贝叶斯分类器
数据挖掘导论，第二版，作者
谭、斯坦贝克、卡帕特内、库马尔
数据挖掘技术
分类:替代技术

贝叶斯分类器
解决分类问题的概率框架
条件概率:
贝叶斯定理:
( )
(|)(|)P X P X Y P Y P Y X
( )
()(|)
( )
()(|)
P Y
P X Y P X Y
P X
P X Y P Y X
·100%
·100%

利用贝叶斯定理进行分类
将每个属性和类别标签视为随机变量
给定一个具有属性的记录(X1，X2，…，Xd)
目标是预测 Y 类
具体来说，我们希望找到最大化 P 的 Y 值(Y| X1，X2，…，Xd)
我们能直接从数据中估算 P(Y| X1，X2，…，Xd)吗?
Tid 退款婚姻
状态
应纳税的
收入逃避
1 是单人 125K 否
2 不结婚 10 万不
3 无单个 70K 否
4 是已婚 12 万否
5 不离婚 95K 是的
6 不结婚 6 万不
7 是离婚 22 万否
8 没有单人 85K 是的
9 不结婚 75K 不
没有单个 90K 是 10
c c c

3

Example Data

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Can we estimate
  P(Evade = Yes | X) and P(Evade = No | X)?

  In the following we will replace
  Evade = Yes by Yes, and
  Evade = No by No

4

02/10/2020 数据挖掘导论，第 2 版 5
利用贝叶斯定理进行分类
方法:
用贝叶斯定理计算后验概率 P(Y | X1，X2，…，Xd)
最大后验概率:选择使 P 最大的 Y(Y | X1，X2，…，Xd)
相当于选择最大化 P(X1，X2，…，Xd|Y) P(Y)的 Y 值
如何估计 P(X1，X2，…，Xd | Y)?
( )
(|) ( ) (|)
1 2
1 2
1 2
d
d
n P X X X
p x x y p y p y x x x

5

Example Data

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Divorced}, \text{Income} = 120\text{K})$$

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Using Bayes Theorem:

- $P(\text{Yes} \mid X) = \frac{P(X \mid \text{Yes})P(\text{Yes})}{P(X)}$
- $P(\text{No} \mid X) = \frac{P(X \mid \text{No})P(\text{No})}{P(X)}$
- How to estimate P(X | Yes) and P(X | No)?

6

02/10/2020 数据挖掘导论，第 2 版 7
朴素贝叶斯分类器
给出等级时，假设 Xi 属性之间的独立性:
P(X1，X2，…，Xd | Yj)= P(X1 | Yj)P(X2 | Yj)…P(Xd | Yj)
现在，我们可以从训练数据中估计所有 Xi 组合和 Yj 组合的 P(Xi | Yj)
如果 P(Yj)　P(Xi | Yj)最大，则新点被分类为 Yj。

7

02/10/2020 数据挖掘导论，第 2 版 8
条件独立性
给定 Z，如果 P(X|YZ) = P(X|Z)，X 和 Y 是条件独立的
示例:手臂长度和阅读技能幼儿的手臂长度较短
与成年人相比，阅读能力有限如果年龄是固定的，手臂长度和阅读能力之间没有明显的关系手臂长度和阅读能力在给定的年龄条件下是独立的

02/10/2020 数据挖掘导论，第 2 版 9
实例数据的朴素贝叶斯
Tid 退款婚姻
状态
应纳税的
收入逃避
1 是单人 125K 否

2 不结婚 10 万不
3 无单个 70K 否
4 是已婚 12 万否
5 不离婚 95K 是的
6 不结婚 6 万不
7 是离婚 22 万否
8 没有单人 85K 是的
9 不结婚 75K 不
没有单个 90K 是 10
c c c
十(退款否，离婚，收入 12 万)
给定一个测试记录:
P(X |是)=
P(退款=否|是)x P(离婚|是)x P(收入= 12 万|是)
P(X |否)=
P(退款=否|否)x P(离婚|否)x P(收入= 12 万|否)

02/10/2020 数据挖掘导论，第二版 10
根据数据估计概率
P(y)= y 类实例的分数
例如，P(否)= 7/10，P(是)= 3/10
对于分类属性:P(Xi =c| y) = nc/ n
其中|Xi =c|是属性值为 Xi =c 且属于 y 类的实例数
示例:
P(状态=已婚|否)= 4/7 P(退款=是|是)=0
Tid 退款婚姻
状态
应纳税的
收入逃避
1 是单人 125K 否
2 不结婚 10 万不
3 无单个 70K 否
4 是已婚 12 万否
5 不离婚 95K 是的
6 不结婚 6 万不
7 是离婚 22 万否
8 没有单人 85K 是的
9 不结婚 75K 不
没有单个 90K 是 10
c c c

