WDS

DataScience@Web

机器学习讨论班

2018年暑期

2. 朴素贝叶斯法

耿苹

介绍内容

- 贝叶斯决策论
- 朴素贝叶斯分类器
- 极大似然估计
- EM（期望最大化）算法
- 贝叶斯网络

分类问题

以多分类任务为例

已知有N个类别 $\gamma = \{c_1, c_2, ..., c_N\}$ ，样本集为 χ

寻找判定准则 $h: \chi \to \gamma$ ，每一个样本 x有且仅有一个类别 c

贝叶斯决策论

基于相关概率和误判损失来进行分类

将真实样本标记为cj的样本误分类为ci的损失

后验概率，样本x类别为cj的概率

条件风险（期望损失） $R(c_i \mid x) = \sum_{j=1}^{N} \lambda_{ij} P(c_j \mid x)$

贝叶斯最优分类器 $h^*(x) = \arg\min_{c \in \gamma} R(c \mid x)$

若λ_{ij}为0-1损失 $\lambda_{ij} = \begin{cases} 0, & if\ i = j; \\ 1, & otherwise \end{cases}$，则

$$h^*(x) = \arg\min_{c \in \gamma} (1 - P(c \mid x)) = \arg\max_{c \in \gamma} P(c \mid x)$$

对每个样本x，选择后验概率 $P(c \mid x)$ 最大的类别标记

贝叶斯决策论

根据贝叶斯定理

$$P(c|\mathbf{x}) = \frac{P(c)P(\mathbf{x}|c)}{P(\mathbf{x})}$$

$P(c)$ 类“先验”概率

$P(\mathbf{x}|c)$ 样本$\mathbf{x}$相对于类标记c的类条件概率（似然函数）

$P(\mathbf{x})$ 用于归一化的“证据”因子，对于所有的c都相同

朴素贝叶斯分类器

- 基于贝叶斯定理 与特征条件独立假设的分类方法
- 特征条件独立假设是指用于分类的特征在类确定的条件下都是条件独立的。

$$P(\mathbf{x} \mid c_i) = P(x^1, x^2 ..., x^d \mid c_i)$$

- 似然函数：

$$= \prod_{j=1}^{d} P(x^j \mid c_i)$$

e.g.

西瓜x　　特征集{x^1(色泽)，x^2(根蒂)，x^3(敲声)，x^4(纹理)...}，　特征条件相互独立

朴素贝叶斯分类器

- 条件独立假设

$$P(\mathbf{x} \mid c_i) = \prod_{j=1}^{d} P(x^j \mid c_i)$$

- 贝叶斯定理

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

- 最大化后验概率

$$h^*(\mathbf{x}) = \arg\max_{c \in \gamma} P(c \mid \mathbf{x})$$

$$h^*(\mathbf{x}) = \arg\max_{c \in \gamma} P(c) \prod_{i=1}^{d} P(x^i \mid c)$$

x	x^1色泽	x^2根蒂	x^3敲声	类别（是否为好瓜）
x1	2	0	1	1
x2	1	1	3	1
x3	2	0	2	1
x4	3	0	3	0
x5	1	1	3	1
x6	2	1	1	0

e.g.

判断 x=(2,1,3)是否为好瓜？

P(c=1)=4/6

P(c=0)=2/6

$P(x^1=2|c=1)=2/4$　　$P(x^1=2|c=0)=1/2$

$P(x^2=1|c=1)=2/4$　　$P(x^2=1|c=0)=1/2$

$P(x^3=3|c=1)=2/4$　　$P(x^3=3|c=0)=1/2$

计算 $P(c)\prod_{i=1}^{d}P(x^i|c)$

$P(c=1)*P(x^1=2|c=1)*P(x^2=1|c=1)*P(x^3=3|c=1)=0.0833$

$P(c=0)*P(x^1=2|c=0)*P(x^2=1|c=0)*P(x^3=3|c=0)=0.0417$

因此，朴素贝叶斯分类器将测试样本判别为c=1

朴素贝叶斯分类器

e.g.

