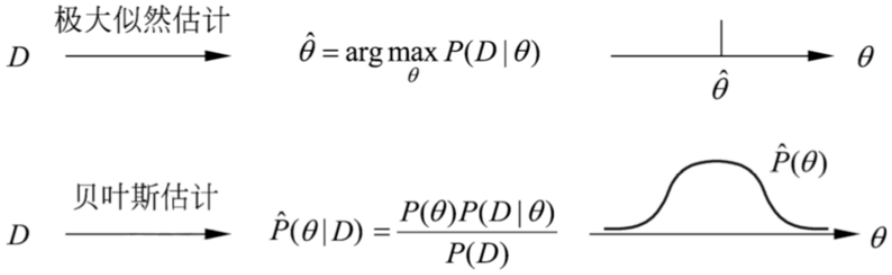
统计学习方法(第二版)-总结

一、第一章 统计方法概论

1. 统计学习基本概念

- 统计学习的对象：
  - 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。
- 统计学习方法的分类：
  - **基本分类：**
    - 监督学习 (Supervised learning)
    - 无监督学习 (Unsupervised learning)
    - 半监督学习 (Semi-supervised learning)
    - 强化学习 (Reinforcement learning)
    - 主动学习 (active learning)
  - **按照模型分类**：(其中x是输入，y是输出)
    - **概率模型** (`生成模型`)：
      - 在监督学习中，取条件概率分布形式$P(y \mid x)$
      - 在无监督学习中，取条件概率分布形式 $P(z \mid x)$ 或 $P(x \mid z)$
      - 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。
    - **非概率模型** (`判别模型`)：
      - 在监督学习中，取函数形式 $y = f(x)$
      - 在无监督学习中，取函数形式 $z = g(x)$
      - 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。
    - 线性模型： 当$y = f(x)$ 或 $z = g(x)$ 是线性函数。
    - 非线性模型：上一条反之即可。
    - 参数化模型：假设模型参数的维度固定，模型可以由有限维参数完全刻画。
    - 非参数化模型：假设模型参数的维度不固定或者说无穷大，随着训练数据量的增加而不断增大。
  - **按照算法分类：**
    - **在线学习** (Online learning)：指每次接受一个样本，进行预测，之后学习模型，并不断重复该操作的机器学习。
    - **批量学习** (Batch learning)：一次接受所有数据，学习模型，之后进行预测。
  - **按照技巧分类：**
    - **贝叶斯学习**：



    - **核方法**:
        - 使用核函数表示和学习非线性模型， 将线性模型学习方法扩展到非线性模型的学习
        - 不显式地定义输入空间到特征空间的映射，而是直接定义核函数, 即映射之后在特征空间 的内积
        - 假设 x_1, x_2 是输入空间的任意两个实例，内积为 $< \mathrm{x}_1, \mathrm{x}_2 >$, 输入空间到特征空间的映射为$\varphi$. 核方法在输入空间中定义核函数 $\mathrm{K}\left(\mathrm{x}_1, \mathrm{x}_2\right)$, 使其满足 $\mathrm{K}\left(\mathrm{x}_1, \mathrm{x}_2\right)=<\varphi\left(\mathrm{x}_1\right), \varphi\left(\mathrm{x}_2\right)>$
- 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**或者**决策函数**
    - **策略**:
        - **损失函数**：度量模型一次预测的好坏。常用损失函数：
            - **0-1损失函数:**

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

            - **平方损失函数:**

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

            - **绝对损失函数:**

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

            - **对数损失函数:**

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

            - **指数损失函数**:

$$L(Y \mid f(X))=\exp[-y f(x)]$$

            - **Hinge 损失函数**:

