三要素

方法 = 模型 + 策略 + 算法

模型

模型存在于假设空间中, 假设空间$\mathcal{F}$可以定义为决策函数的集合(非概率模型)

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

决策函数是由参数向量决定的函数族. 同时也可以定义为条件概率的集合(概率模型)

$$\mathcal{F} = \{P|P_\theta(Y|X), \theta \in \mathbf{R}^n\}$$

策略

统计学习的目标: 从假设空间中选组最优模型.

策略意味着选区最优模型的准则. 损失函数和风险函数: 损失函数度量模型一次预测的好坏, 风险函数度量**平均意义**下模型度量的好坏.

损失/风险函数

损失/代价函数用来度量预测错误程度

1. 0-1 loss function
2. quadraric loss function 平方损失函数
3. absolute loss function 绝对损失函数
4. logarithmic loss function or log-likelihood loss function 对数损失函数或对数似然损失函数

$$L(Y, P(Y|X)) = -\log P(Y|X)$$

由于输入数据是随机变量, 假设遵循联合分布$P(X, Y)$, 风险函数**risk function** 或期望损失 expected loss 被定义为联合分布$P(X, Y)$ 意义下的平均损失

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

联合分布是学习的目标. 一方面需要根据期望风险最小学习模型需要用到联合分布, 另一方面联合分布未知, 因此监督学习是病态问题 ill-formed problem.

因此引入经验风险 **empirical risk** 或经验损失 **empirical loss**

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

根据**大数定律**: 当N 趋向无穷时, 经验风险趋向于期望风险, 但是通常样本量难以满足, 因此采用**经验风险最小化** 和 **结构风险最小化** 对经验风险进行校正.

经验风险最小化 Empirical risk minimization, ERM

策略认为: 经验风险最小的模型就是最优模型. 在样本容量足够大时效果比较好, 极大似然估计是经验风险最小化的例子: 此时, 损失函数是**对数损失函数**, 模型是条件概率分布.

ERM 在小样本容易产生过拟合现象

结构风险最小化 Structural risk minimization, SRM

结构风险最小化等价于正则化, 在经验风险上面加入正则化项.

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

$J(f)$ 为模型复杂度.

例子: 贝叶斯估计中的最大后验概率估计(maximum posterior probability extimation, MAP)

模型条件概率分布, 损失函数对数损失函数, 模型复杂度由模型先验概率分布表示时, 结构风险最小化等价于最大后验概率估计.

算法

指模型的具体计算方法.

模型评估与模型选择

误差

误差分为训练误差和测试误差.

训练误差 training error 是关于训练数据集的平均损失

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

测试误差 test error 同理.

但是当损失函数为 0-1 损失是, 测试误差变成测试数据集上的误差率 error rate

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

I 为指示函数, indicator function. 当预测结果与标枪相等是为 1, 其他未 0

对应准确率 arruracy 为

$$r_{\text{text}} = \frac{1}{N'} \sum_{i=1}^{N'} I(y_i = \hat{f}(x_i))$$

模型选择相关

模型选择目的: 避免过拟合并且提高模型的预测能力

正则化

结构风险最小化策略的实现, 在经验风险上面加正则化项.

正则化项一般是模型复杂度的单调递增函数, 如, 模型参数向量的范数

交叉验证

数据集分为训练/验证/测试三个部分:

1. 简单交叉验证: 数据集随机划分为两个部分, 训练集和测试集, 训练集用于训练模型, 选在在测试集上误差最小的模型
2. S 折交叉验证: S-fold cross validation, 数据集划分为 S 个互不相交的子集, 用 S-1 个子集训练模型, 其余的子集测试模型. 重复进行 S 次, 选择平均测试误差最小的模型
3. 留一交叉验证 leave-one-out cross validation, S=N 是的 S 折交叉验证, 适用于数据缺乏情况

泛化能力评估

泛化能力 generalizationability 指模型对未知数据的预测能力. 定义为

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

泛化误差为模型的期望风险

文中给了一定理, 泛化误差和经验误差的关系:

$$R(f) \leq \hat{R}(f) + \epsilon(d, N, \delta)$$
$$\epsilon(d, N, \delta) = \sqrt{\frac{1}{2N}\left(\log d + log\frac{1}{\delta}\right)}$$

d为模型函数空间的容量 例如有 5 函数, $d = 5$, 有至少$1 - \sigma$的概率不等式成立

生成 vs 判别

生成模型由数据联合概率分布$P(X,Y)$, 然后求出条件概率分布 $P(Y|X)$ 作为模型的预测:

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

模型表示了输入和输出的生成关系, 所以为生成模型, 经典的有模型有: **朴素贝叶斯方法, 隐马尔科夫模型**

判别模型:由数据直接学习决策函数或者条件概率分布. 经典的模型有: **k 近邻法, 感知机, 决策树, logistics 回归模型, 最大熵模型, 支持向量机, 提升方法 和条件随机场等**

分类任务的指标

TP - 正类预测为正

FN - 正类预测为负

FP - 付类预测为正

TN - 负类预测为负

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

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

$$\frac{2}{F1}=\frac{1}{P}+\frac{1}{R}$$

习题

> *说明伯努利模型的极大似估计以及贝叶斯估计中的统计学习方法三要素. 伯努利模型是定义在 0 与 1 上的随机变量的概率分布, 假设观测到伯努利模型 n 次独立的数据生成结果, 其中 k 次结果为 1, 这时可以用极大似然估计或贝叶斯估计来估计结果为 1 的概率*

假设为 1 的概率为p

$$P(X=x)=(p)^x(1-p)^{1-x}$$

根据观测结果有

$$\sum_{i=0}^{n}X_i=k$$

极大似然估计

极大似然估计的策略师经验风险最小化

似然函数

$$L(x_1,x_2,\dots,x_n|p)=p^k(1-p)^{n-k}$$

取对数

$$\ln L=k\ln p+(n-k)ln(1-p)$$

对p求偏导

$$\frac{\partial \ln L}{\partial p}=\frac{k}{p}-\frac{n-k}{1-p}=\frac{k-np}{p(1-p)}$$

偏导为 0:

$$p=\frac{k}{n}$$

贝叶斯估计

贝叶斯估计的策略是结构风险最小化