華東師範大學
EAST CHINA NORMAL UNIVERSITY

朴素贝叶斯

武蔡丽
51184506045

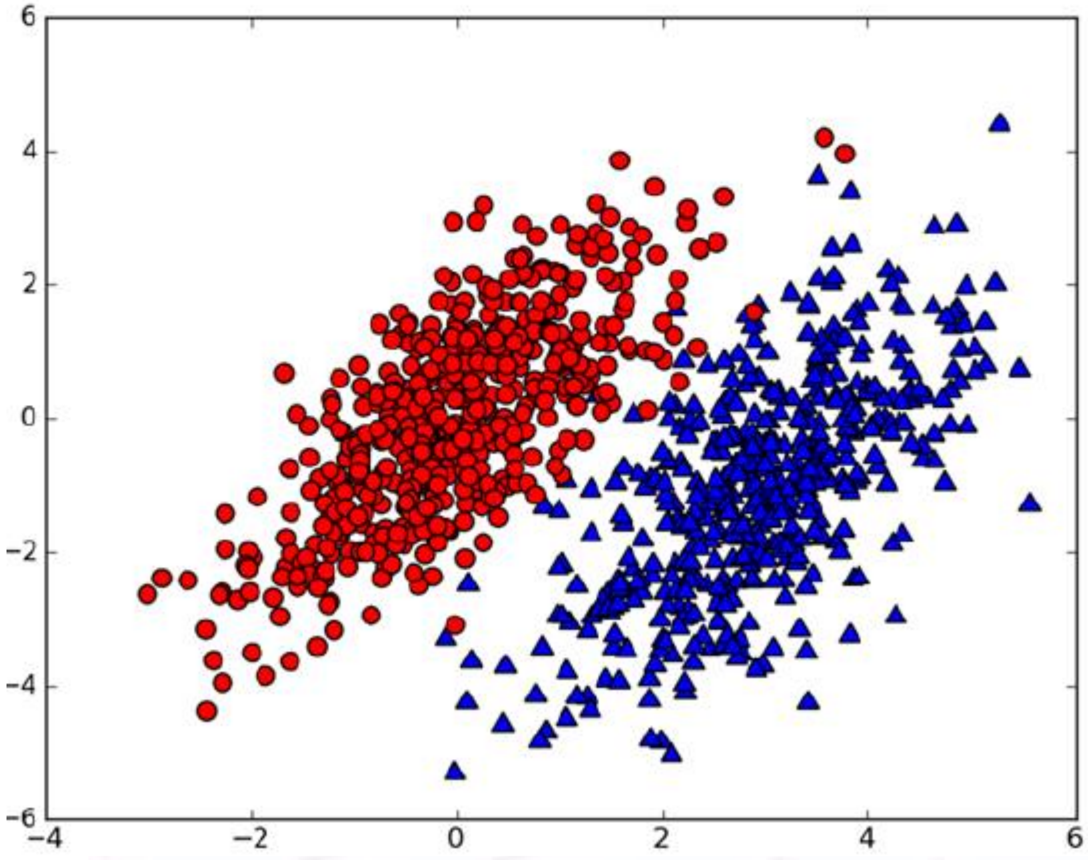
Content

- 原理
- 实例
- 作业

Content

- 原理
- 实例
- 作业

贝叶斯决策理论



如果p1(x, y) > p2(x, y)，类别为1
如果p2(x, y) > p1(x, y)，类别为2

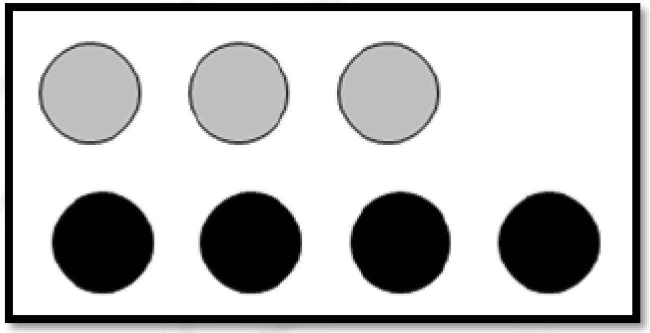
简而言之，选择高概率对应的类别；

即，做错误率最小的决策

贝叶斯概率（Bayesian Probability）

贝叶斯概率以18世纪的一位神学家托马斯·贝叶斯命名。贝叶斯概率引入先验知识和逻辑推理来处理不确定命题。另一种概率理论则被称为频数概率（frequency probability），只从数据本身获得结论，并不考虑逻辑推理及先验知识。