判断 样本x={1,1,3}的类别呢？

$P(x^1=1|c=0) = 0 / 2 = 0$

修正后：

$P(c=0)=(2+1)/(6+2)=3/8$

$P(x1=1|c=0) =(0+1)/(2+3)$

$= 1/5$

拉普拉斯修正:

$$P(c)=\frac{|D_c|+1}{|D|+N}$$

$$P(x^i|c)=\frac{|D_{cx_i}|+1}{|D|+N_i}$$

N：训练集D中可能的类别数

Ni：第i个属性可能的取值数

朴素贝叶斯分类器

- 多种使用方式：

  (1)“查表”

  计算所有概率估值并存储，通过“查表”进行判别

  (2)懒惰学习

  先不进行任何训练，待收到预测请求时在根据当前数据集进行概率估值

  (3)增量学习

  训练集不断增加，在现有估值基础上计算进行计数修正

  (4)半朴素贝叶斯

  常用策略是假设每个属性在类别之外最多依赖于一个其他属性

朴素贝叶斯分类器

- 应用场景：垃圾邮件过滤，文本分类，新闻分类，Query分类，商品分类等

- 优点：

1）生成式模型，通过计算概率来进行分类，可以用来处理多分类问题

2）对小规模的数据表现很好

3）适合增量式训练

4）算法也比较简单，是经典常用的分类算法

朴素贝叶斯分类器

- 缺点：

1）对缺失数据不太敏感，常用于文本分类

2）输入数据的表达形式很敏感

3）特征条件独立假设在实际应用中往往是不成立的。可以考虑**半朴素贝叶斯**方法

4）需要知道先验概率，且先验概率很多时候取决于假设

5）分类决策存在错误率

朴素贝叶斯分类器

- 增量学习过程

1. 观察样本前，先验概率为 $p(c \mid D^0) = p(c)$

   (D表示每一次的训练集，$D^n=\{x_1,x_2,...,x_n\}$)

2. 观察样本 $\mathbf{x}_1$ $p(c \mid D^1) = \dfrac{p(c \mid D^0)p(\mathbf{x}_1 \mid c)}{p(x_1 \mid D^0)}$

3. 观察样本 $\mathbf{x}_2$ $p(c \mid D^2) = \dfrac{p(c \mid D^1)p(\mathbf{x}_2 \mid c)}{p(x_2 \mid D^1)}$

……

n. 观察样本 $\mathbf{x}_n$ $p(c \mid D^n) = \dfrac{p(c \mid D^n)p(\mathbf{x}_{n-1} \mid c)}{p(x_{n-1} \mid D^n)}$

上一步的后验概率
作为下一步的先验

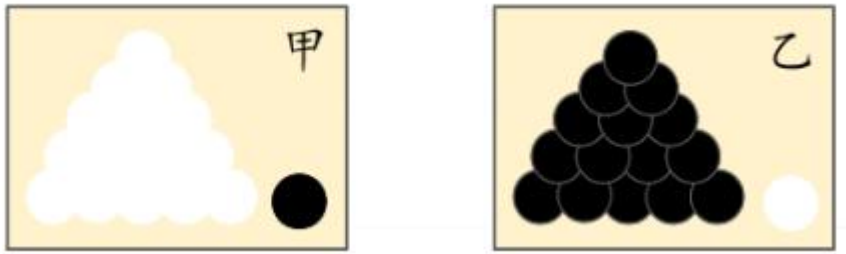
最大似然估计

- 利用已知的样本结果，反推最有可能（最大概率）导致这样结果的参数值

e.g.已知甲乙两个箱子，甲装有99白球1黑球，乙装有99黑球1白球。

问：随机取一个黑球，是从哪一个盒子中取出来的？

P（黑|甲）=0.01　　P（黑|乙）=0.99



最大似然估计

e.g.

随机找100个男生和100个女生。按照性别划分为两组，然后先统计抽样得到的100个男生的身高。假设身高服从高斯分布，但分布的均值μ和方差**Σ**未知。

估计样本分布参数，记作θ=[μ, **Σ**]

$$L(\theta) = P(x_1, x_2, ..., x_n \mid \theta) = \prod_{i=1}^{n} P(x_n \mid \theta)$$

$$\hat{\theta} = \arg\max linL(\theta)$$

使**似然函数**达到**最大**的参数值θ,就是极大似然估计值

EM算法

- 不完全数据或有数据丢失的数据集（存在隐含变量）
- 求解概率模型参数的最大似然估计方法

e.g.

