机器学习

4. 朴素贝叶斯

- 把分类回归问题用统计框架描述

主要内容

- 贝叶斯分类器
- NB基本原理
- MLE vs. MAP
- 垃圾邮件分类
- Bag of Words
- 字符识别

主要内容

- 贝叶斯分类器
- NB基本原理
- MLE vs. MAP
- 垃圾邮件分类
- Bag of Words
- 字符识别

贝叶斯分类器

- 分类问题目标：

学习预测函数f: Ω → {0,1}，使得某个风险函数(表现度量) R(f)在某个学习机器上达到最小。

X

体育
娱乐
科学
。。。

Y

概率误差：$R(f) = P(f(X) \neq Y)$

贝叶斯分类器

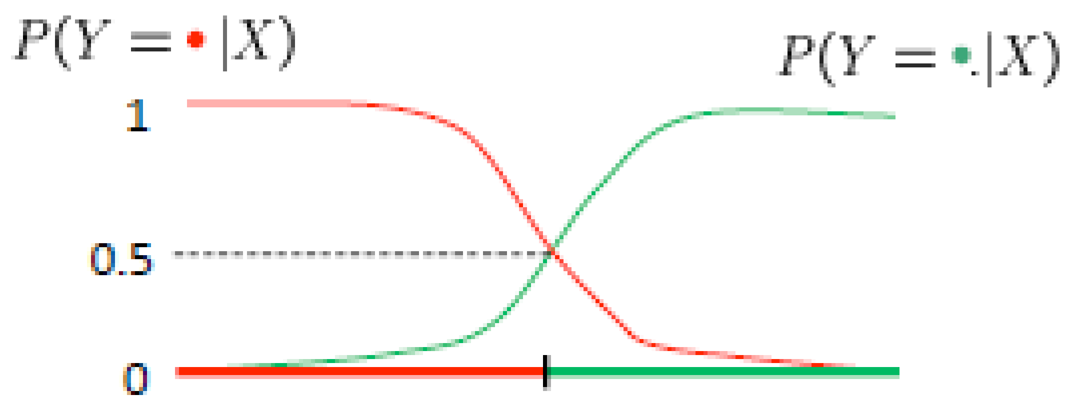
- 最优决策器（贝叶斯分类器）：

$$f^* = \arg\min_f P(f(X) \neq Y)$$

Bayes Classifier

- 等价于:

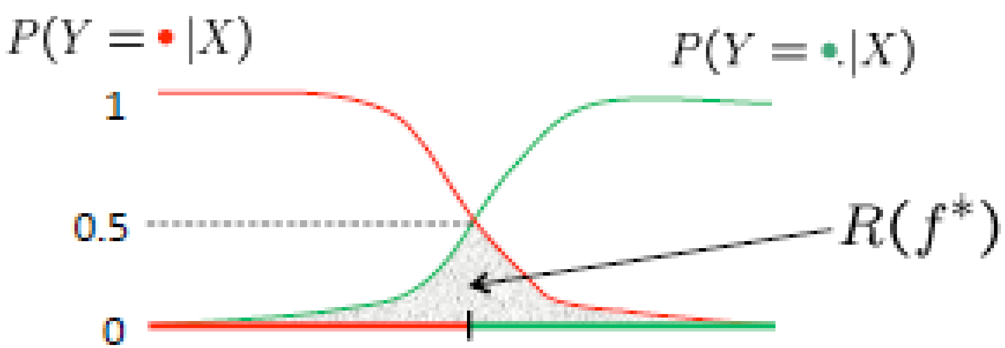
$$f^*(x) = \arg\max_{Y=y} P(Y = y | X = x)$$



贝叶斯分类器

- 最优决策器（贝叶斯分类器）：

$$f^* = \arg\min_f P(f(X) \neq Y)$$



贝叶斯误差

- 即使最优分类器也会使概率误差>0
- 最优的分类器依赖于未知却固定的分布P(X,Y)

