CH01 统计学习及监督学习概论

前言

章节目录

1. 统计学习
2. 统计学习的分类
    a. 基本分类
    b. 按模型分类

    c. 按算法分类
    d. 按技巧分类
3. 统计学习三要素
    a. 模型
    b. 策略
    c. 算法
4. 模型评估与模型选择
    a. 训练误差与测试误差
    b. 过拟合与模型选择
5. 正则化与交叉验证
    a. 正则化
    b. 交叉验证
6. 泛化能力
    a. 泛化误差
    b. 泛化误差上界
7. 生成模型与判别模型
8. 监督学习应用
    a. 分类问题
    b. 标注问题
    c. 回归问题

导读

- ~~直接看目录结构，会感觉有点乱，就层级结构来讲感觉并不整齐。~~第二版重新梳理了这部分目录结构，舒服多了，尤其之前的分类，回归与标注因为出现了1.10造成目录中这部分不对齐，非常不爽。结果，第二版改了。赞
- 本章最后的三个部分，这三个问题可以对比着看，如果暂时没有概念，略过也可以，回头对各个算法有了感觉回头再看这里。
  这三部分怎么对比，三部分都有个图来说明，仔细看下差异，本文后面会对此展开。
- 关于损失函数，风险函数与目标函数注意体会差异
- 后面插点从深度学习角度拿到的点
    - 关于机器学习三要素, 复旦大学邱锡鹏教授也有解读[^2]: 模型, 学习准则, 优化算法. 这个定义比较接近代码. 以Tensorflow为例. 通常会定义一个网络(模型), 定义Loss(学习准则), 定义优化算法(Optimizer), 然后开Session, 不停的把数据带入用Opitmizer去最小化Loss.
    - Losses, Metrics, 在Keras里面划分了两个模块, 解释是Losses是BP过程用到的, 而Metrics实际和损失函数类似, 用来评价模型的性能, 但是不参与反向传播. 从源码也能看到, Metrics里面import了很多Loss算法
- 书中例子1.1可以参考PRML中对应的表述， 更详细些。
- 在监督学习中输入和输出对称为样本，在无监督学习中输入是样本。
- 注意在介绍输入空间，输出空间等概念的时候，以及这一章的很多部分都会有个帽子，监督学习中， 书中也明确了本书主要讨论监督学习的问题，最后的概要总结部分对监督学习有这样的描述：监督学习可以概括如下：从给定有限的训练数据出发，假设数据是独立同分布的，而且假设模型属于某个假设空间，应用某一评价准则，从假设空间中选取一个最优的模

型，使它对已给的训练数据以及未知测试数据在给定评价标准意义下有最准确的预测。，理解下这里的假设。

- 在贝叶斯学习部分，提到将模型、为观测要素及其参数用变量表示，使用模型的先验分布是贝叶斯学习的特点。注意这里面先验是**模型**的先验分布。
- 在泛化误差部分，用了$\hat{f}$表示最优估计，这个有时候也会用f^*表示意思差不多。有时候表示向量又要表示估计值，用$*$可能好看一点，比如$\vec{x}^*$，但是通常没本书都有自己的符号体系，向量可以通过字体表示，具体可以从书中的符号表部分了解。关于这一点，在第二版第一章就有所体现，监督和无监督学习中，模型用hat表示，在强化学习中，最优解用*表示。
- 提一下参考文献，几个大部头都在，ESL，PRML，DL，PGM，西瓜书，还有Sutton的强化学习，不过这本书2018年出了第二版，感兴趣的话可以看新版。

实现统计学习方法的步骤

统计学习方法三要素：模型，策略，算法

> 1. *得到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**，即学习模型的集合*
> 3. *确定模型选择的准则，即学习的**策略***
> 4. *实现求解最优模型的算法，即学习的**算法***
> 5. *通过学习方法选择最优的模型*
> 6. *利用学习的最优模型对新数据进行预测或分析*

