Learning And Mining from DatA
http://lamda.nju.edu.cn

机器学习导论

习题课

詹德川
zhandc@lamda.nju.edu.cn
2018年6月20日

大纲

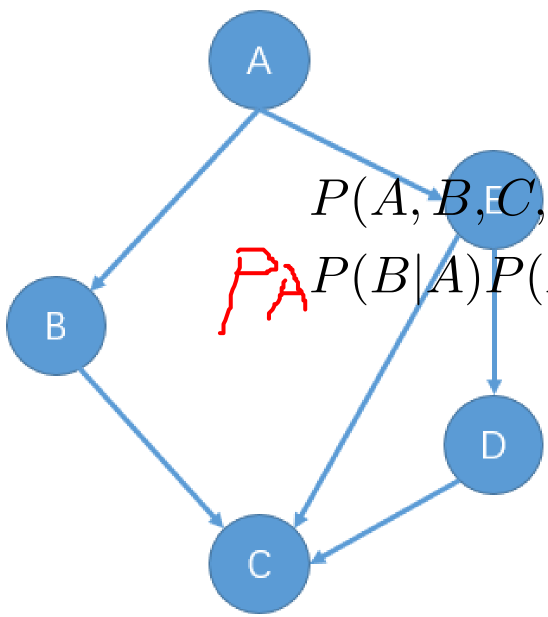
- 第五次作业
  - PS1 - Conditional Independence in Bayesian Network
  - PS2 - Naive Bayes Classifier
  - PS3 - Ensemble Methods in Practice

PS1 – Conditional Independence in Bayesian Network

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

Solution.



