统计学习方法

李航 著

- 第1章 统计学习方法概论
  - 1.1 统计学习
  - 1.2 监督学习
  - 1.3 统计学习三要素
    - 1.3.1 模型
      - 非概率模型：$f=F_{\theta\theta}$
      - 概率模型：$f=P(y|x)$
    - 1.3.2 策略
      - 1．损失函数和风险函数
        - （1）0-1损失函数（0-1 loss function）

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

        - （2）平方损失函数（quadratic loss function）

$$L(Y,f(X))=(Y-f(X))^2$$

        - （3）绝对损失函数（absolute loss function）

$$L(Y,f(X))=|Y-f(X)|$$

        - （4）对数损失函数（logarithmic loss function）或对数似然损失函数（loglikelihood loss function）

$$L(Y,P(Y|X))=-\log P(Y|X) \quad (1.8)$$

        - 这是理论上模型f(X)关于联合分布P(X,Y)的平均意义下的损失，称为风险函数（risk function）或期望损失（expected loss）。

$$R_{\exp}(f)=E_P[L(Y,f(X))]=\int_{X\times Y}L(y,f(x))P(x,y)\mathrm{d}x\mathrm{d}y \quad (1.9)$$

        - 模型f(X)关于训练数据集的平均损失称为经验风险（empirical risk）或经验损失（empirical loss），记作R_{emp}：

$$R_{emp}(f)=\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i)) \quad (1.10)$$

      - 2．经验风险最小化与结构风险最小化
        - $\min_{f\in\mathcal{F}}\ \frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))$
        - 用。比如，极大似然估计（maximum likelihood estimation）就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，经验风险最小化就等价于极大似然估计。
          但是，当样本容量很小时，经验风险最小化学习的效果就未必很好，会产生后面将要叙述的“过拟合(over-fitting)”现象。
        - $\min_{f\in\mathcal{F}}\ \frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))+\lambda J(f)$
        - 结构风险最小化（structural risk minimization，SRM）是为了防止过拟合而提出来的策略。结构风险最小化等价于正则化（regularization）。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）或罚项（penalty term）。在假设空间、损失函数以及训练数据集确定的情况下，结构风险的定义是
    - 1.3.3 算法
  - 1.4 模型评估与模型选择
    - 1.4.1 训练误差与测试误差
    - 1.4.2 过拟合与模型选择
  - 1.5 正则化与交叉验证
    - 1.5.1 正则化
      - $\min_{f\in\mathcal{F}}\ \frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))+\lambda J(f)$
      - 以是参数向量的L_2范数：

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

    - 1.5.2 交叉验证
      - 1．简单交叉验证
      - 2．S折交叉验证
      - 3．留一交叉验证
  - 1.6 泛化能力
    - 1.6.1 泛化误差
      - 首先给出泛化误差的定义。如果学到的模型是$\hat{f}$，那么用这个模型对未知数据预测的误差即为泛化误差（generalization error）

$$R_{\exp}(\hat{f})=E_P[L(Y,\hat{f}(X))]=\int_{X\times Y}L(y,\hat{f}(x))P(x,y)\mathrm{d}x\mathrm{d}y \quad (1.20)$$

      - 泛化误差反映了学习方法的泛化能力，如果一种方法学习的模型比另一种方法学习的模型具有更小的泛化误差，那么这种方法就更有效。事实上，泛化误差就是所学习到的模型的期望风险。
    - 1.6.2 泛化误差上界
      - 定理**1.1**（泛化误差上界） 对二类分类问题，当假设空间是有限个函数的集合$\mathcal{F}=\{f_1,f_2,\cdots,f_d\}$时，对任意一个函数$f\in\mathcal{F}$，至少以概率$1-\delta$，以下不等式成立：

$$R(f)\leqslant\hat{R}(f)+\varepsilon(d,N,\delta) \quad (1.25)$$

        其中，

$$\varepsilon(d,N,\delta)=\sqrt{\frac{1}{2N}\left(\log d+\log\frac{1}{\delta}\right)} \quad (1.26)$$

  - 1.7 生成模型与判别模型
    - 监督学习方法又可以分为生成方法（generative approach）和判别方法（discriminative approach）。所学到的模型分别称为生成模型（generative model）和判别模型（discriminative model）。
    - 生成方法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，即生成模型：