西瓜的特征有{x^1(色泽)，x^2(根蒂)，x^3(敲声)，x^4(纹理)...}，但是瓜根蒂掉了，不知道是“蜷缩”还是“硬挺”。

随机找100个男生和100个女生测量身高，但样本记录时没有记录性别。

这些未观测的变量称为“隐变量”

EM算法

- **X**表示已观测变量集，**Z**表示隐变量集，θ表示模型参数

$$L(\theta \mid \mathbf{X}) = \ln P(\mathbf{X} \mid \theta) \quad \rightarrow \quad L(\theta \mid \mathbf{X}, \mathbf{Z}) = \ln P(\mathbf{X}, \mathbf{Z} \mid \theta)$$

由于Z是隐变量，这个式子无法求解

- 可以通过对Z计算期望，来最大化一观测数据的“边际似然”

$$L(\theta \mid \mathbf{X}) = \ln P(\mathbf{X} \mid \theta) = \ln \sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z} \mid \theta)$$

- 如果不是直接计算**Z**的期望，而是**Z**的概率分布，再计算L(θ)关于**Z**的期望

$$Q(\theta \mid \mathbf{X}) = E_{\mathbf{Z}|\mathbf{X}} L(\theta \mid \mathbf{X}, \mathbf{Z})$$

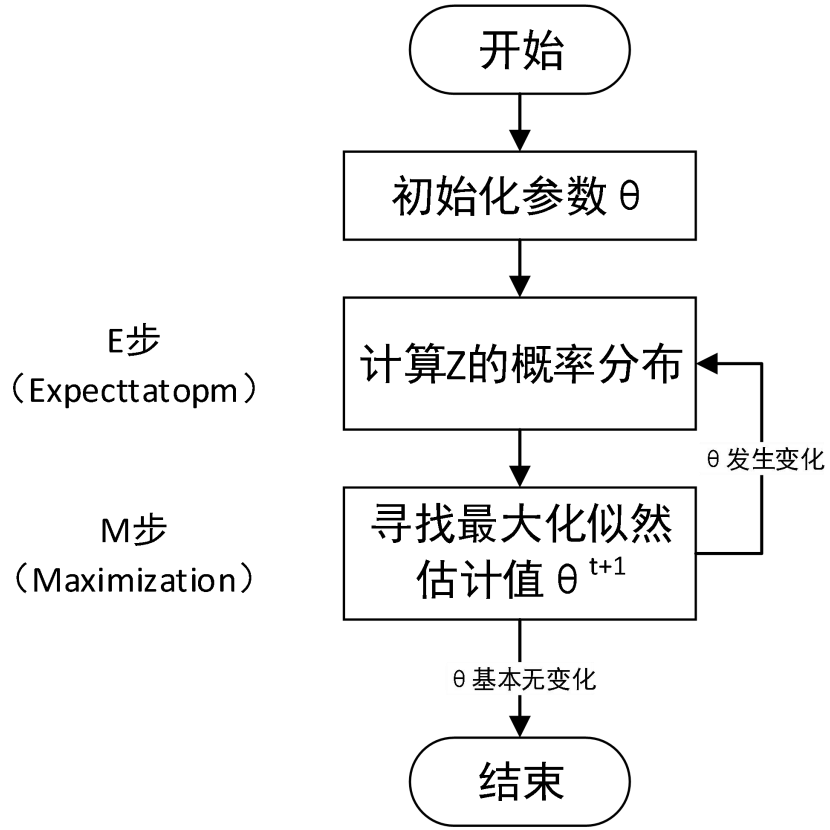
EM算法

E步（Expectation）：以当前参数θ^{t}推断隐变量分布P（Z|X,θ），并计算对数似然L(θ|X,Z)关于Z的期望

$$Q(\theta \mid \theta^t) = E_{\mathbf{Z}|\mathbf{X},\theta^t} L(\theta \mid \mathbf{X}, \mathbf{Z})$$