贝叶斯分类器

- 生成模型: 能够基于模型重新产生数据

$$P(x,y) \Longrightarrow P(x), P(y), P(x|y), P(y|x)$$

Generative Model

- 决策模型: 直接学习决策分布

$$P(y|x)$$

Discrininative Model

- 确定性模型: 直接学习决策函数

$$f: R \rightarrow \{0,1\}$$

贝叶斯分类器

- 贝叶斯规则： $P(Y|X)=\dfrac{P(X|Y)P(Y)}{P(X)}$

$$P(Y=y|X=x) \;=\; \frac{P(X=x|Y=y)P(Y=y)}{P(X=x)}$$

- 最优分类器：

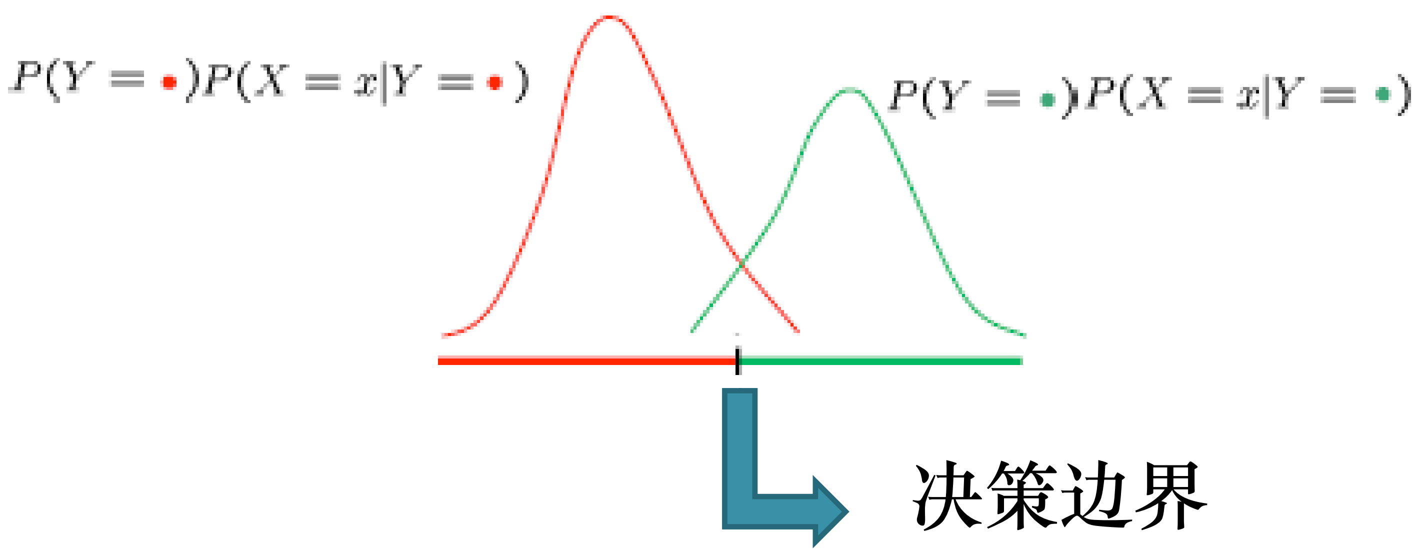
$$\begin{aligned} f^*(x) &= \arg\max_{Y=y} P(Y=y|X=x) \\ &= \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{类条件密度}}\underbrace{P(Y=y)}_{\text{类先验}} \end{aligned}$$