条件概率（Conditional probability）

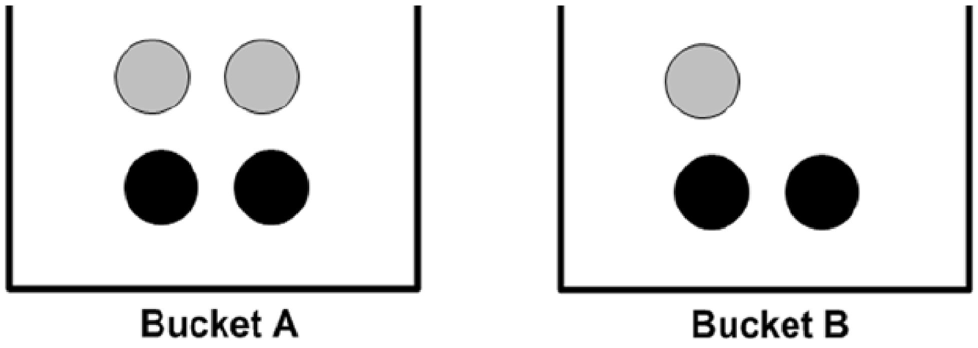


假设有个盒子，里面有三个灰球，四个黑球，那么我们随手抓一个，抓到灰球的概率是多少？抓到黑球的概率又是多少？

P(灰球) = 3/7　　P(黑球) = 4/7

条件概率（Conditional probability）

现在，7个球如下图所示，被放入两个盒子；上述的概率又该如何计算呢?



把问题更具体一点，随手从B盒中抓一个，抓到灰球的概率是多少?

在“已知是从B盒里抽取的条件下，取出灰球的概率”，这便被称为“条件概率”。
在 事件B 发生条件下 事件A 发生的概率如下:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

即，

P(灰球| B盒) = P(抽中B盒中的灰球) / P(B盒)

贝叶斯定理（Bayes’rule）

在事件B发生条件下事件A发生的条件概率：

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

那么，在事件A发生条件下事件B发生的条件概率为：

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

所以，我们可以这样计算：

$$P(A|B)\,P(B) = P(A \cap B) = P(B|A)\,P(A).$$

我们对这个引理进行变换，两边同除 P(A)，若 P(A) 不为零，便得到了著名的贝叶斯定理：

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}.$$

贝叶斯定理（Bayes’rule）

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}.$$

其中，P(B) 被称为 “先验概率”，在A事件发生之前，对B事件概率的一个判断。
P(B | A) 则被称为 “后验概率”，是在A事件发生之后，对B事件的重新评估。
P(A | B) / P(A) 则被称为 “可能性函数”，是一个调整因子，使得预估概率更接近真实概率。

朴素贝叶斯（Naïve Bayes)

- 朴素贝叶斯是基于贝叶斯定理与特征条件独立假设的分类方法。
- 对于给定的训练集，首先基于特征条件独立立假设学习输入 / 输出的联合概率分布
- 然后基于此模型，对给定的输入x，利用贝叶斯定理求出后验概率最大的输出y。
- 因为我们假设每个特征条件之间是独立的，比如一个单词出现的概率和其他相邻单词没有关系，所以这个算法就被称为“朴素”的贝叶斯。
- 然而虽然Naive，但朴素贝叶斯实现简单，学习和预测的效率都很高，所以应用很广泛。

華東師範大學
EAST CHINA NORMAL UNIVERSITY

Content

- 原理
- 实例
- 作业

朴素贝叶斯应用实例一

- 朴素贝叶斯的一个常见用途是区分垃圾邮件。
- 用S表示垃圾邮件，H表示正常邮件。
- 随机抽取一封邮件，抽到垃圾邮件的概率是 P(S) ,正常邮件的概率是 P(H)
- 给出一封邮件 D，D= { W1, W2, ..., Wn }

