Chapter 8 – Naïve Bayes

Instructor: Zach Zhizhong ZHOU,
Shanghai Jiao Tong University
主讲教师：周志中，上海交通大学

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

Characteristics 特征

Data-driven, not model-driven 数据驱动而非模型驱动

Make no assumptions about the data 对数据未做任何假设。

Naïve Bayes: The Basic Idea
朴素贝叶斯：基本思路

For a given new record to be classified, find other records like it (i.e., same values for the predictors) 对于一个需要分类的新纪录，找到类似该记录的其他记录（比如，预测因子取值相同）。

What is the prevalent class among those records? 与新纪录类似的这些记录中较多的类别是什么？

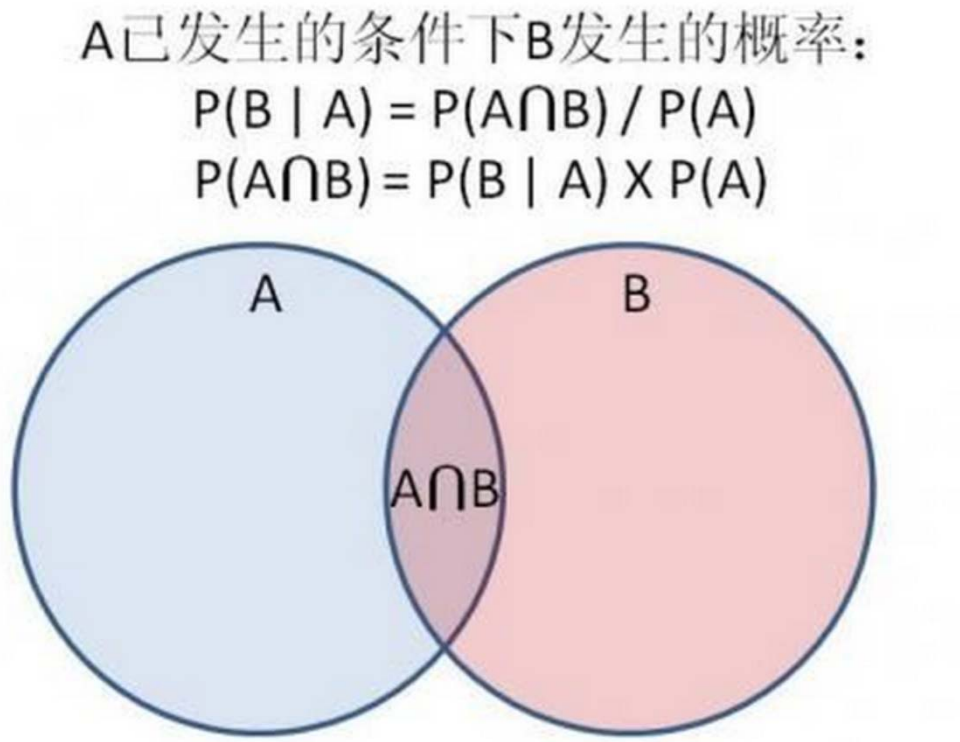
Assign that class to your new record 把新纪录归入这个类别。

Usage 应用

- Requires categorical variables 要求使用类别型变量
- Numerical variable must be binned and converted to categorical 数值型变量必须进行归类并转换成类别型变量。
- Can be used with very large data sets 可以用在非常大的数据集上。
- Example: Spell check programs assign your misspelled word to an established “class” (i.e., correctly spelled word) 例子：拼写检查程序把你拼错的词归入一个事先建立好的类别（例如：正确拼写的词）。