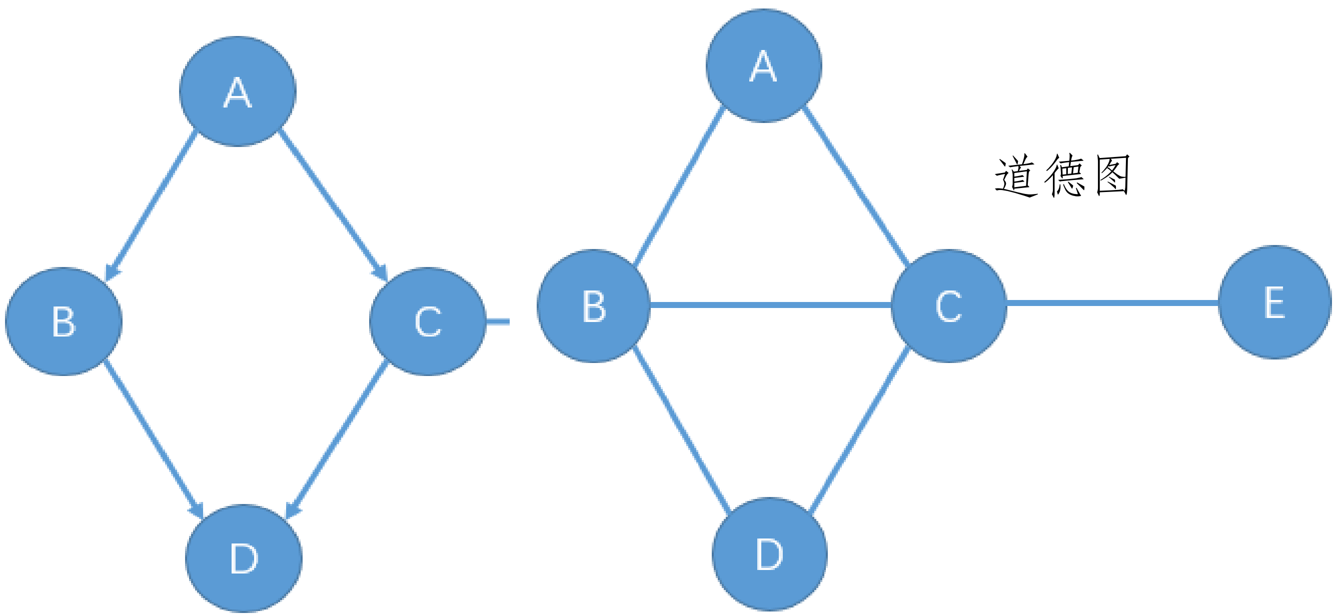
$$P(A,B,C,D,E) =$$

PA $P(B|A)P(E|A)P(D|E)P(C|B,D,E)$

PS1 – Conditional Independence in Bayesian Network

(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

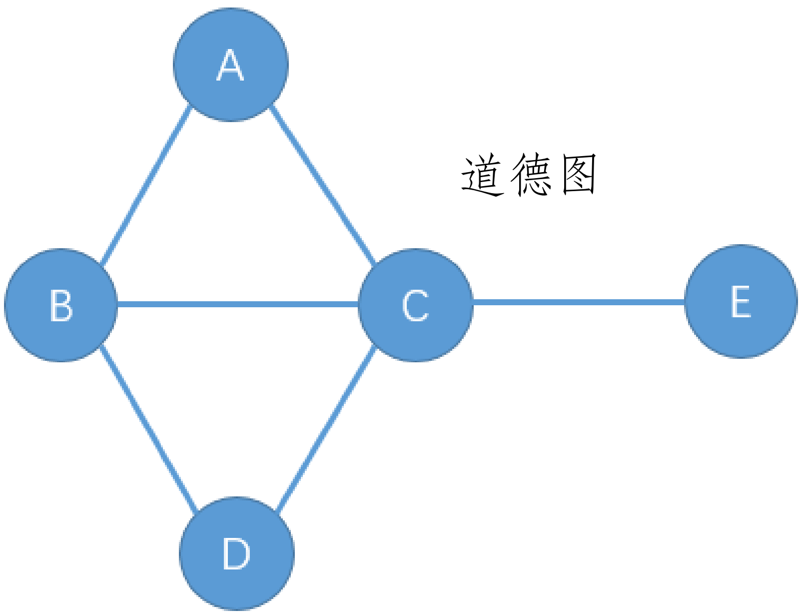
Solution.



PS1 – Conditional Independence in Bayesian Network

Learning And Mining from DatA

(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



道德图

$$(A \perp\!\!\!\perp D \mid B, C)$$

$$(A \perp\!\!\!\perp D \mid B, C, E)$$

$$(E \perp\!\!\!\perp others \mid C)$$

PS1 – Conditional Independence in Bayesian Network

(3) [10pts] 请在这里，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；请根据定理1，判断第二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)，只考虑X和Y是单变量即可

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 $i \to z \to j$ 或同父结构 $i \leftarrow z \to j$，结点 z 包含在集合 Z 中；
2. P 中存在 V 型结构 $i \to z \leftarrow j$，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 $X \perp\!\!\!\perp Y | Z$。

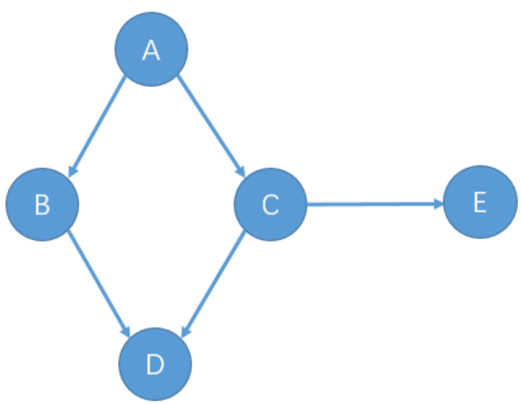
PS1 – Conditional Independence in Bayesian Network

(3)

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 $i \to z \to j$ 或同父结构 $i \leftarrow z \to j$, 结点 z 包含在集合 Z 中;
2. P 中存在 V 型结构 $i \to z \leftarrow j$, 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 $X \perp\!\!\!\perp Y \mid Z$。



Solution.

回顾第二问的结论:

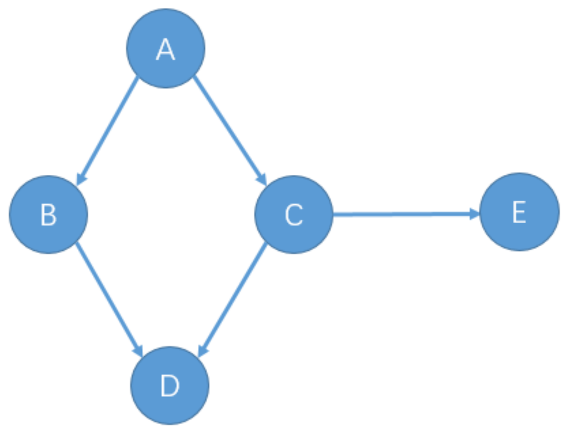
$$(A \perp\!\!\!\perp D \mid B, C)$$

$$(A \perp\!\!\!\perp D \mid B, C, E)$$

$$(E \perp\!\!\!\perp others \mid C)$$

PS1 – Conditional Independence in Bayesian Network

(3)