02/10/2020 数据挖掘导论，第 2 版 11
根据数据估计概率
对于连续属性:

离散化:将范围划分为箱:
用面元值替换连续值
-属性从连续更改为序数
概率密度估计:
假设属性遵循正态分布，☒使用数据来估计分布参数
(例如，均值和标准差) 一旦概率分布已知，就用它来估计条件概率 P(Xi |Y)

11

根据数据估计概率
正态分布:
每对(Xi，易)一个(收入，类别=否):如果类别=否
样本均值= 110 样本方差= 2975
Tid 退款婚姻
状态
应纳税的
收入逃避
1 是单人 125K 否
2 不结婚 10 万不
3 无单个 70K 否
4 是已婚 12 万否
5 不离婚 95K 是的
6 不结婚 6 万不
7 是离婚 22 万否
8 没有单人 85K 是的
9 不结婚 75K 不
没有单个 90K 是 10
2
2
2
( )
2
(|) ij
Xi ij
颈内

·100%
0.0072
2 (54.54)
(120 |)()()《☒☒损益表》

朴素贝叶斯分类器示例
十(退款否，离婚，收入 12 万)

P(X |否)= P(退款=否|否)
P(离异|否) P(收入=12 万|否)= 4/7 1/7 0.0072 = 0.0006
P(X |是)= P(退款=否|是)
P(离婚|是) P(收入=12 万|是)= 1 1/3 1.2 10= 4 10
因为 P(X |否)P(否)> P(X |是)P(是)因此 P(否|X) > P(是|X)
= >类=否
给定一个测试记录:
朴素贝叶斯分类器:
P(退款=是|否)= 3/7 P(退款=否|否)= 4/7 P(退款=是|是)= 0 P(退款=否|是)= 1
P(婚姻状况=单身|否)= 2/7 P(婚姻状况=离婚|否)= 1/7 P(婚姻状况=已婚|否)= 4/7 P(婚姻状况=单身|是)
= 2/3 P(婚姻状况=离婚|是)= 1/3 P(婚姻状况=已婚|是)= 0 对于应纳税收入:
如果类别=否:样本平均值= 110
样本方差= 2975
如果类别=是:样本平均值= 90
样本方差= 25

13

02/10/2020 数据挖掘导论，第 2 版 14
朴素贝叶斯分类器可以利用测试记录中属性的部分信息进行决策
P(是)= 3/10 P(否)= 7/10
如果我们只知道婚姻状况是离婚，那么:P(是|离婚)= 1/3 x 3/10 / P(离婚)P(否|离婚)= 1/7 x 7/10 / P(离婚)
如果我们也知道退款=否，那么
P(是|退款=否，离婚)= 1 x 1/3 x 3/10 /
(离婚，退款=否)
p(否|退款=否，离婚)= 4/7 x 1/7 x 7/10 /
(离婚，退款=否)
如果我们也知道应税收入= 120，那么
P(是|退款=否，离婚，收入= 120)= 1
1.2 x10x 1 x 1/3 x 3/10 /
P(离婚，退款=否，收入= 120) P(否|退款=否，离婚收入= 120) =
0.0072 x 4/7 x 1/7 x 7/10 /
(离婚，退款=否，收入= 120)
即使没有关于任何属性的信息，我们也可以使用类变量的先验概率:
朴素贝叶斯分类器:
P(退款=是|否)= 3/7 P(退款=否|否)= 4/7 P(退款=是|是)= 0 P(退款=否|是)= 1
P(婚姻状况=单身|否)= 2/7 P(婚姻状况=离婚|否)= 1/7 P(婚姻状况=已婚|否)= 4/7 P(婚姻状况=单身|是)
= 2/3 P(婚姻状况=离婚|是)= 1/3 P(婚姻状况=已婚|是)= 0 对于应纳税收入:
如果类别=否:样本平均值= 110
样本方差= 2975
如果类别=是:样本平均值= 90
样本方差= 25