D是垃圾邮件的概率：

$$P(S|D) = \frac{P(S)P(D|S)}{P(D)}$$

D是正常邮件的概率：

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)}$$

如果 P(S | D) > P(H | D) ,则邮件D是垃圾邮件，

反之，若 P(S | D) < P(H | D) ,则邮件D是正常邮件。

先验概率 P(S) & 条件概率 P(D | S) 如何计算?

先验概率 P(S)：极大似然估计

- 垃圾邮件的概率是 P(S) ,正常邮件的概率是 P(H) , P(S) + P(H) = 1
- 从网上所有的邮件中抽取 m+n 封邮件，其中 m 封垃圾邮件，n 封正常邮件的概率为：

$$P = P(S)^m P(H)^n = P(S)^m (1 - P(S))^n$$

- 假设有100封邮件的训练样本，其中30封垃圾邮件，70封正常邮件，它是由上述的概率模型产生生的，那么我们就可以依靠这个样本来估计参数 P(S)，这个估计基于这样的思想：我们所估计的模型参数，要使得产生这个样本集的可能性最大。
- 所以我们所求出的 P(S)，要使 P 最大：

$$P = P(S)^{30} (1 - P(S))^{70}$$

- 对上式求导，令其等于 0 ，得：

$$30P(S)^{29}(1 - P(S))^{70} - 70(1 - P(S))^{69} P(S)^{30} = 0$$

$$30P(S)^{29}(1 - P(S))^{70} = 70(1 - P(S))^{69} P(S)^{30} \rightarrow \frac{1 - P(S)}{P(S)} = \frac{7}{3}$$

- 所以 P(S) = 0.3

条件概率 P(D | S)：特征条件独立假设

- 给出一封邮件 D，D= { W1, W2, ..., Wn }
- P(D | S) = P(W1, W2, ..., Wn | S)
- 根据一般乘法公式，可改写为：

  P(D | S) = P(W1| S) P(W2 | S,W1) P(W3 | S,W1,W2) ... P(Wn | S,W1,...,Wn-1)
- 为什么是“朴素”贝叶斯

  P(D | S) = P(W1| S) P(W2 | S) P(W3 | S) ... P(Wn | S)

$$P(W_i|S) = \frac{P(W_i,S)}{P(S)}$$

$$P(S) = \frac{\text{垃圾邮件的个数}}{\text{邮件总数}} \qquad P(W_i,S) = \frac{\text{包含词}W_i\text{的垃圾邮件个数}}{\text{邮件总数}}$$

$$P(W_i|S) = \frac{\text{包含词}W_i\text{的垃圾邮件个数}}{\text{垃圾邮件的个数}}$$

拉普拉斯平滑

- 给出一封邮件 D，$D=\{W_1,W_2,\dots W_k\}$，$(k\neq 1,2,..n)$

$$P(W_k|S)=\frac{包含w_k的垃圾邮件数量}{垃圾邮件数量}=0\quad P(W_k|H)=\frac{包含w_k的正常邮件数量}{正常邮件数量}=0$$

- 即，P(S | D) = 0 & P(H | D) = 0

正确吗！！！

$$P(\mathrm{W_k}|\mathrm{S})=\frac{包含W_k的垃圾邮件数量+1}{垃圾邮件数量+1}\neq 0$$

$$P(\mathrm{W_k}|\mathrm{H})=\frac{包含W_k的正常邮件数量+1}{正常邮件数量+1}\neq 0$$

朴素贝叶斯应用实例二

- 朴素贝叶斯还可用于多分类任务。
- 给定成绩等级划分规则如下：

  语文：大于等于120分为A；大于等于105分，小于120分为B；大于等于90分，小于105分为C；小于90分为D。

  数学、英语： 大于等于135分为A；大于等于120分，小于135分为B；大于等于105分，小于90分为C；小于90分为D。

  总分：大于等于650分为A；大于等于550分，小于650分为B；大于等于550分，小于450分为C；小于450分为D。

- 目标是通过语数英的等级预测总成绩的等级

朴素贝叶斯应用实例二

要计算的值为：

P(总成绩等级为A)*P(语文成绩为A|总成绩为A)*(数学成绩为A|总成绩为A)*(英语成绩为B|总成绩为A)

