七、贝叶斯分类器

主讲教师：周志华

贝叶斯决策论 (Bayesian decision theory)

概率框架下实施决策的基本理论

给定 N 个类别，令 λ_{ij} 代表将第 j 类样本误分类为第 i 类所产生的损失，则基于后验概率将样本 $\boldsymbol{x}$ 分到第 i 类的条件风险为：

$$R(c_i \mid \boldsymbol{x}) = \sum_{j=1}^{N} \lambda_{ij} P(c_j \mid \boldsymbol{x})$$

贝叶斯判定准则 (Bayes decision rule)：

$$h^*(\boldsymbol{x}) = \underset{c \in \mathcal{Y}}{\arg\min}\, R(c \mid \boldsymbol{x})$$

- h^* 称为贝叶斯最优分类器(Bayes optimal classifier)，其总体风险称为贝叶斯风险 (Bayes risk)
- 反映了学习性能的理论上限

判别式 vs. 生成式

$P(c \mid \boldsymbol{x})$ 在现实中通常难以直接获得

从这个角度来看，机器学习所要实现的是基于有限的训练样本尽可能准确地估计出后验概率

两种基本策略：

判别式 (discriminative) 模型

思路：直接对 $P(c \mid \boldsymbol{x})$ 建模

代表：
- 决策树
- BP 神经网络
- SVM

生成式 (generative) 模型

思路：先对联合概率分布 $P(\boldsymbol{x}, c)$ 建模，再由此获得 $P(c \mid \boldsymbol{x})$

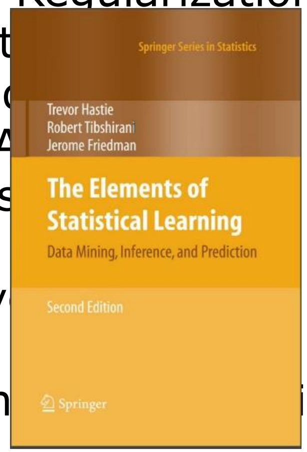
$$P(c \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x}, c)}{P(\boldsymbol{x})}$$

代表：贝叶斯分类器

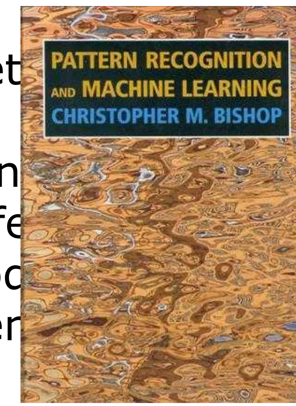
注意：贝叶斯分类器 ≠ 贝叶斯学习 (Bayesian learning)

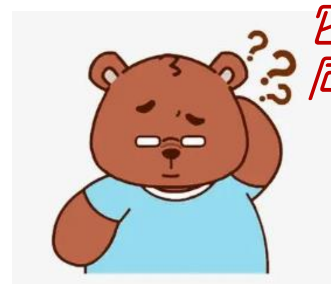
机器学习中不同学派的视野差别极大，以两本名著的目录为例

1. Introduction
2. Overview of Supervised Learning
3. Linear Methods for Regression
4. Linear Methods for Classification
5. Basic Expansions and Regularization
6. Kernel Smoothing Met
7. Model Assessment and
8. Model Inference and A
9. Additive Models, Trees
Methods
10. Boosting and Additiv
11. Neural Networks
12. Support Vector Mach ible
Discriminants
13. Prototype Methods and Nearest-Neighbors
14. Unsupervised Learning
15. Random Forests
16. Ensemble Learning
17. Undirected Graphical Models
18. High-dimensional problems



1. Introduction
2. Probability Distributions
3. Linear Models for Regression
4. Linear Models for Classification
5. Neural Networks
6. Kernel Methods
7. Sparse Kernel Met
8. Graphical Models
9. Mixture Models an
10. Approximate Infe
11. Sampling Methoc
12. Continuous Later
13. Sequential Data
14. Combining Models





它们真的是在介绍同一个学科领域？

贝叶斯定理

Thomas Bayes
(1701?-1761)

$$P(c \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x}, c)}{P(\boldsymbol{x})}$$

根据贝叶斯定理，有

$$P(c \mid \boldsymbol{x}) = \frac{\boxed{P(c)}\,\boxed{P(\boldsymbol{x} \mid c)}}{\boxed{P(\boldsymbol{x})}}$$

先验概率（prior）
样本空间中各类样本所占的比例，可通过各类样本出现的频率估计（大数定律）

证据（evidence）因子，与类别无关

样本相对于类标记的类条件概率（class-conditional probability），亦称 似然 (likelihood)

主要困难在于估计似然

$$P(\boldsymbol{x} \mid c)$$

极大似然估计

先假设某种概率分布形式，再基于训练样例对参数进行估计

假定 $P(\boldsymbol{x} \mid c)$具有确定的概率分布形式，且被参数 $\boldsymbol{\theta}_c$ 唯一确定，则任务就是利用训练集 D 来估计参数 $\boldsymbol{\theta}_c$

$\boldsymbol{\theta}_c$ 对于训练集 D 中第 c 类样本组成的集合 D_c 的似然(likelihood)为

$$P(D_c \mid \boldsymbol{\theta}_c) = \prod_{\boldsymbol{x} \in D_c} P(\boldsymbol{x} \mid \boldsymbol{\theta}_c)$$

连乘易造成下溢，因此通常使用对数似然 (log-likelihood)

$$LL(\boldsymbol{\theta}_c) = \log\ P(D_c \mid \boldsymbol{\theta}_c) = \sum_{\boldsymbol{x} \in D_c} \log P(\boldsymbol{x} \mid \boldsymbol{\theta}_c)$$

于是，$\boldsymbol{\theta}_c$ 的极大似然估计为 $\hat{\boldsymbol{\theta}}_c = \underset{\boldsymbol{\theta}_c}{\arg\max}\ LL(\boldsymbol{\theta}_c)$

估计结果的准确性严重依赖于所假设的概率分布形式是否符合潜在的真实分布

To be continued