条件概率

- $P(A \cap B) = P(A,B) = P(B|A)*P(A)$

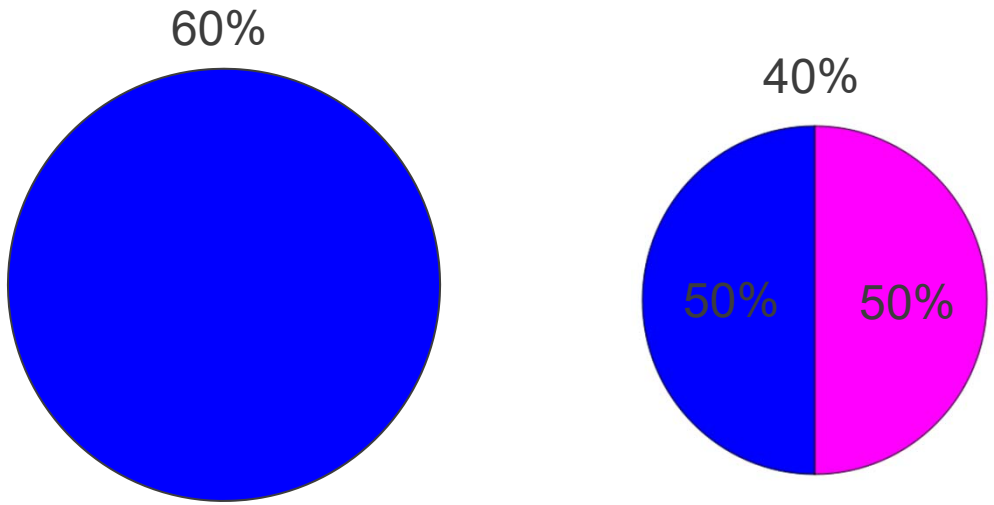


贝叶斯法则

- $P(A \cap B) = P(A,B) = P(B|A)*P(A) = P(A|B) * P(B)$
- $P(A|B) = P(A)* P(B|A) / P(B)$
- P(A): A的先验概率（或者边缘概率）
- P(A|B): 发生B之后发生A的条件概率，或者称之为A的后验概率。
- P(B): B的先验概率
- P(B|A): 发生A之后发生B的条件概率，或者称之为B的后验概率。

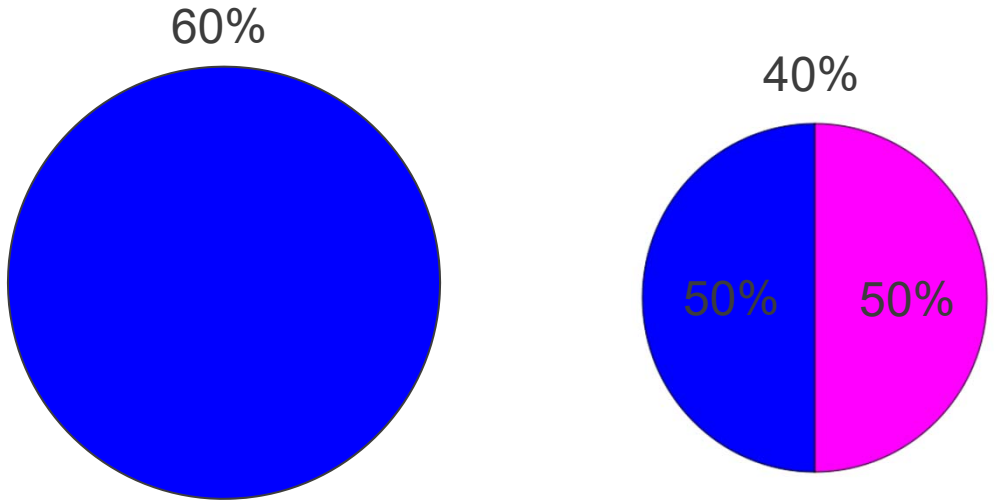
贝叶斯法则

- 一所学校里面有 60% 的男生，40% 的女生。男生总是穿长裤，女生则一半穿长裤一半穿裙子。如果你看到一个学生穿了长裤，请问该学生是男生的概率是多大?



贝叶斯法则

- 一所学校里面有 60% 的男生，40% 的女生。男生总是穿长裤，女生则一半穿长裤一半穿裙子。如果你看到一个学生穿了长裤，请问该学生是男生的概率是多大?



- 考虑10个学生，6人是男生，4人是女生，其中2女生穿长裤，2女生穿裙子。你看到穿长裤的人是8个学生之一（6男生、2女生）。在8个学生当中，男生占3/4。因此你看到了穿长裤的学生，该学生为男生概率是75%。

贝叶斯法则

- M: 男生，F: 女生，L: 长裤，S: 裙子
- P(M) = 60%, P(F) = 40%, P(L|M) = 1, P(S|M)=0, P(L|F)=50%, P(S|F) = 50%， 求P(M|L)。
- P(M|L) = P(M) * P(L|M) / P(L) = 0.6 * 1 / P(L)
- P(F|L) = P(F) * P(L|F)/P(L) = 0.4*0.5/P(L)
- 已知 P(M|L)+P(F|L)=1, 所以： P(M|L) = 0.6/(0.6+0.4*0.5) = 75%
- 也可以用全概率公式算出 P(L) = P (L|M) * P(M) + P (L|F)*P(F)
- P(L) = 1*0.6 + 0.5*0.4 = 0.8
- 所以，P(M|L) = 0.6 / P(L) = 0.6/0.8 = 75%。

三门问题（Monty Hall Problem）

- 你参加电视台的一个抽奖节目。台上有三个门，一个后边有汽车，其余后边是山羊。主持人让你任意选择其一。然后他打开其余两个门中的一个，你看到是山羊。这时，他给你机会让你可以重选，也就是你可以换选另一个剩下的门。那么，你换不换?

