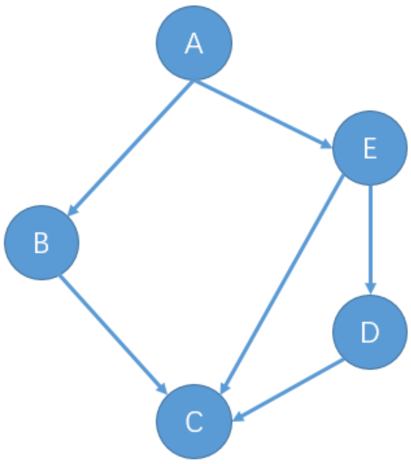
机器学习导论
作业五

151220097，孙旭东，248381185@qq.com

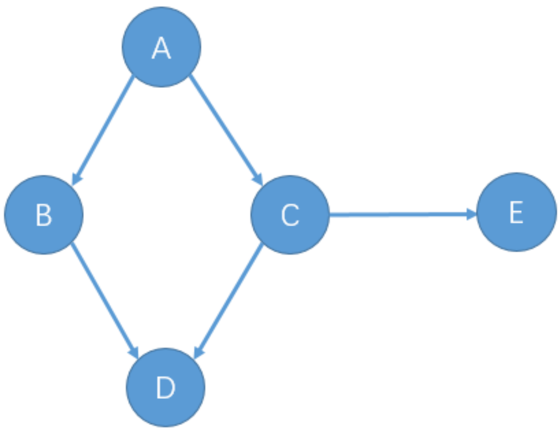
2018 年 6 月 13 日

1 [30pts] Conditional Independence in Bayesian Network

(1) **[5pts]** 请给出图中贝叶斯网结构的联合概率分布的分解表达式。



(2) **[5pts]** 请给出下图中按照道德化方法可以找到的所有条件独立的组合(即哪些变量关于哪些变量或者变量集条件独立)，独立也算做条件独立的一种特例。



(3) **[10pts]** 在这里，首先我们将给出关于“阻塞”的概念，然后我们根据“阻塞”的概念给出条件独立的充要条件。(大家也可以参考这个网站)

> **定义 1** (阻塞). *设X, Y, Z分别是一个有向无环图G里互没有交集的结点集，Z阻塞X中的一结点到Y中一结点的通路P(关于“通路”，在这里只要连通就算一条通路，对路中每条边的方向无任何要求)，当且仅当满足以下条件之一：*
>
> *1. P中存在顺序结构$x \to z \to y$或同父结构$x \leftarrow z \to y$，结点z包含在集合Z中；*
>
> *2. P中存在V型结构$x \to z \leftarrow y$，结点z及其孩子结点不包含在集合Z中。*

> **定理 1** (条件独立). *设X, Y, Z分别是一个有向无环图G里互没有交集的结点集，如果集合Z阻塞X到Y的任何一条道路，则X和Y在给定Z时条件独立，即$X \perp\!\!\!\perp Y | Z$。*

请根据定理1，判断第一问中有哪些条件独立的组合（独立也算条件独立的一种特例）。

(4) **[10pts]** 由以上两问我们可知，道德化方法中的“除去集合z后，x和y分属两个连通分支”并不构成条件独立性的充要条件。如果对道德化方法稍加修改，在连接V型结构父结点前，我们只保留图中X, Y, Z及他们的非孩子结点，之后的步骤则相同。请问你认为用修改后的方法可以保证得到全部的正确的条件独立集合吗？如果可以，请说明理由；如果不能，请给出反例。

Proof. *此处用于写证明(中英文均可)*

(1) 定义x_A表示A结点的属性值，x_B, x_C, x_D, x_E同理。

$$
\begin{aligned}
&P(x_A, x_B, x_C, x_D, x_E) \\
=&P(x_A)P(x_B|x_A)P(x_C|x_B, x_D, x_E)P(x_D|x_E)P(x_E|x_A)
\end{aligned}
\tag{1.1}
$$

(2) 条件独立的组合：

$$
\begin{aligned}
x_E &\perp x_A | x_C \\
x_E &\perp x_B | x_C \\
x_E &\perp x_D | x_C \\
x_A &\perp x_D | x_B, x_C
\end{aligned}
\tag{1.2}
$$