$$L(y, f(X))=\max(0,1-y f(X))$$

        - **交叉熵损失函数**：x为样本，y为标签，a为预测结果，n表示样本总数

$$C=-\frac{1}{n}\sum_x[y\ln a+(1-y)\ln(1-a)]$$

        - **风险函数**：模型 $f(X)$关于联合分布 $P(X,Y)$ 的平均意义下的损失。

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

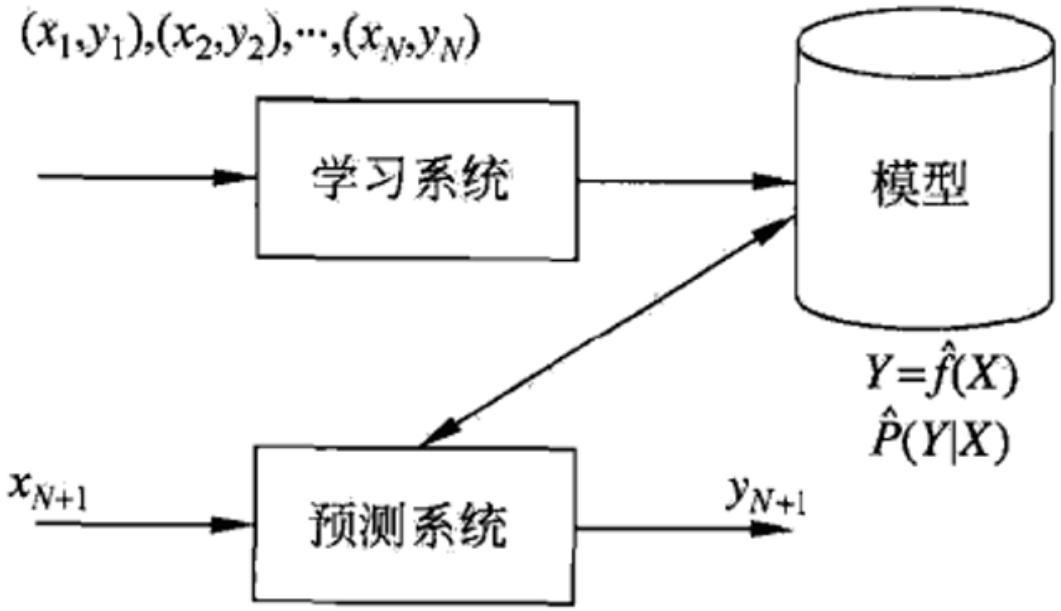
            - **经验风险最小化：**`(ERM)`
                - **极大似然估计**是经验风险最小化的一个例子。当模型是条件概率分布，损失函数是对数损失函数时，经验风险最小化等价于极大似然估计。
            - **结构风险最小化：**`(SRM)`
                - 等价于**正则化**；
                - **贝叶斯估计**中的**最大后验概率估计**是结构风险最小化的一个例子。当模型是条件概率分布，损失函数是对数损失函数，**模型复杂度由模型的先验概率表示**时，结构风险最小化等价于最大后验概率估计。
    - **算法**：
        - 指学习模型的具体计算方法。统计学习基于训练数据集，根据学习策略，从假设空间中选择最优模型，最后需要考虑用什么杨的计算方法来求解最优模型。

2. 统计学习方法的实现步骤：

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。

3. 监督学习

- 定义：指从`标注数据`中学习预测模型的机器学习问题，其本质是学习`输入到输出的映射`的统计规律。
- 输入变量和输出变量：
    - 分类问题：输出变量为有限个离散变量的预测问题。
    - 回归问题：输入与输出均为连续变量的预测问题。
    - 标注问题：输入与输出变量均为变量序列的预测问题。