M步（Maximization）:寻找参数最大化期望似然，即

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta \mid \theta^t)$$



EM算法

- 应用领域：参数估计，计算机视觉的数据集聚
- 缺点

  (1)迭代速度慢，次数多，不适合大规模数据集和高维数据

  (2)当目标函数不是凸函数时，容易给出局部最佳解，而不是最优解
- 优点

  (1)简单稳定

  (2)自收敛的分类算法，不需要事先设定类别

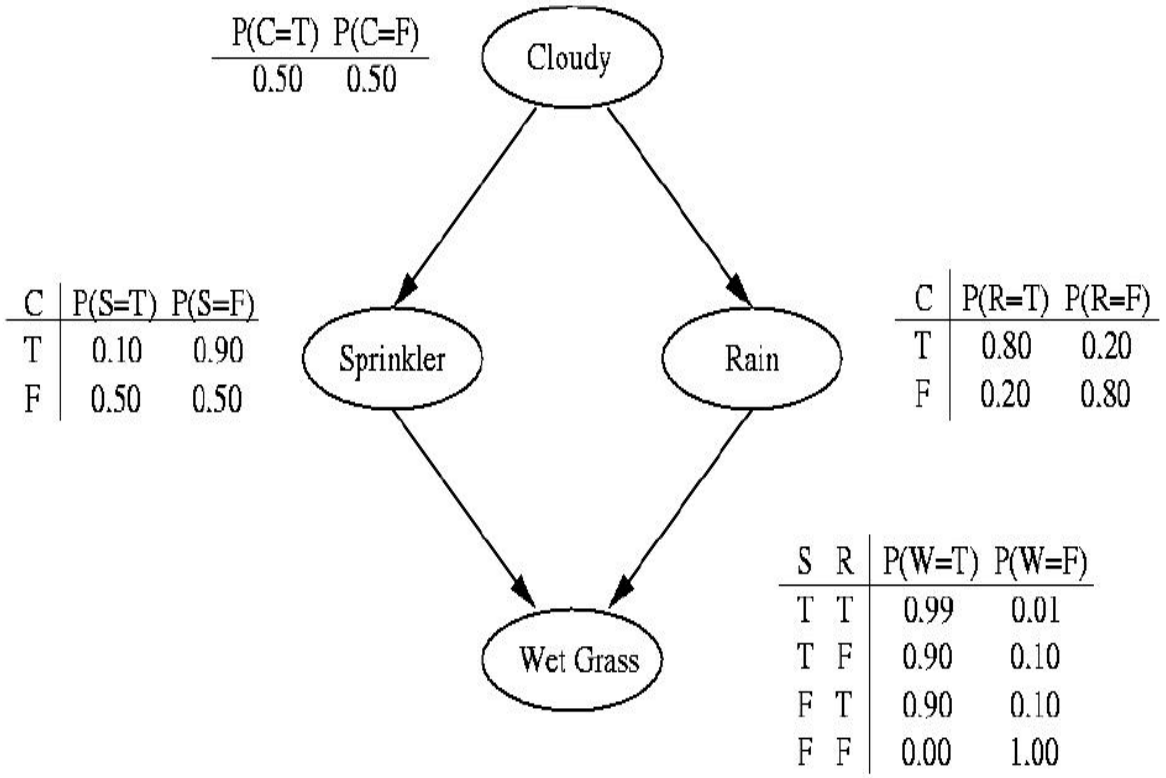
贝叶斯网络

借助有向无环图刻画属性之间的依赖关系
使用条件概率表描述属性的联合概率分布

一个贝叶斯网络B由有向无环图G和条件概率表 Θ两部分构成

$$B = < G, \Theta >$$

e.g.草地变湿可能有两个原因：
洒水装置打开过，或者下过雨



贝叶斯网络

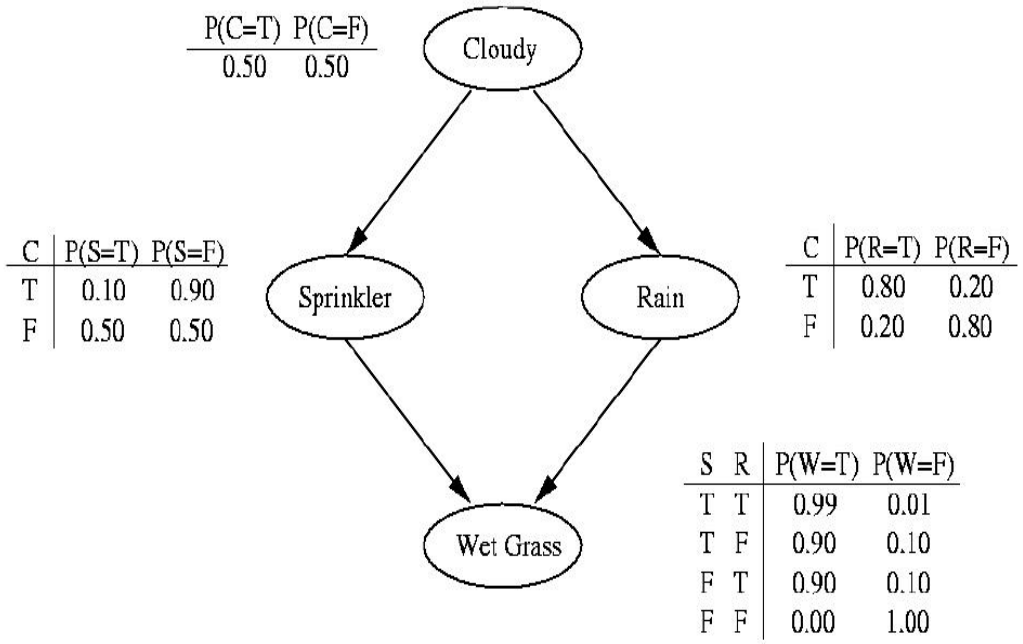
e.g. 如果看到草地是湿的，判断洒水装置和下雨哪个原因更为可能？

$$P(S \mid W) = \frac{P(S,W)}{P(W)} = \frac{\sum_c \sum_r P(W \mid S,r)P(S \mid c)P(r \mid c)P(c)}{P(S,W) + P(\neg S,W)}$$

$$= \frac{0.2781}{0.2781 + 0.369} = 0.430$$