统计学习分类

基本分类

这部分内容新增了无监督学习和强化学习。值得注意的一个点，之前因为只写了监督学习，样本表示(x, y)对，在无监督学习里面，样本就是x。

监督学习

无监督学习

强化学习

按模型分类

概率模型与非概率模型

在监督学习中，概率模型是生成模型，非概率模型是判别模型。

按算法分类

在线学习和批量学习，在线学习通常比批量学习更难。

按技巧分类

贝叶斯学习

核方法

统计学习方法三要素

模型

模型是什么？

在监督学习过程中，模型就是所要学习的**条件概率分布**或者**决策函数**。

注意书中的这部分描述，整理了一下到表格里：

	假设空间$\mathcal{F}$	输入空间$\mathcal{X}$	输出空间$\mathcal{Y}$	参数空间
决策函数	$\mathcal{F}=\{f_\theta \mid Y=f_\theta(x),\theta\in\mathbf{R}^n\}$	变量	变量	$\mathbf{R}^n$
条件概率分布	$\mathcal{F}=\{P \mid P_\theta(Y\mid X),\theta\in\mathbf{R}^n\}$	随机变量	随机变量	$\mathbf{R}^n$

书中描述的时候，有提到**条件概率分布族**，这个留一下，后面CH06有提到确认逻辑斯谛分布属于指数分布族。

策略

损失函数与风险函数

> *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预测的好坏。*

1. 损失函数(loss function)或代价函数(cost function)
   损失函数定义为给定输入X的**预测值**$f(X)$和**真实值**Y之间的**非负实值**函数，记作$L(Y,f(X))$

2. 风险函数(risk function)或期望损失(expected loss)
   这个和模型的泛化误差的形式是一样的
   $R_{exp}(f)=E_p[L(Y,f(X))]=\int_{\mathcal{X}\times\mathcal{Y}}L(y,f(x))P(x,y)\,\mathrm{d}x\mathrm{d}y$
   模型$f(X)$关于联合分布$P(X,Y)$的**平均意义下的**损失(**期望**损失)，但是因为$P(X,Y)$是未知的，所以前面的用词是**期望**，以及**平均意义下的**。

   这个表示其实就是损失的均值，反映了对整个数据的预测效果的好坏，P(x,y)转换成\frac {\nu(X=x, Y=y)}{N}$更容易直观理解, 可以参考CH09，6.2.2节的部分描述来理解，但是真实的数据N是无穷的。
3. 经验风险(empirical risk)或经验损失(empirical loss)
   $R_{emp}(f)=\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))$
   模型f关于**训练样本集**的平均损失
   根据大数定律，当样本容量N趋于无穷大时，经验风险趋于期望风险
4. 结构风险(structural risk)
   $R_{srm}(f)=\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))+\lambda J(f)$
   $J(f)$为模型复杂度, $\lambda\geqslant 0$是系数，用以权衡经验风险和模型复杂度。

常用损失函数

损失函数数值越小，模型就越好

1. 0-1损失
   $L(Y,f(X))=\begin{cases}1, Y\neq f(X)\\0, Y=f(X)\end{cases}$
2. 平方损失
   $L(Y,f(X))=(Y-f(X))^2$
3. 绝对损失
   $L(Y,f(X))=|Y-f(X)|$
4. 对数损失
   这里$P(Y|X)\leqslant 1$，对应的对数是负值，所以对数损失中包含一个负号，为什么不是绝对值？因为肯定是负的。
   $L(Y,P(Y|X))=-\log P(Y|X)$

ERM与SRM

经验风险最小化(ERM)与结构风险最小化(SRM)

1. **极大似然估计**是经验风险最小化的一个例子
   当模型是条件概率分布，损失函数是对数损失函数时，经验风险最小化等价于极大似然估计
2. **贝叶斯估计**中的**最大后验概率估计**是结构风险最小化的一个例子
   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