(3) 条件独立的组合：

$$
\begin{aligned}
x_A &\perp\!\!\!\perp x_C | x_B, x_E \\
x_A &\perp\!\!\!\perp x_D | x_B, x_E \\
x_B &\perp\!\!\!\perp x_D | x_A, x_E \\
x_B &\perp\!\!\!\perp x_E | x_A, x_D
\end{aligned}
\tag{1.3}
$$

2 [20pts] Naive Bayes Classifier

通过对课本的学习，我们了解了采用“属性条件独立性假设”的朴素贝叶斯分类器。现在我们有如下表所示的一个数据集：

表 1: 数据集

编号	x_1	x_2	x_3	x_4	y
样本1	1	1	1	0	1
样本2	1	1	0	0	0
样本3	0	0	1	1	0
样本4	1	0	1	1	1
样本5	0	0	1	1	1

(1) **[10pts]** 试计算：$\Pr\{y=1|\mathbf{x}=(1,1,0,1)\}$ 与 $\Pr\{y=0|\mathbf{x}=(1,1,0,1)\}$ 的值；

(2) **[10pts]** 使用“拉普拉斯修正”之后，再重新计算上一问中的值。

Solution.

(1)

$$\Pr\{y=1|\mathbf{x}=(1,1,0,1)\} = \frac{\Pr\{\mathbf{x}=(1,1,0,1)|y=1\}\Pr\{y=1\}}{\Pr\{\mathbf{x}=(1,1,0,1)\}} \tag{2.1}$$

根据“属性条件独立性假设”的思想，有：

$$\begin{aligned}
&\Pr\{\mathbf{x}=(1,1,0,1)|y=1\}\\
=&\Pr\{x_1=1|y=1\}\Pr\{x_2=1|y=1\}\Pr\{x_3=0|y=1\}\Pr\{x_4=1|y=1\}\\
=&\frac{2}{3}\times\frac{1}{3}\times\frac{0}{3}\times\frac{2}{3}\\
=&0
\end{aligned} \tag{2.2}$$

以及

$$\begin{aligned}
&\Pr\{\mathbf{x}=(1,1,0,1)\}\\
=&\Pr\{x_1=1\}\Pr\{x_2=1\}\Pr\{x_3=0\}\Pr\{x_4=1\}\\
=&\frac{3}{5}\times\frac{2}{5}\times\frac{1}{5}\times\frac{3}{5}\\
=&\frac{18}{625}
\end{aligned} \tag{2.3}$$

所以可得：

$$\Pr\{y=1|\mathbf{x}=(1,1,0,1)\} = 0 \tag{2.4}$$

类似的

$$\Pr\{y=0|\mathbf{x}=(1,1,0,1)\} = \frac{\Pr\{\mathbf{x}=(1,1,0,1)|y=0\}\Pr\{y=0\}}{\Pr\{\mathbf{x}=(1,1,0,1)\}} \tag{2.5}$$

根据“属性条件独立性假设”的思想，有：

$$\begin{aligned}&\Pr\{\mathbf{x}=(1,1,0,1)|y=0\}\\=&\Pr\{x_1=1|y=0\}\Pr\{x_2=1|y=0\}\Pr\{x_3=0|y=0\}\Pr\{x_4=1|y=0\}\\=&\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\\=&\frac{1}{16}\end{aligned}\tag{2.6}$$

以及

$$\Pr\{\mathbf{x}=(1,1,0,1)\}=\frac{18}{625}\tag{2.7}$$

又因为

$$\Pr\{y=0\}=\frac{2}{5}\tag{2.8}$$

综上可得

$$\begin{aligned}&\Pr\{y=0|\mathbf{x}=(1,1,0,1)\}\\=&\frac{\Pr\{\mathbf{x}=(1,1,0,1)|y=0\}\Pr\{y=0\}}{\Pr\{\mathbf{x}=(1,1,0,1)\}}\\=&\frac{\frac{1}{16}\times\frac{2}{5}}{\frac{18}{625}}\\=&\frac{125}{144}\end{aligned}\tag{2.9}$$