三门问题（Monty Hall Problem）

- 设P(X)为A、B、C三道门后面有车的概率，则P(A)=P(B)=P(C)=1/3

- 假定：游戏者任选了一道门A，而主持人（HOST）知道那个门后面有车，他肯定会打开一道后面是羊的门（假设是C门）。

- 如果车在A门后面，主持人有B、C两种选择，打开C门（“羊”门）的概率为　P(Host opens C|A) = 1/2

- 如果车在B门后面，主持人没有选择，只能打开C门　P(Host opens C|B) = 1

- 如果车在C门后面，主持人一样没得选择，绝对不能开C门　P(Host opens C|C) = 0

- 所以，主持人打开C门的概率为　P(Host opens C) = P(A)*P(H.o. C|A) + P(B)*P(H.o. C|B) + P(C)*P(H.o. C|C) = 1/6 + 1/3+ 0 = 1/2

三门问题（Monty Hall Problem）

- 根据贝叶斯公式，在主持人打开C门的条件下，A、B两门后面是车的概率分别为

- P(A|Host opens C) = P(A)*P(Host opens C|A) / P(Host opens C) = (1/6) / (1/2) = 1/3

- P(B|Host opens C) = P(B)*P(Host opens C|B) / P(Host opens C) = (1/3) / (1/2) = 2/3

检测SNS社区中不真实账号

- 检测指标：a1：日志数量/注册天数，a2：好友数量/注册天数，a3：是否使用真实头像。

- a1：{a1<=0.05 ♣, 0.05<a1<0.2 ♥, a1>=0.2 ♠}，a2：{a2<=0.1 ▼, 0.1<a2<0.8 ►, a2>=0.8 ▲}，a3：{a3=0（不是）■，a3=1（是）□}。

- C0表示真实账号（☺），C1表示不真实账号（ ☻ ）。

- P(C0) = P(☺) = 0.89, P(C1) = P(☻) = 0.11

检测SNS社区中不真实账号

	♣	♥	♠
☺	0.3	0.5	0.2
☻	0.8	0.1	0.1
	▼	▸	▲
☺	0.1	0.7	0.2
☻	0.7	0.2	0.1
	■	□	
☺	0.2	0.8	
☻	0.9	0.1	

解释：P(♣ | ☺) = P(a1<=0.05 | C0) = 0.3

问题：P(☺ | ■) = ？如果观察到一个用户没用真实头像，该用户是真实账号的概率有多大?

P(☺ | ■) = P (☺) * P(■ | ☺) / P(■) = 0.89*0.2 / P(■)

P(■) = P(■ | ☺)*P(☺) + P(■ | ☻)*P(☻) = 0.89*0.2 + 0.11*0.9

所以，P(☺ | ■) = 64.26%

检测SNS社区中不真实账号

- 问题：P(☺|♥, ►,■) = ？如果观察到一个用户满足0.05<a1<0.2, 0.1<a2<0.8，没用真实头像，你倾向于将该用户归入真实账号还是虚假账号?
- P(☺|♥, ►,■) = P (☺) * P(♥, ►,■| ☺) / P(♥, ►,■) =

  (注意：第二个等号用了独立性假设)

  P (☺) * P(♥| ☺)*P(►| ☺)* P(■| ☺) / P(■, ♥, ►) =

  0.89*0.5*0.7*0.2/P(■, ♥, ►) =0.0623/P(■, ♥, ►)
- P(☻|♥, ►,■) = P (☻) * P(♥, ►,■| ☻) / P(♥, ►,■) =

  P (☻) * P(♥| ☻)*P(►| ☻)* P(■ | ☻) / P(■, ♥, ►) =

  0.11*0.1*0.2*0.9/P(■, ♥, ►) = 0.00198/P(■, ♥, ►)
- 倾向于把该记录归入真实账号。
- P(☺|♥, ►,■) = 0.0623 / (0.0623+0.00198) = 97%
- P(☻|♥, ►,■) = 0.00198 / (0.0623+0.00198) = 3%

Exact Bayes Classifier 精确贝叶斯分类器

Relies on finding other records that share same predictor values as record-to-be-classified. 对欲被分类的记录，在其他记录中找到与该记录的预测因子相同的记录。

Want to find "probability of belonging to class *C*, given specified values of predictors." 给定预测因子取值情况下，希望算出"属于C类的概率"

Even with large data sets, may be hard to find other records that exactly match your record, in terms of predictor values. 但是就算有很大的数据集，也很难在原数据集中找到与新纪录的预测因子恰好吻合的记录。

Solution – Naïve Bayes
解决方案：朴素贝叶斯

- Assume independence of predictor variables (within each class) 假设预测变量在每个分类当中相互独立。

- Use multiplication rule 使用乘法准则。

- Find same probability that record belongs to class C, given predictor values, without limiting calculation to records that share all those same values 给定预测变量并算出新记录属于C类的概率，如果从C类已经存在的记录中可以找到“属于C类的概率”与新纪录“属于C类的概率”相同则将新纪录归入C类。我们无需在C类中找到与新纪录预测因子完全相同的记录。