算法

这章里面简单提了一下，具体可以参考CH12表格中关于学习算法的描述。

模型评估与模型选择

训练误差和测试误差是模型关于数据集的平均损失。

提到一句，`统计学习方法具体采用的损失函数未必是评估时使用的损失函数`，这句理解下。参考下在数据科学比赛中给出的评分标准，与实际学习采用的损失函数之间的关系。

过拟合与模型选择

这部分讲到了最小二乘法，给了PRML中的一个例子。

这个问题中**训练数据**为$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$

模型为

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

经验风险最小化策略下

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

将模型和训练数据带入到上式得到

$$L(w) = \frac{1}{2} \sum_{i=1}^{N} \left(\sum_{j=0}^{M} w_j x_i^j - y_i \right)^2 = \frac{1}{2} \sum_{i=1}^{N} (w \cdot x_i - y_i)^2$$

这个问题要求$w = (w_0^*, w_1^*, \cdots, w_M^*)$

对w求偏导令其为零，得到一系列方程，求解可以用梯度下降或者矩阵分解。

求解线性方程组$Ax = b$，可以表示为$x = A/b$，问题展开之后可以涉及到**矩阵分解**。

TODO: 这个例子展开一下

正则化与交叉验证

正则化

模型选择的典型方法是正则化

交叉验证

另一种常用的模型选择方法是交叉验证

- 简单
- S折(K折, K-Fold)[1]
- 留一法

关于交叉验证，这里补充一点。

数据集的划分这个问题，书中有提到数据充足的情况下，将数据划分为三个部分，训练集，验证集和测试集。看到这里，不知道大家会不会有一样的问题：**验证集和测试集有什么区别？**

注意这里，在算法学习的过程中，测试集可能是固定的，但是验证集和训练集可能是变化的。比如K折交叉验证的情况下，分成K折之后，其中的K-1折作为训练集，1折作为验证集，这样针对每一个模型操作K次，计算平均测试误差，最后选择平均测试误差最小的模型。这个过程中用来验证模型效果的那一折数据就是**验证集**。交叉**验证**，就是这样一个使用**验证**集测试模型好坏的过程。他允许我们在模型选择的过程中，使用一部分数据（验证集）“偷窥”一下模型的效果。

泛化能力

- 现实中采用最多的方法是通过测试误差来评价学习方法的泛化能力
- 统计学习理论试图从理论上对学习方法的泛化能力进行分析
- 学习方法的泛化能力往往是通过研究泛化误差的**概率上界**进行的, 简称为泛化误差上界(generalization error bound)

  这本书里面讨论的不多，在CH08里面有讨论提升方法的误差分析, 提到$AdaBoost$不需要知道下界γ。在CH02中讨论算法的收敛性的时候有提到误分类次数的上界.

注意泛化误差的定义，书中有说**事实上，泛化误差就是所学习到的模型的期望风险**

生成模型与判别模型

监督学习方法可分为**生成方法**(generative approach)与**判别方法**(discriminative approach)

生成方法

generative approach

- 可以还原出**联合概率分布**$P(X,Y)$
- 收敛速度快, 当样本容量增加时, 学到的模型可以更快收敛到真实模型
- 当存在隐变量时仍可以用

判别方法

discriminative approach

- 直接学习**条件概率**$P(Y|X)$或者**决策函数**$f(X)$
- 直接面对预测, 往往学习准确率更高
- 可以对数据进行各种程度的抽象, 定义特征并使用特征, 可以简化学习问题

分类问题、标注问题、回归问题

Classification, Tagging, Regression

- 图1.4和图1.5除了分类系统和标注系统的差异外，没看到其他差异，但实际上这两幅图中对应的输入数据有差异，序列数据的 $x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)})^T$ 对应了
- 图1.5和图1.6，回归问题的产出为 $Y = \hat{f}(X)$

参考

参考文献都是大部头，ESL，PRML在列

1.
2.

↑ top

---

1. ESL:7.10.1:K-Forld Cross Validation↩