(2) 首先有：

$$\Pr\{\mathbf{x}=(1,1,0,1)\}=\frac{18}{625}\tag{2.10}$$

根据上题，结合“拉普拉斯修正”，可得：

$$\begin{aligned}&\Pr\{\mathbf{x}=(1,1,0,1)|y=1\}\\=&\Pr\{x_1=1|y=1\}\Pr\{x_2=1|y=1\}\Pr\{x_3=0|y=1\}\Pr\{x_4=1|y=1\}\\=&\frac{3}{5}\times\frac{2}{5}\times\frac{1}{5}\times\frac{3}{5}\\=&\frac{18}{625}\end{aligned}\tag{2.11}$$

还有：

$$\Pr\{y=1\}=\frac{4}{7}\tag{2.12}$$

所以得到：

$$\Pr\{y=1|\mathbf{x}=(1,1,0,1)\}=\frac{4}{7}\tag{2.13}$$

类似的，根据”拉普拉斯修正”，可得：

$$\begin{aligned}&\Pr\{\mathbf{x}=(1,1,0,1)|y=0\}\\=&\Pr\{x_1=1|y=0\}\Pr\{x_2=1|y=0\}\Pr\{x_3=0|y=0\}\Pr\{x_4=1|y=0\}\\=&\frac{2}{4}\times\frac{2}{4}\times\frac{2}{4}\times\frac{2}{4}\\=&\frac{1}{16}\end{aligned}\tag{2.14}$$

还有：

$$\Pr\{y=0\}=\frac{3}{7} \tag{2.15}$$

所以得到：

$$\Pr\{y=0|\mathbf{x}=(1,1,0,1)\}=\frac{625}{672} \tag{2.16}$$

3 [50pts] Ensemble Methods in Practice

由于出色的性能和良好的鲁棒性，集成学习方法 (Ensemble methods) 成为了极受欢迎的机器学习方法，在各大机器学习比赛中也经常出现集成学习的身影。在本次实验中我们将结合两种经典的集成学习思想：Boosting和Bagging，对集成学习方法进行实践。

本次实验选取UCI数据集Adult，此数据集为一个二分类数据集，具体信息可参照链接，为了方便大家使用数据集，已经提前对数据集稍作处理，并划分为训练集和测试集，大家可通过此链接进行下载。

由于Adult是一个类别不平衡数据集，本次实验选用AUC作为评价分类器性能的评价指标，AUC指标的计算可调用sklearn算法包。

(1) **[5pts]** 本次实验要求使用Python 3或者Matlab编写，要求代码分布于两个文件中，BoostMain.py、RandomForestMain.py (Python) 或 BoostMain.m、RandomForestMain.m (Matlab)，调用这两个文件就能完成一次所实现分类器的训练和测试；

(2) **[35pts]** 本次实验要求编程实现如下功能：

- **[10pts]** 结合教材8.2节中图8.3所示的算法伪代码实现AdaBoost算法，基分类器选用决策树，基分类器可调用sklearn中决策树的实现；
- **[10pts]** 结合教材8.3.2节所述，实现随机森林算法，基分类器仍可调用sklearn中决策树的实现，当然也可以自行手动实现，在实验报告中请给出随机森林的算法伪代码；
- **[10pts]** 结合AdaBoost和随机森林的实现，调查基学习器数量对分类器训练效果的影响 (参数调查)，具体操作如下：分别对AdaBoost和随机森林，给定基分类器数目，在训练数据集上用5折交叉验证得到验证AUC评价。在实验报告中用折线图的形式报告实验结果，折线图横轴为基分类器数目，纵轴为AUC指标，图中有两条线分别对应AdaBoost和随机森林，基分类器数目选取范围请自行决定；
- **[5pts]** 根据参数调查结果，对AdaBoost和随机森林选取最好的基分类器数目，在训练数据集上进行训练，在实验报告中报告在测试集上的AUC指标；

(3) **[10pts]** 在实验报告中，除了报告上述要求报告的内容外还需要展现实验过程，实验报告需要有层次和条理性，能让读者仅通过实验报告便能了解实验的目的，过程和结果。

实验报告.