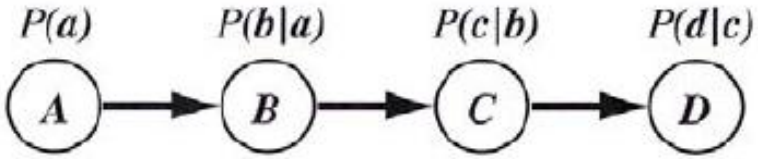
$$P(R \mid W) = \frac{P(R,W)}{P(W)} = \frac{\sum_c \sum_s P(W \mid s,R)P(s \mid c)P(R \mid c)P(c)}{P(R,W) + P(\neg R,W)}$$

$$= \frac{0.4581}{0.6471} = 0.708$$



因为 $P(S \mid W) < P(R \mid W)$ ，所以下雨造成草地湿的可能性更大

贝叶斯网络



$$P(a,b,c,d) = P(a)P(b \mid a)P(c \mid b)P(d \mid c)$$

$$P(b,c,d) = P(c \mid b)P(d \mid c)\sum_{a} P(a)P(b \mid a)$$

$$P(c,d) = P(d \mid c)\sum_{a}\sum_{b} P(a)P(b \mid a)P(c \mid b)$$

课后练习：

1.用Python实现朴素贝叶斯分类器对垃圾邮件的分类，识别文本中的侮辱性文本

（参考 https://www.cnblogs.com/pursued-deer/p/7783459.html）

2.贝叶斯置信网的应用

（参考 RichardO.Duda, Duda, Hart. 模式分类[M]. 机械工业出版社, 2004. P47）

3.SKlearn库EM算法

（参考 https://blog.csdn.net/lihou1987/article/details/70833229)

谢 谢