统计机器学习

*李航 《统计学习方法》

推荐书籍

- 机器学习　　周志华
- 统计学习方法　　李航
- 深度学习　　Lan Goodfellow等

- 作业提交：ketangpai.com 加课码：VN46H6

统计学习

- 统计学习是关于计算机基于数据构建概率统计模型并运用模型对数据进行预测与分析的一门学科。统计学习也称为统计机器学习。
- 统计学习的对象
  - ✓ 数据：计算机及互联网上的各种数字、文字、图像、视频、音频数据以及它们的组合。
  - ✓ 数据的基本假设是同类数据具有一定的统计规律性。
- 统计学习的目的
  - ✓ 用于对数据（特别是未知数据）进行预测和分析。
  - ✓ 从数据出发，提取数据的特征，抽象出数据的模型，发现数据中的知识，又回到对数据的分析与预测中。

统计学习的方法

- 监督学习　　Supervised learning
- 非监督学习　Unsupervised learning
- 半监督学习　Semi-supervised learning
- 强化学习　　Reinforcement learning

统计学习方法的步骤

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. 确定模型的选择的准则，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模型
6. 利用学习的最优模型对新数据进行预测或分析

监督学习

- 监督学习的任务
  - ✓ 学习一个模型，使模型能够对任意给定的输入，对其相应的输出做出一个好的预测。
- 将输入与输出所有可能取值的集合分别称为输入空间与输出空间。
- 每个具体的输入是一个实例（instance），通常由特征向量（feature vector）表示。
- 所有特征向量存在的空间称为特征空间(feature space)
- 在监督学习过程中，将输入与输出看作是定义在输入（特征）空间与输出空间上的随机变量的取值。

监督学习

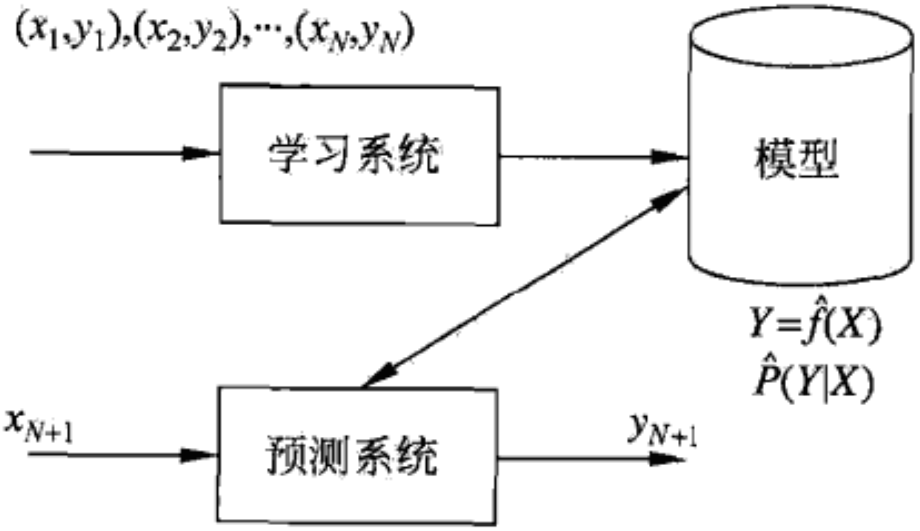
- 输入变量X和输出变量Y有不同的类型，可以是连续的，也可以是离散的。
- 回归问题
  - ✓输入变量与输出变量均为连续变量的预测问题。
- 分类问题
  - ✓输出变量为有限个离散变量的预测问题。
- 标注问题
  - ✓输入变量与输出变量均为变量序列的预测问题。

监督学习

- 联合概率分布
  - ✓ 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)为分布函数或分布密度函数
  - ✓ 对于学习系统来说，联合概率分布是未知的，
  - ✓ 训练数据和测试数据被看作是依联合概率分布P(X, Y)独立同分布产生的。
- 假设空间
  - ✓ 监督学习目的是学习一个由输入到输出的映射，称为模型
  - ✓ 模型的集合就是假设空间（hypothesis space）
  - ✓ 概率模型或非概率模型：条件概率分布P(Y|X)，决策函数：Y=f(X)