02/10/2020 数据挖掘导论，第 2 版 15
朴素贝叶斯分类器示例

十(退款否，离婚，收入 12 万)
P(是)= 3/10 P(否)= 7/10
P(是|离婚)= 1/3 x 3/10 / P(离婚)P(否|离婚)= 1/7 x 7/10 / P(离婚)
P(是|退款=否，离婚)= 1 x 1/3 x 3/10 /
(离婚，退款=否)
p(否|退款=否，离婚)= 4/7 x 1/7 x 7/10 /
(离婚，退款=否)
给定一个测试记录:
朴素贝叶斯分类器:
P(退款=是|否)= 3/7 P(退款=否|否)= 4/7 P(退款=是|是)= 0 P(退款=否|是)= 1
P(婚姻状况=单身|否)= 2/7 P(婚姻状况=离婚|否)= 1/7 P(婚姻状况=已婚|否)= 4/7 P(婚姻状况=单身|是)
= 2/3 P(婚姻状况=离婚|是)= 1/3 P(婚姻状况=已婚|是)= 0 对于应纳税收入:
如果类别=否:样本平均值= 110
样本方差= 2975
如果类别=是:样本平均值= 90
样本方差= 25

朴素贝叶斯分类器的问题
P(是)= 3/10 P(否)= 7/10
P(是|已婚)= 0 x 3/10 / P(已婚)P(否|已婚)= 4/7 x 7/10 / P(已婚)
朴素贝叶斯分类器:
P(退款=是|否)= 3/7 P(退款=否|否)= 4/7 P(退款=是|是)= 0 P(退款=否|是)= 1
P(婚姻状况=单身|否)= 2/7 P(婚姻状况=离婚|否)= 1/7 P(婚姻状况=已婚|否)= 4/7 P(婚姻状况=单身|是)
= 2/3 P(婚姻状况=离婚|是)= 1/3 P(婚姻状况=已婚|是)= 0 对于应纳税收入:
如果类别=否:样本平均值= 110
样本方差= 2975
如果类别=是:样本平均值= 90
样本方差= 25

朴素贝叶斯分类器的问题
Tid 退款婚姻
状态
应纳税的
收入逃避
1 是单人 125K 否
2 不结婚 10 万不
3 无单个 70K 否
4 是已婚 12 万否
5 不离婚 95K 是的
6 不结婚 6 万不
7 是离婚 22 万否
8 没有单人 85K 是的

9 不结婚 75K 不
没有单个 90K 是 10
朴素贝叶斯分类器:
P(退款=是|否)= 2/6 P(退款=否|否)= 4/6 P(退款=是|是)= 0 P(退款=否|是)= 1
P(婚姻状况=单身|否)= 2/6 P(婚姻状况=离婚|否)= 0 P(婚姻状况=已婚|否)= 4/6 P(婚姻状况=单身|是)= 2/3 P(婚姻状况=离婚|是)= 1/3 P(婚姻状况=已婚|是)= 0/3 应税收入:
如果类别=否:样本平均值= 91
样本方差= 685
如果类别=否:样本平均值= 90
样本方差= 25
考虑删除 Tid = 7 的表
给定 X =(退款=是，离婚，12 万)
P(X |否)= 2/6 X 0 X 0.0083 = 0 P(X |是)= 0 X 1/3 X 1.2 X 10= 0
天真的贝叶斯无法将 X 分类为是或否!

朴素贝叶斯分类器的问题
如果其中一个条件概率为零，则整个表达式为零
需要使用简单分数以外的其他条件概率估计概率估计:
n:属于 Y 类 nc 的训练实例数:Xi = c 且 Y = y 的实例数
v:Xi 可以接受的属性值总数 P:对(P(Xi = c|y)的初始估计已知先验 m:我们对 P 的信心的超参数
拉普拉斯估计:$PX = cy)$= 1
$n + 1\ n + v$
m 估计值:$PX = cy)$= 1
$n + mp\ n + m$
原件:$PX = cy)$= 1
$n\ n$