Calculations 计算过程

1. Take a record, and note its predictor values 拿到一个新纪录并注意到它的预测变量取值。

2. Find the probabilities those predictor values occur across all records in C1 找到这些预测变量取值在类C1中发生的概率。

3. Multiply them together, then by proportion of records belonging to C1 将这些概率值连乘，然后乘以数据集中记录属于C1的概率。

Calculations 计算过程

4. Same for C2, C3, etc. 对类别C2，C3做同样的操作。

5. Prob. of belonging to C1 is value from step (3) divide by sum of all such values C1 … Cn 新纪录属于C1类的概率等于第（3）步的对C1类得到的连乘取值除以对所有C1…Cn类得到的连乘取值的求和。

6. Establish & adjust a “cutoff” prob. for class of interest 根据对类别感兴趣的程度建立并调整分配个各个类别的概率截值。

Example: Financial Fraud 例子：金融欺诈

Target variable: Audit finds fraud, no fraud 目标变量：审计找到的欺诈，没有欺诈

Predictors: 预测变量：

Prior pending legal charges (yes/no) 存在悬而未决的诉讼案件（是/否）

Size of firm (small/large) 公司大小（小/大）

Charges?	Size	Outcome
y	small	truthful
n	small	truthful
n	large	truthful
n	large	truthful
n	small	truthful
n	small	truthful
y	small	fraud
y	large	fraud
n	large	fraud
y	large	fraud

Exact Bayes Calculations
精确贝叶斯分类器的计算

Goal: classify (as “fraudulent” or as “truthful”) a small firm with charges filed 目标：将一个有案在身的小公司进行分类。

There are 2 firms like that, one fraudulent and the other truthful 有2个公司是有案在身的小公司，一个是骗子一个不是。

P(fraud | charges=y, size=small) = ½ = 0.50

Note: calculation is limited to the two firms matching those characteristics 注：计算基于特征与新记录相同的2个公司。

Naïve Bayes Calculations
朴素贝叶斯分类器计算

Same goal as before 目标与前同

Compute 2 quantities: 计算2个数值

Proportion of "charges = y" among frauds, times proportion of "small" among frauds, times proportion frauds = 3/4 * 1/4 * 4/10 = 0.075

Prop "charges = y" among truthfuls, times prop. "small" among truthfuls, times prop. truthfuls = 1/6 * 4/6 * 6/10 = 0.067

P(fraud | charges, small) = 0.075/(0.075+0.067)

= 0.53

Naïve Bayes, cont.

- Note that probability estimate does not differ greatly from exact 算出来的概率和精确贝叶斯分类器得到的概率并没有很大的不同。
- All records are used in calculations, not just those matching predictor values 但是所有记录都用于计算，而不是那些预测变量与新纪录的预测变量吻合的记录才用于计算。
- This makes calculations practical in most circumstances 这使得绝大多数情况下朴素贝叶斯分类器更实用。
- Relies on assumption of independence between predictor variables within each class 但朴素贝叶斯分类器依赖于独立性假设：每个类别中所有预测变量相互独立。

Independence Assumption 独立性假设

- Not strictly justified (variables often correlated with one another) 严格意义上不合理（变量经常与其他变量有相关性）

- Often “good enough” 不过这个假设经常是“足够好”的假设（也就是说我们用起来不会有太大误差）。

Advantages 优点

- Handles purely categorical data well 能够很好处理纯分类型数据。
- Works well with very large data sets 处理大数据集效果很好。
- Simple & computationally efficient 简单，计算效率好。

Shortcomings 缺点

- Requires large number of records 要求有大量的数据

- Problematic when a predictor category is not present in training data 如果一个预测因子的某个类型不在训练数据集该算法就有问题

  Assigns 0 probability of response, ignoring information in other variables

  导致对新数据处理时忽略其他变量中的有用信息。

- 例子：用owns yacht预测bought high-value life insurance。如果训练数据集中没有owns yacht=1的记录，那么对于将来所有owns yacht=1的记录，朴素贝叶斯分类器都会给bought high-value life insurance赋值为0。

On the other hand… 在另外一方面

- Probability rankings are more accurate than the actual probability estimates 朴素贝叶斯分类器对概率进行排序比对概率进行预测更为准确。
  - Good for applications using lift (e.g. response to mailing), less so for applications requiring probabilities (e.g. credit scoring) 用在使用提升水平的应用（如对推销邮件的回应）效果较好，但用在需要概率值的应用（如信用评分）则效果不太好。

Summary 总结

- No statistical models involved 不需要建立统计模型

- Naïve Bayes (like KNN) pays attention to complex interactions and local structure 朴素贝叶斯分类器（类似KNN）关注（预测变量间）复杂的交互关系以及局部数据结构。

- Computational challenges remain 仍然存在计算量上的挑战