监督学习

- 问题的形式化



$$y_{N+1} = \arg\max_{y_{N+1}} \hat{P}(y_{N+1} \mid x_{N+1})$$

$$y_{N+1} = \hat{f}(x_{N+1})$$

统计学习三要素

方法=模型+策略+算法

- 模型
  - ✓在监督学习过程中，模型就是所要学习的条件概率分布或决策函数
- 策略
  - ✓按照什么样的准则学习或者选择最优的模型。
  - ✓统计学习的目标在于从假设空间中选取最优模型。
- 算法
  - ✓学习模型的具体计算方法。
- 统计学习基于训练数据集，根于学习策略，从假设空间中选择最优模型，最后需要考虑用什么样的计算方法求解最优模型。

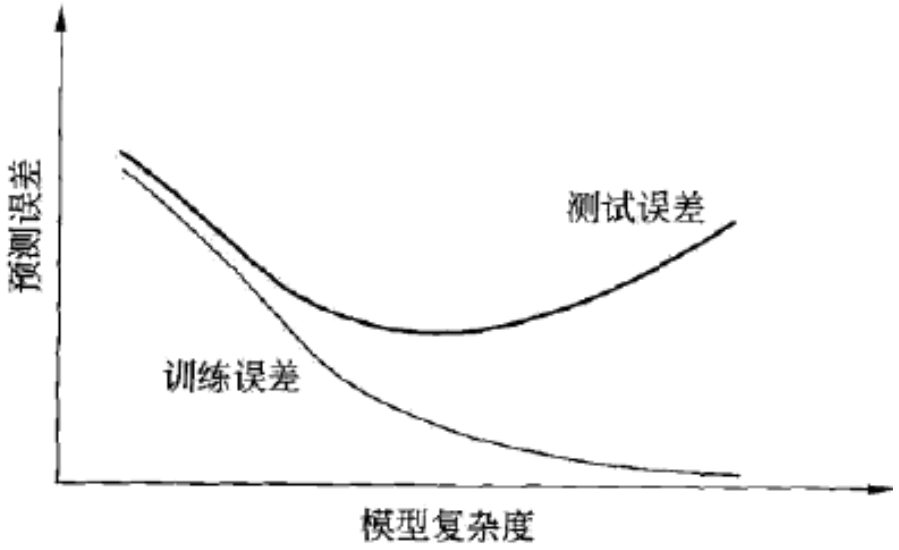
模型评估

- 统计学习的目的是使学到的模型不仅对已知数据而且对未知数据都能有很好的预测能力。
- 当损失函数给定时，基于损失函数的模型的训练误差和模型的测试误差就成为学习方法评估的标准。

- 测试数据集上误差率 $$e_{\text{test}} = \frac{1}{N'} \sum_{i=1}^{N'} I(y_i \neq \hat{f}(x_i))$$
- 测试数据集上的准确率 $$r_{\text{test}} = \frac{1}{N'} \sum_{i=1}^{N'} I(y_i = \hat{f}(x_i))$$

过拟合与模型选择

- 过拟合
  - ✓ 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得很好，但对未知数据预测得很差的现象。
- 模型选择旨在避免过拟合并提高模型的预测能力。



正则化、泛化能力

- 正则化
  - ✓ 模型选择的典型方法
  - ✓ 是结构风险最小化策略的实现，是在经验风险上加一个正则化项或惩罚项。
  - ✓ 正则化的一般形式：

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

- 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，是学习方法本质上重要的性质。
  - ✓ 通过测试误差来评价学习方法的泛化能力。
  - ✓ 模型对未知数据预测得误差即泛化误差，是所学习到的模型的期望风险。