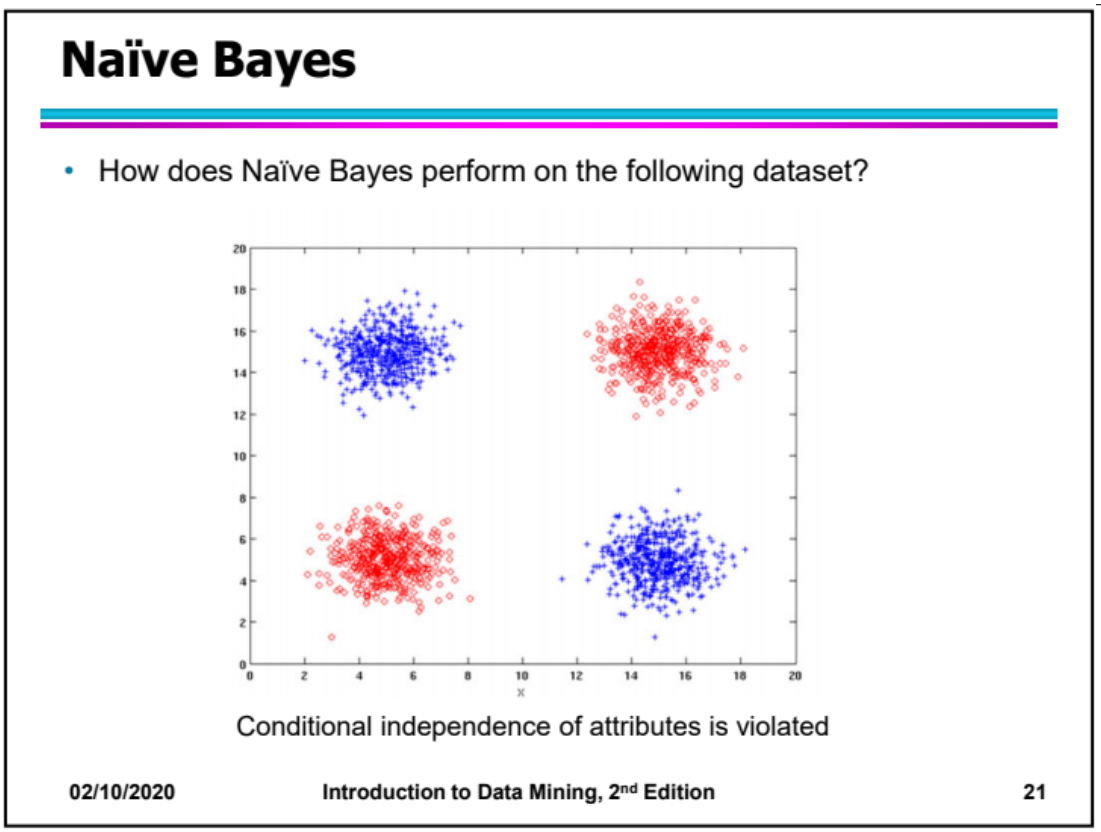
朴素贝叶斯分类器示例
名生能飞能在水中生存有腿类
人类是不是不是哺乳动物
蟒蛇不不不不非哺乳动物
鲑鱼不，不，是的，不，非哺乳动物
鲸鱼是不是是不是哺乳动物
青蛙不不有时是的非哺乳动物
科莫多不不不是的非哺乳动物
蝙蝠是是是否是哺乳动物
鸽子否是否是非哺乳动物
猫是不是不是哺乳动物
豹鲨是否是否非哺乳动物
乌龟不，不，有时是的，非哺乳动物
企鹅不不有时是的非哺乳动物
豪猪是不是不是哺乳动物

鳗鱼不，不，是的，不，非哺乳动物
蝾螈不，不，有时是的，非哺乳动物
吉拉怪物不不不是的非哺乳动物
鸭嘴兽不不不是的哺乳动物
猫头鹰否是否是非哺乳动物
海豚是否是否哺乳动物
鹰否是否是非哺乳动物
生孩子可以在水中生活有腿类
是不是是不是?
0.0027 20
(|) ( ) 0.004
0.021 20
(|) ( ) 0.06
0.0042 13
4
13
3
13
10 13
(|)
0.06 7
2 7
2 7
6 7
(|)

名词短语名词短语
下午
附注
下午
属性 M:哺乳动物 N:非哺乳动物
P(A|M)P(M) > P(A|N)P(N) = >哺乳动物

19

02/10/2020 数据挖掘导论，第二版 20
朴素贝叶斯(摘要)
对隔离噪声点具有鲁棒性
通过在概率估计计算过程中忽略实例来处理缺失值对不相关的属性具有鲁棒性
冗余和相关属性将违反类别条件假设
使用其他技术，如贝叶斯信念网络(BBN)



21

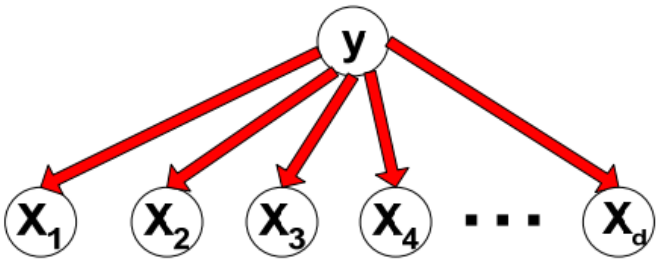
02/10/2020 数据挖掘导论，第 2 版 22
贝叶斯信念网络
提供一组随机变量之间概率关系的图形表示。包括:
有向无环图
节点对应一个变量，⊠弧对应一对变量之间的依赖关系
将每个节点与其直接父节点相关联的概率表
甲乙
C