P(总成绩等级为B)*P(语文成绩为A|总成绩为B)*(数学成绩为A|总成绩为B)*(英语成绩为B|总成绩为B)

P(总成绩等级为C)*P(语文成绩为A|总成绩为C)*(数学成绩为A|总成绩为C)*(英语成绩为B|总成绩为C)

P(总成绩等级为D)*P(语文成绩为A|总成绩为D)*(数学成绩为A|总成绩为D)*(英语成绩为B|总成绩为D)

取值最高的一个对应的等级即为预测结果。

朴素贝叶斯的优点

- 朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。
- 对小规模的数据表现很好，能个处理多分类任务，适合增量式训练，尤其是数据量超出内存时，我们可以一批批的去增量训练。
- 对缺失数据不太敏感，算法也比较简单，常用于文本分类

朴素贝叶斯的缺点

- 理论上，朴素贝叶斯模型与其他分类方法相比具有最小的误差率。但是实际上并非总是如此，这是因为朴素贝叶斯模型假设属性之间相互独立，这个假设在实际应用中往往是不成立的。
- 举个例子，在文本分类中，朴素贝叶斯会假设单词出现概率是不相关的，但angry后出现person的概率明显高于sofa。
- 在属性个数比较多或者属性之间相关性较大时，分类效果不好。而在属性相关性较小时，朴素贝叶斯性能最为良好。对于这一点，有半朴素贝叶斯之类的算法通过考虑部分关联性适度改进。

$$h(x) = \max(P(c)\prod_{i=1}^{d} P(x_i \mid c, pa_i))$$

朴素贝叶斯的缺点

- 需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳。
- 举个例子：假设一个暗箱中有白球、黑球共两个，虽然不知道具体的颜色分布情况、但是知道这两个球是完全一样的。现在有放回地从箱子里抽了 2 个球，发现两次抽出来的结果是 1 黑 1 白，那么该如何估计箱子里面球的颜色？从直观上来说似乎箱子中也是 1 黑 1 白会比较合理（频率估计概率），朴素贝叶斯则用极大似然估计来得到“1 黑 1 白”这个先验概率。

$$p(\tilde{x}|\theta) = \theta^{X_1+X_2}(1-\theta)^{2-X_1-X_2}$$

華東師範大學
EAST CHINA NORMAL UNIVERSITY

朴素贝叶斯的缺点

- 由于我们是通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率。
- 对输入数据的表达形式很敏感。在分类变量的情况下表现良好，若是数值变量，则需要假设其为正态分布

Content

- 原理
- 实例
- 作业

作业

基于给定的数据集，抽取特征，训练一个可以分类垃圾邮件的贝叶斯分类器。

数据集分两部分：训练集（spam_train.txt，共5000封邮件），测试集（spam_test.txt，共1000封邮件)，其中正常邮件和垃圾邮件比例为3：1。

数据格式说明：

训练集：每封邮件占一行，格式为 标签+空格+内容，标签1代表是垃圾邮件，0代表是正常邮件。

测试集：每封邮件占一行，只有内容，无标签

提交要求

1.代码部分

文本处理得到向量的的代码，包括去停用词，文本向量化（词包和词集模式均可，要指明）等；

贝叶斯模型的代码，包括训练和测试部分。

2.模型评价部分

提交1000条测试集文本的预测结果（即1000个0/1标签，放于txt文件中，以空格分割）；

对算法及模型进行简要说明，给出在训练集上的错误率结果。

评分标准

1. 基础分50分，每晚交12小时减10分；
2. Naïve Bayes算法代码17分；
3. 预处理到特征向量的生成17分；
4. 模型评价16分

	Healthy	Spam
Healthy	681	12
Spam	28	279