回顾第二问的结论：

$$(A \perp\!\!\!\perp D \mid B, C)$$

$$(A \perp\!\!\!\perp D \mid B, C, E)$$

$$(E \perp\!\!\!\perp others \mid C)$$

除了第一问之外，还有 $B \perp\!\!\!\perp C \mid A$

B 与 C 有两条通路，$B - A - C$ 和 $B - D - C$。A 点阻断了 $B \leftarrow A \rightarrow C$ 的路（它在同父结构中心），同时没有出现在 $B \rightarrow D \leftarrow C$ 的 V 型结构中心和中心子节点，所以也阻断了路 $B - D - C$。所以 A 点阻断了所有 B、C 的通路，B 和 C 在给定 A 的条件下独立。（这个题旨在让大家掌握贝叶斯网变量间条件独立性的充要条件）

PS2 – Naive Bayes Classifier

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：

Table 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。

PS2 – Naive Bayes Classifier

(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；

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

Solution.

$$\Pr\{y=1|\mathbf{x}=(1,1,0,1)\} = \frac{P(y)}{P(\mathbf{x})}\prod_{i=1}^{4}P(x_i|y)$$

$$= \frac{\Pr\{y=1\}}{\Pr\{\mathbf{x}=(1,1,0,1)\}}\Pr\{x_1=1|y=1\}\Pr\{x_2=1|y=1\}\Pr\{x_3=0|y=1\}\Pr\{x_4=1|y=1\}$$

$$= \frac{3/5}{P(\mathbf{x})}\times\frac{2}{3}\times\frac{1}{3}\times\frac{0}{3}\times\frac{2}{3}$$

$$= 0/P(\mathbf{x})$$

PS2 – Naive Bayes Classifier

(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；

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

$$\Pr\{y=0|\mathbf{x}=(1,1,0,1)\}=\frac{P(y)}{P(\mathbf{x})}\prod_{i=1}^{4}P(x_i|y)$$

$$=\frac{\Pr\{y=0\}}{\Pr\{\mathbf{x}=(1,1,0,1)\}}\Pr\{x_1=1|y=0\}\Pr\{x_2=1|y=0\}\Pr\{x_3=0|y=0\}\Pr\{x_4=1|y=0\}$$

$$=\frac{2/5}{P(\mathbf{x})}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}$$

$$=\frac{1}{40}/P(\mathbf{x})$$

PS2 – Naive Bayes Classifier

Learning And Mining from DatA

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

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

Solution.

$$\Pr\{y=1|\mathbf{x}=(1,1,0,1)\} = \frac{\hat{P}(y)}{\hat{P}(\mathbf{x})}\prod_{i=1}^{4}\hat{P}(x_i|y)$$

$$= \frac{\hat{\Pr}\{y=1\}}{\hat{\Pr}\{\mathbf{x}=(1,1,0,1)\}}\hat{\Pr}\{x_1=1|y=1\}\hat{\Pr}\{x_2=1|y=1\}\hat{\Pr}\{x_3=0|y=1\}\hat{\Pr}\{x_4=1|y=1\}$$

$$= \frac{4/7}{P(\mathbf{x})}\times\frac{3}{5}\times\frac{2}{5}\times\frac{1}{5}\times\frac{3}{5}$$

$$= \frac{72}{4375}/P(\mathbf{x})$$

$$= \frac{0.0164}{P(\mathbf{x})}$$

PS2 – Naive Bayes Classifier

Learning And Mining from DatA

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

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

$$\Pr\{y=0|\mathbf{x}=(1,1,0,1)\} = \frac{\hat{P}(y)}{\hat{P}(\mathbf{x})}\prod_{i=1}^{4}\hat{P}(x_i|y)$$

$$= \frac{\hat{\Pr}\{y=0\}}{\hat{\Pr}\{\mathbf{x}=(1,1,0,1)\}}\hat{\Pr}\{x_1=1|y=0\}\hat{\Pr}\{x_2=1|y=0\}\hat{\Pr}\{x_3=0|y=0\}\hat{\Pr}\{x_4=1|y=0\}$$

$$= \frac{3/7}{P(\mathbf{x})}\times\frac{2}{4}\times\frac{2}{4}\times\frac{2}{4}\times\frac{2}{4}$$

$$= \frac{3}{112}/P(\mathbf{x})$$

$$= \frac{0.268}{P(\mathbf{x})}$$

PS3 - Ensemble Methods in Practice

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；

PS3 - Ensemble Methods in Practice

(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；

PS3 - Ensemble Methods in Practice

(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。

- **编程题的注意事项**
  - 在进行ensemble时，要灵活调用已有的工具，如可调用sklearn中的基本算法，利用sklearn中的参数设置，实现集成方法
  - 超参数对集成的影响很大，可以通过k-折交叉验证来进行参数选择，减轻过拟合的影响，提高训练效果
  - 实验报告中需要明确说明实验的目的，较为清楚的简述实验方法，直观的展示实验结果，并根据实验结果进行分析，发现实验的不足提出不足的产生原因

Q & A

Thanks!