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

生成模型与判别模型

- 监督学习的任务就是学习一个模型，应用这个模型，对给定的输入预测相应的输出。这个模型的一般形式为：

  决策函数 $Y = f(X)$

  或 条件概率分布 $P(Y \mid X)$

- 监督学习方法又可以分为生成方法和判别方法，所学到的模型称为生成模型和判别模型。

生成模型与判别模型

- 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：

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

- 为什么称为生成方法？
  - ✓ 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、隐马尔科夫模型

生成模型与判别模型

- 判别方法由数据直接学习决策函数f(X)或条件概率分布P(Y|X)作为预测的模型，即判别模型。

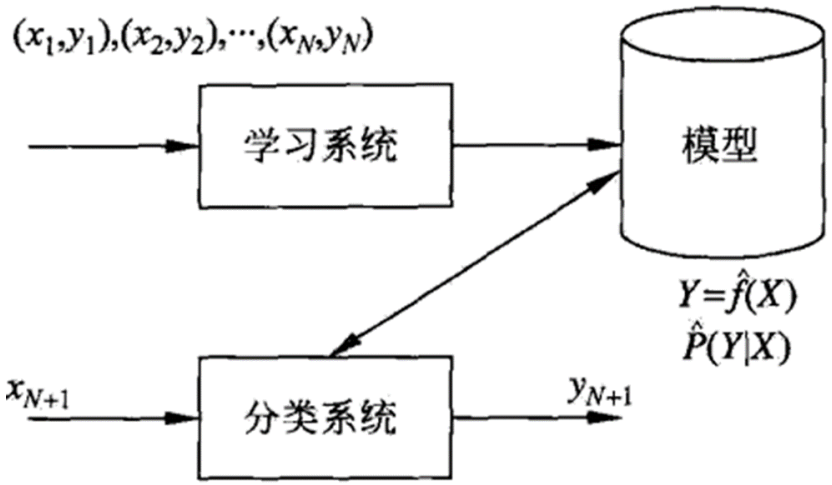
- 判别方法关心的是对给定的输入X，应该预测什么样的输出Y。
- 典型的判别模型：K近邻法、感知机、决策树、逻辑斯蒂回归模型、最大熵模型、支持向量机、条件随机场等

生成模型与判别模型

- 生成方法的特点：可还原出联合概率分布P(X, Y)，而判别方法不能。生成方法的收敛速度更快，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，仍可以使用生成方法，而判别方法则不能用。
- 判别方法的特点：直接学习到条件概率或决策函数，直接进行预测，往往学习的准确率更高；由于直接学习Y=f(X)或P(Y|X)，可对数据进行各种程度上的抽象、定义特征并使用特征，因此可以简化学习过程。

分类问题

- 在监督学习中，当输出变量Y取有限个离散值时，预测问题便成为分类问题。
- 监督学习从数据中学习一个分类模型或分类决策函数，称为分类器。
- 分类器对新的输入进行输出的预测，称为分类，可能的输出称为类，分类的类别为多个时，称为多分类问题。



分类问题

- 二分类评价指标

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 $P = \frac{TP}{TP + FP}$
- 召回率 $R = \frac{TP}{TP + FN}$
- F_1值 $\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R}$

- 常用于分类的统计学习方法：K近邻法、感知机、朴素贝叶斯法、决策树、支持向量机、神经网络等

标注问题

- 标注问题的输入是一个观测序列，输出是一个标记序列或状态序列。
- 标注问题的目标在于学习一个模型，使它能够对观测序列给出标记序列作为预测。
- 标注问题分为学习和标注两个过程。
- 常用方法：隐马尔科夫模型、条件随机场

回归问题

- 回归用于预测输入变量（自变量）和输出变量（因变量）之间的关系，特别是当输入变量的值发生变化时，输出变量的值随之发生的变化。
- 回归模型正是表示从输入变量到输出变量之间映射的函数。
- 回归问题的学习等价于函数拟合：选择一条函数曲线使其很好地拟合已知数据且很好地预测未知数据。
- 回归问题：股价预测问题、市场趋势预测、投资风险分析等。