贝叶斯分类器

- 例子：高斯类条件密度（1维情形）

$$P(X=x|Y=y)=\frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)$$

- 二分类问题：

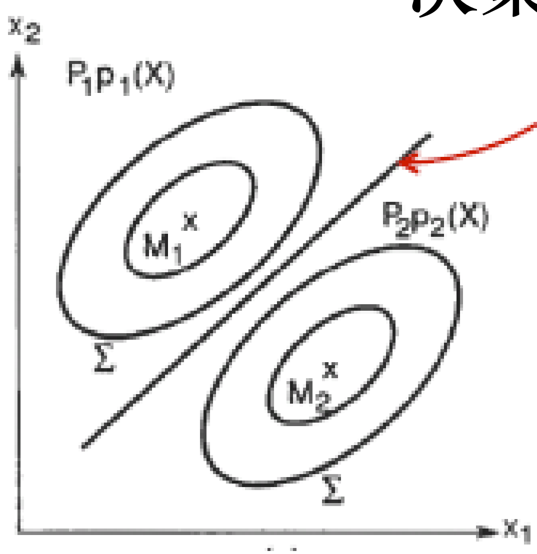


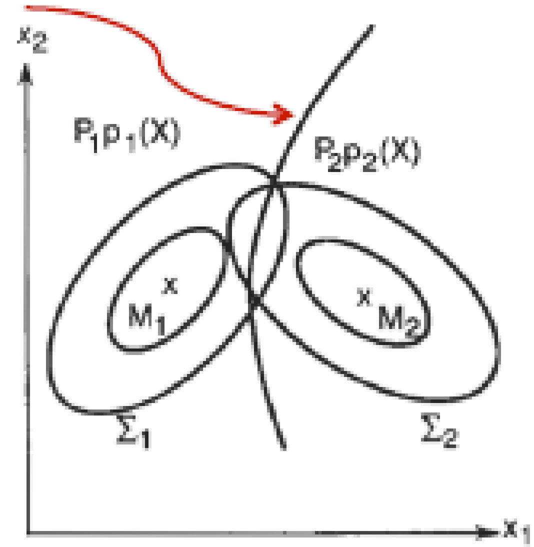
贝叶斯分类器

- 例子：高斯类条件密度（2维情形）

$$P(X=x|Y=y)=\frac{1}{\sqrt{2\pi|\Sigma_y|}}\exp\left(-\frac{(x-\mu_y)\Sigma_y^{-1}(x-\mu_y)'}{2}\right)$$

决策边界





贝叶斯分类器

- 最优分类器：

$$
\begin{aligned}
f^*(x) &= \arg\max_{Y=y} P(Y=y|X=x) \\
&= \arg\max_{Y=y} \underbrace{P(X=x|Y=y)}_{\text{类条件密度}}\underbrace{P(Y=y)}_{\text{类先验}}
\end{aligned}
$$

- 需要知道的信息：
  - 类先验： $P(Y=y)$
  - 似然： $P(X=x|Y=y)$

主要内容

- 贝叶斯分类器
- NB基本原理
- MLE vs. MAP
- 垃圾邮件分类
- Bag of Words
- 字符识别

NB基本原理

- 任务：测试是否一个男生是受欢迎的
- 训练数据：

$X = (X_1 \quad X_2 \quad X_3 \quad \ldots \quad \ldots \quad X_d) \quad Y$

幽默	高大	家境好	成绩好	有思想	做事靠谱	受欢迎
是	是	否	是	是	是	是
是	否	是	否	是	是	是
否	否	否	是	否	否	否
否	是	否	是	否	否	否

- 目标：学习P(Y|X) 多少参数需要学习?
- 类先验：$P(Y = y)$ 若有K个类，K-1
- 似然：$P(X=x \mid Y = y)$ 若每个特征为2类特征，(2^d-1)K

NB基本原理

- 任务：测试是否一个男生是受欢迎的
- 训练数据：

$X = (X_1 \quad X_2 \quad X_3 \quad \ldots \quad \ldots \quad X_d) \quad Y$

幽默	高大	家境好	成绩好	有思想	做事靠谱	受欢迎
是	是	否	是	是	是	是
是	否	是	否	是	是	是
否	否	否	是	否	否	否
否	是	否	是	否	否	否

- 目标：学习P(Y|X) 多少参数需要学习?
- 共2^dK-1个参数需要估计!
- 需要远大于共2^dK-1个训练数据来训练所有变量!

NB基本原理

- 条件独立：

$$(\forall x, y, z) P(X = x | Y = y, Z = z) = P(X = x | Z = z)$$

- 等价于：

$$P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z)$$

- 例：P(会用决策树|会用朴素贝叶斯，上过ML课) = P(会用决策树|上过ML课)

NB基本原理

- 分析： A=会用决策树, B=会用朴素贝叶斯,C=上过ML课

  P(会用决策树|会用朴素贝叶斯，上过ML课) = P(会用决策树|上过ML课)

- ✓ 当未知C时，A,B相关，因为会用一种机器学习方法的学者通常会用另一种的概率更大
- ✓ 当知道C时，A,B完全由学生自己的能力、悟性与努力程度有关，因此变成独立事件

NB基本原理

P(会用决策树，会用朴素贝叶斯|上过ML课)
P(A,B|C)

- 问题：根据会不会使用决策树与朴素贝叶斯，预测是否上过ML课
- 从两个条件独立的特征入手：A , B

估计类概率密度（似然）需要估计的参数个数:
P(A,B|C) ： (2^2-1)*2 = 6
利用条件概率假设:
P(A,B|C) = P(A|C) P(B|C) : (2-1)*2+(2-1)*2=4

NB基本原理

- 朴素贝叶斯假设：
  - 所有特征在给定类下条件独立

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

  - 更一般的：

$$P(X_1 \ldots X_d|Y) = \prod_{i=1}^{d} P(X_i|Y)$$

  - 需要估计多少参数?