02/10/2020 数据挖掘导论，第 2 版 23
条件独立性
贝叶斯网络中的一个节点有条件地独立于它的所有非感知节点，如果它的父节点是已知的
甲是丙的孩子
乙是丁的后代丁是甲的祖先

23

Conditional Independence

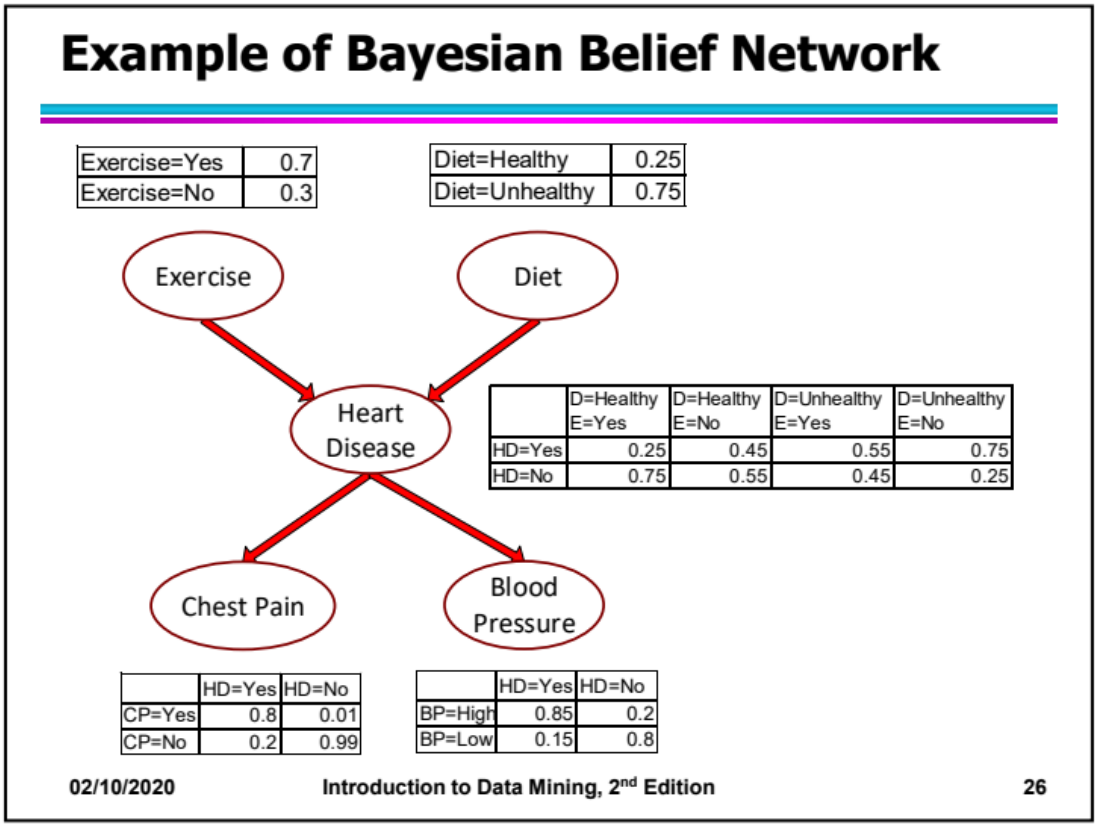
- Naïve Bayes assumption:



24

02/10/2020 数据挖掘导论，第 2 版 25
概率表
如果 X 没有任何父代，表中包含先验概率 P(X)
如果 X 只有一个父(Y)，则表包含条件概率 P(X|Y)
如果 X 有多个父代(Y1，Y2，…，Yk)，则表包含条件概率 P(X|Y1，Y2，…，Yk)

25



26

02/10/2020 数据挖掘导论，第 2 版 27

使用 BBN 推理的例子

给定:X =(E =否，D =是，CP =是，BP =高)计算 P(HD|E，D，CP，BP)？ P(高=是| E =否，D =是)= 0.55 P(低=是|高=是)= 0.8 P(高=高|高=是)= 0.85

p(HD =是| e =否，d =是，cp =是，bp =高)　0.55　0.8　0.85 = 0.374

P(高密度=否| E =否，D =是)= 0.45 P(低密度=是|高密度=否)= 0.01 P(高密度=否)= 0.2

p(HD =否| e =否，d =是，cp =是，bp =高)　0.45　0.01　0.2 = 0.0009

将 X 分类为是

27