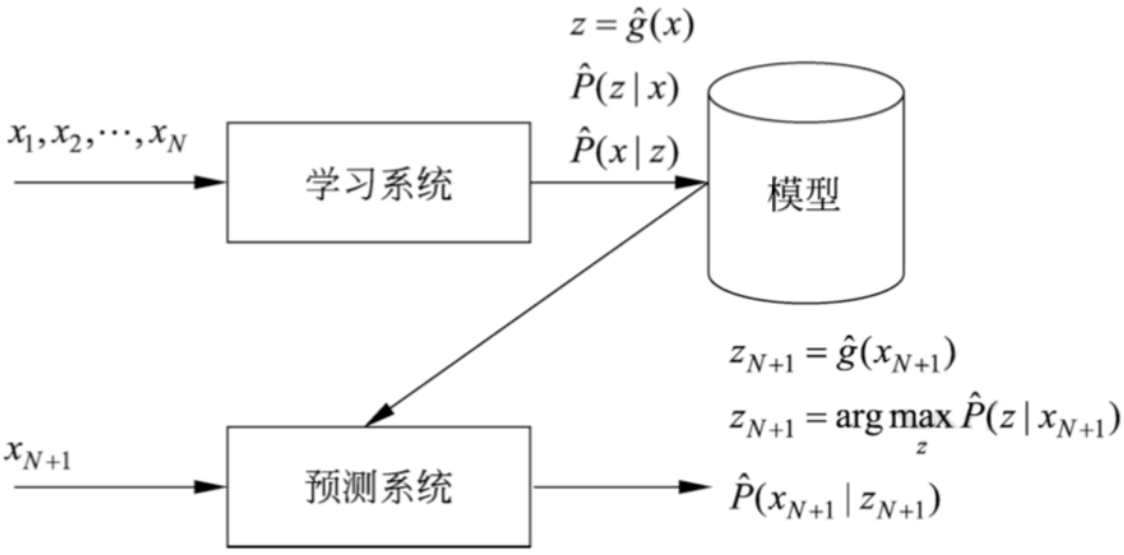


$$y_{N+1} = \arg\max_{y_{N+1}} \hat{P}(y_{N+1} \mid x_{N+1})$$

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

4. 无监督学习

- 定义：指从无标注数据中学习预测模型的机器学习问题，其本质是学习数据中的统计规律或内在结构。
- 旨在从假设空间中选出在给定评价标准下的最优模型，模型可以实现对数据的聚类、降维或是概率估计。



5. 半监督学习

- 定义：指利用**标注**数据和**未标注**数据学习预测模型地机器学习问题。
- 利用未标注数据中的信息，辅助标注数据，进行监督学习，
- 以较低的成本达到较好的学习效果。

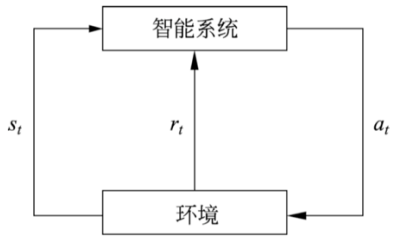
6. 强化学习

- 定义：指智能系统在与环境的连续互动中学习最优行为策略的机器学习问题，其本质是学习最优的序贯决策。
- 智能系统的目标不是短期奖励的最大化，而是长期累积奖励的最大化。
- 强化学习过程中，系统不断地试错，以达到学习最优策略地目的。

强化学习的马尔可夫决策过程是状态、奖励、动作序列上的随机过程，由五元组 $\langle S, A, P, r, \gamma \rangle$ 组成。

- S 是有限状态（state）的集合
- A 是有限动作（action）的集合
- P 是状态转移概率（transition probability）函数:

$$P(s'|s,a) = P(s_{t+1} = s'|s_t = s, a_t = a)$$



- r 是奖励函数（reward function）：$r(s,a) = E(r_{t+1}|s_t = s, a_t = a)$
- γ 是衰减系数（discount factor）：$\gamma \in [0,1]$

7. 主动学习

- 定义：指机器不断主动给出实例让教师进行标注，然后利用标注数据学习预测模型的机器学习问题。
- 目标是找出对学习最有帮助的实例让教师标注，以较小的标注代价达到较好的学习效果。

8. 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。
- 注意：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。
- **过拟合**：指学习时选择的模型所包含的参数过多，以至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。可以说模型选择旨在避免过拟合并提高模型的预测能力。

9. 正则化、交叉验证、泛化能力

- **正则化**：
  - 模型选择的典型方法是**正则化**.。正则化项一般是模型复杂度的单调递增函数，模型越复杂，正则化就越大。比如，正则化项可以是模型参数向量的范数。
  - $$L(w) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i;w) - y_i)^2 + \frac{\lambda}{2}|w|^2$$
  - $|w|$表示向量w的L_2范数
  - 正则化符合奥卡姆剃刀原理：如无必要，勿增实体。在应用于模型选择中时，可以理解为：在所有可能选择的模型中，能够很好地解释已知数据并且十分简单的才是最好的模型，也是应该选择的模型。
- **交叉验证**：
  - 的基本想法时重复地利用数据；把给定地数据进行切分，将切分的数据集组合为训练集和测试集，在此基础上反复地进行训练、测试以及模型选择。
  - 主要有简单交叉验证，S折交叉验证，留一交叉验证三种。
  - 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S折交叉验证的情况下，分成S折之后，其中的S-1折作为训练集，1折作为验证集，计算这S个模型

每个模型的平均测试误差，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。

- **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。
  - 多通过**测试误差**来评价学习方法的泛化能力。但这种评价是依赖于测试数据集的。但测试数据集是有限的，评价结果是不可靠性大。
  - 业内试图从理论上对学习方法的泛化能力进行分析。
  - **泛化误差**：
    - 定义：若学到的模型是 $\hat{f}$，那么用这个模型对未知数据预测的误差；

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

    - 泛化误差就是所学习到的**模型的期望风险**。