(2-1）dK vs. (2^d-1)K!!

主要内容

- 贝叶斯分类器
- NB基本原理
- MLE vs. MAP
- 垃圾邮件分类
- Bag of Words
- 字符识别

MLE vs. MAP

- 贝叶斯分类器判别过程:
- 给定:
  - 类先验P(Y)
  - 对每个特征Xi,似然P(Xi|Y)
- 决策规则:

$$
\begin{aligned}
f_{NB}(\mathbf{x}) &= \arg\max_y P(x_1, \ldots, x_d \mid y) P(y) \\
&= \arg\max_y \prod_{i=1}^{d} P(x_i|y) P(y)
\end{aligned}
$$

- 若条件独立假设成立，NB解为贝叶斯分类器!

MLE vs. MAP

- 贝叶斯分类器训练过程：
- 训练数据：

$$\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n} \qquad X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$$

- MLE估计：
  - 类先验： $\widehat{P}(y) = \dfrac{\{\#j : Y^{(j)} = y\}}{n}$
  - 似然： $P(x_i|y) = \dfrac{\{\#j : X_i^{(j)} = x_i, Y^{(j)} = y\}/n}{\{\#j : Y^{(j)} = y\}/n}$

MLE vs. MAP

- 如果你在y=b类中未观察到样本俱备特征X1=a，会发生什么情况?

- 无论X2,…,Xd取什么值，定有：

$$P(Y=b \mid X_1=a, X_2, \ldots, X_d) = 0$$

$$\boxed{P(X_1=a, X_2 \ldots X_n|Y) = P(X_1=a|Y)\prod_{i=2}^{d} P(X_i|Y)}$$

- 该怎么办?

MLE vs. MAP

- 最大似然估计：

$$\theta_{MLE} = \frac{n_Z}{n_Z + n_B}$$

- 当投掷硬币次数太少时，
可能估计为 P(正面)=0

- 最大后验估计： $\theta_{MAP} = \frac{\beta_Z + n_Z - 1}{\beta_Z + n_Z + \beta_B + n_B - 2}$

- 相当于模拟增加了硬币的投掷次数
- 当投掷硬币次数很少时，避免出现
P(正面)=1/0的异常情况

MLE vs. MAP

- 训练数据： $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$
- 最大后验估计：增加“虚拟”样本

$$P(X_i = a|Y = b) = \frac{\{\#j: X_i^{(j)} = a, Y^{(j)} = b\} + \beta_i(a, b) - 1}{\{\#j: Y^{(j)} = b\} + \sum_a (\beta_i(a, b) - 1)}$$

- 此时，即使在某类中未观察到某一特征，后验概率也一定非0.

主要内容

- 贝叶斯分类器
- NB基本原理
- MLE vs. MAP
- 垃圾邮件分类
- Bag of Words
- 字符识别

垃圾邮件分类

- 文本分类：
  - Y = {垃圾邮件，正常邮件}
  - Y = {文章的主题}
  - Y = {主页的类型}
  - 。。。

- 特征X该如何表达?

垃圾邮件分类

- X是整个文件！
- 怎样表示?

西安交大“学者教授进中学”系列科普讲座火热进行中

来源：交大新闻网 日期 2015-12-31 18:11 点击：1703

发送图片到手机

月29日，郑南宁院士在江苏淮阴中学为同学们作了一场主题为“太空探索与机器人”的科普讲座。至此，大2015-2016学年“学者教授进中学”系列科普讲座活动已在全国各省市区70所中学完成77场讲座，参加