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

      这样的方法之所以称为生成方法，是因为模型表示了给定输入X产生输出Y的生成关系。典型的生成模型有：朴素贝叶斯法和隐马尔可夫模型，将在后面章节进行相关讲述。
    - 判别方法由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。判别方法关心的是对给定的输入X，应该预测什么样的输出Y。典型的判别模型包括：k近邻法、感知机、决策树、逻辑斯谛回归模型、最大熵模型、支持向量机、提升方法和条件随机场等，将在后面章节讲述。
    - 生成方法的特点：生成方法可以还原出联合概率分布P(X,Y)，而判别方法则不能；生成方法的学习收敛速度更快，即当样本容量增加的时候，学到的模型可以更快地收敛于真实模型；当存在隐变量时，仍可以用生成方法学习，此时判别方法就不能用。
      判别方法的特点：判别方法直接学习的是条件概率P(Y|X)或决策函数f(X)，直接面对预测，往往学习的准确率更高；由于直接学习P(Y|X)或f(X)，可以对数据进行各种程度上的抽象、定义特征并使用特征，因此可以简化学习问题。
  - 1.8 分类问题
    - 对于二类分类问题常用的评价指标是精确率（precision）与召回率（recall）。通常以关注的类为正类，其他类为负类，分类器在测试数据集上的预测或正确或不正确，4种情况出现的总数分别记作：
    - TP——将正类预测为正类数；
      FN——将正类预测为负类数；
      FP——将负类预测为正类数；
      TN——将负类预测为负类数。
      精确率定义为

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

      召回率定义为

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

      此外，还有1F值，是精确率和召回率的调和均值，即

$$\frac{2}{F_1}=\frac{1}{P}+\frac{1}{R}$$

$$F_1=\frac{2TP}{2TP+FP+FN}$$

    - 许多统计学习方法可以用于分类，包括k近邻法、感知机、朴素贝叶斯法、决策树、决策列表、逻辑斯谛回归模型、支持向量机、提升方法、贝叶斯网络、神经网络、Winnow等。本书将讲述其中一些主要方法。
  - 1.9 标注问题
    - 标注问题分为学习和标注两个过程（如图1.5所示）。首先给定一个训练数据集

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

      这里，$x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)})^{\mathrm{T}}$，i=1,2,…,N，是输入观测序列，
      $y_i=(y_i^{(1)},y_i^{(2)},\cdots,y_i^{(n)})^{\mathrm{T}}$是相应的输出标记序列，n是序列的长度，对不同样本可以有不同的值。学习系统基于训练数据集构建一个模型，表示为条件概率分布：

$$P(Y^{(1)},Y^{(2)},\cdots,Y^{(n)}\,|\,X^{(1)},X^{(2)},\cdots,X^{(n)})$$

      这里，每一个$X^{(i)}$（i=1,2,…,n）取值为所有可能的观测，每一个$Y^{(i)}$（i=1,2,…,n）取值为所有可能的标记，一般n=N。标注系统按照学习得到的条件概率分布模型，对新的输入观测序列找到相应的输出标记序列。具体地，对一个观测序列$x_{N+1}=(x_{N+1}^{(1)},x_{N+1}^{(2)},\cdots,x_{N+1}^{(n)})^{\mathrm{T}}$找到使条件概率$P((y_{N+1}^{(1)},y_{N+1}^{(2)},\cdots,y_{N+1}^{(n)})^{\mathrm{T}}\,|\,(x_{N+1}^{(1)},x_{N+1}^{(2)},\cdots,x_{N+1}^{(n)})^{\mathrm{T}})$最大的标记序列
      $y_{N+1}=(y_{N+1}^{(1)},y_{N+1}^{(2)},\cdots,y_{N+1}^{(n)})^{\mathrm{T}}$。
    - 标注常用的统计学习方法有：隐马尔可夫模型、条件随机场。
  - 1.10 回归问题
  - 习题
    - 1.1 说明伯努利模型的极大似然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取值为0与1的随机变量上的概率分布。假设观测到伯努利模型n次独立的数据生成结果，其中k次的结果为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。
    - 1.2 通过经验风险最小化推导极大似然估计。证明模型是条件概率分布，当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。