- 【讲座预告】沉醉不知归路-油画专业...
- 【讲座预告】创源论坛机械学院专场报告
- 【讲座预告】“学而”讲坛——教授...
- 国梦 少年强》央视报道交大少年...
- 【讲座预告】万达副总宁奇峰——转...
- 【讲座预告】“学而”讲坛——全球...
- 【讲座预告】“创源”论坛机械学院...
- 【讲座预告】“馥莉论坛”第五期：...
- 【讲座预告】“学而”讲坛——教授...
- 过程装备与控制工程高端学术论坛会议

栏目新闻

- 西安交通大学新年贺辞
- 西安交通大学2015年度十大新闻
- 西安交大举办2014—2015学年学生表...
- 西安交大4人入选国家 “万人计划”...
- 【双甲子校庆】中国农业银行捐赠西...
- 【双甲子校庆】中国银行全力支持西...
- 陕西省法学会企业经济法治研究会在...
- 中组部、省委组织部来校调研基层党...

主要内容

- 贝叶斯分类器
- NB基本原理
- MLE vs. MAP
- 垃圾邮件分类
- Bag of Words
- 字符识别

Bag of Words

- 表达方式
  - 有一个字典集，构成关键字集合，如包含1000个元素
  - Xi代表在字典中的第i个字在文本中出现次数，因此，特征共1000维
- 朴素贝叶斯执行模式
  - 估计先验P(y) ,对所有类别
  - 估计每一个P(Xi|y),对所有类别y与所有特征Xi

Bag of Words

- 基本思想：不管字的排序，只管出现次数
- 看上去似乎非常简易，实际中却往往非常有效！

Bag of Words

$$\prod_{i=1}^{LengthDoc} P(x_i|y) = \prod_{w=1}^{W} P(w|y)^{count_w}$$

Bag of Words

Bag of Words

- 基本思想：不管字的排序，只管出现次数
- 看上去似乎非常简易，实际中却往往非常有效！

$$\prod_{i=1}^{LengthDoc} P(x_i|y) = \prod_{w=1}^{W} P(w|y)^{count_w}$$

Bag of Words

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

Bag of Words

- BoW学习过程：通过大量训练文本，获得
  - 类先验P(Y)
  - 对每个特征Xi,似然P(Xi|Y)

- BoW决策过程：
  - 对一个测试文本，使用NB决策规则：

$$
\begin{aligned}
h_{NB}(\mathbf{x}) &= \arg\max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y) \\
&= \arg\max_y P(y) \prod_{w=1}^{W} P(w|y)^{count_w}
\end{aligned}
$$

Bag of Words

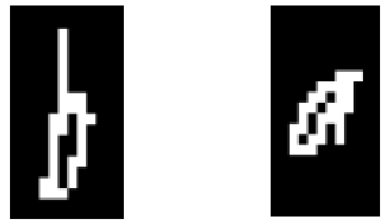
- 准确率可达到95%以上！
- 是最常用的垃圾邮件分类器之一

主要内容

- 贝叶斯分类器
- NB基本原理
- MLE vs. MAP
- 垃圾邮件分类
- Bag of Words
- 字符识别

字符识别

- Xi是在第i个像素处的灰度值



- 高斯朴素贝叶斯:

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \, e^{\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

- 每个类别，每个像素不同的均值与方差变量
  - 有时假设方差:
    - 与Y无关
    - 与X无关
    - 或与两者均无关

字符识别

- 最大似然估计：

$$\boxed{\widehat{\mu}_{MLE} = \frac{1}{N}\sum_{j=1}^{N} x_j}$$

$$\widehat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

$$\boxed{\widehat{\sigma}^2_{unbiased} = \frac{1}{N-1}\sum_{j=1}^{N} (x_j - \widehat{\mu})^2}$$

$$\widehat{\sigma}^2_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k) - 1} \sum_j (X_i^j - \widehat{\mu}_{ik})^2 \delta(Y^j = y_k)$$

要求

1. 贝叶斯分类器的基本概念与原理

2. NB方法的：假设，动机，怎样训练和预测，MAP重要性

3. 文本分类与BoW原理

4. 高斯NB原理

5. 若有少数Xi相关，能否改造NB方法?

阅读：

[1] The Elements of Statistical Learning: Data Mining, Inference and Prediction. Hastie, Tibshirani, Friedman. Springer, 2008. 6.6.3 The Naïve Bayes Classifier

[2] On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes, Andrew Y. Ng and Michael Jordan. In NIPS 14, 2002.