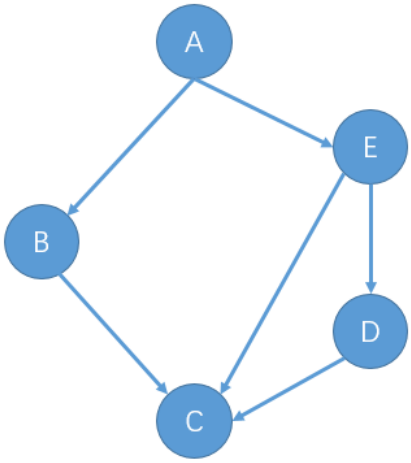
机器学习导论
作业五

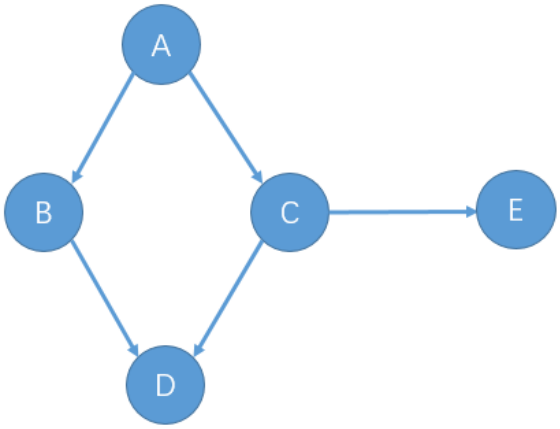
2018 年 6 月 30 日

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），只考虑X和Y是单变量即可。

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

(2) $A \perp\!\!\!\perp E|C, B \perp\!\!\!\perp E|C, D \perp\!\!\!\perp E|C, A \perp\!\!\!\perp D|\{B, C\}$.

(3) $B \perp\!\!\!\perp D|\{A\}, B \perp\!\!\!\perp D|\{E\}, B \perp\!\!\!\perp D|\{A, E\}$;
$B \perp\!\!\!\perp E|\{A\}, B \perp\!\!\!\perp E|\{A, D\}$;
$A \perp\!\!\!\perp C|\{B, E\}, A \perp\!\!\!\perp C|\{B, E, D\}$;
$A \perp\!\!\!\perp D|\{E\}, A \perp\!\!\!\perp D|\{E, B\}$.

(4) 不能保证，反例如图1所示.

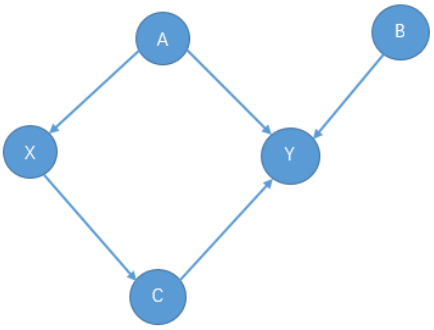


图 1: 反例

根据修改后的道德化方法，在考察变量X和Y的独立性时，连接V型结构父结点前应当去除节点C(因为C是X的孩子)，从而得到的无向图中A可以分离X和Y，得到$X \perp\!\!\!\perp Y|\{A\}$，但是根据定理1可知，$\{A\}$没有将道路$XCY$阻塞因此该条件独立集合不正确。

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. 此处用于写解答*(中英文均可)*

(1) 首先估计类条件概率，显然有

$$\begin{aligned} P(y=1) &= \frac{3}{5} \\ P(y=0) &= \frac{2}{5} \end{aligned} \tag{2.1}$$

然后估计每个属性的条件概率

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

所以可估计后验概率

$$\begin{aligned}\Pr\{y=1|\mathbf{x}=(1,1,0,1)\} &= P(y=1)P(x_1=1|y=1)P(x_2=1|y=1)P(x_3=0|y=1)P(x_4=1|y=1)\\ &= 0\\ \Pr\{y=0|\mathbf{x}=(1,1,0,1)\} &= P(y=0)P(x_1=1|y=0)P(x_2=1|y=0)P(x_3=0|y=0)P(x_4=1|y=0)\\ &= 1/40\\ &= 0.025\end{aligned} \tag{2.3}$$

(2) 首先估计类条件概率，有

$$\begin{aligned}P(y=1) &= \frac{3+1}{5+2} = \frac{4}{7}\\ P(y=0) &= \frac{2+1}{5+2} = \frac{3}{7}\end{aligned} \tag{2.4}$$

然后估计每个属性的条件概率

$$\begin{aligned}P(x_1=1|y=1) &= \frac{2+1}{3+2} = \frac{3}{5}\\ P(x_2=1|y=1) &= \frac{1+1}{3+2} = \frac{2}{5}\\ P(x_3=0|y=1) &= \frac{0+1}{3+2} = \frac{1}{5}\\ P(x_4=1|y=1) &= \frac{2+1}{3+2} = \frac{3}{5}\\ P(x_1=1|y=0) &= \frac{1+1}{2+2} = \frac{1}{2}\\ P(x_2=1|y=0) &= \frac{1+1}{2+2} = \frac{1}{2}\\ P(x_3=0|y=0) &= \frac{1+1}{2+2} = \frac{1}{2}\\ P(x_4=1|y=0) &= \frac{1+1}{2+2} = \frac{1}{2}\end{aligned} \tag{2.5}$$

所以可估计后验概率

$$\begin{aligned}\Pr\{y=1|\mathbf{x}=(1,1,0,1)\} &= P(y=1)P(x_1=1|y=1)P(x_2=1|y=1)P(x_3=0|y=1)P(x_4=1|y=1)\\ &\approx 0.01646\\ \Pr\{y=0|\mathbf{x}=(1,1,0,1)\} &= P(y=0)P(x_1=1|y=0)P(x_2=1|y=0)P(x_3=0|y=0)P(x_4=1|y=0)\\ &\approx 0.02679\end{aligned} \tag{2.6}$$

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。

4 实验报告.

4.1 实验目的

对集成学习方法的两种经典方法(Boosting和Bagging) 进行实践，探究基学习期数量对模型性能的影响，对比两种不同的算法，体会其中的集成学习思想。

4.2 实验过程

4.2.1 AdaBoost

1. [**算法框架**] 依据教材8.2节中图8.3所示的AdaBoost算法伪代码。
2. [**基学习器**] 选用决策树，调用sklearn中的DecisionTreeClassifier实现，最大深度设为14，其他值为default。
3. [**参数分析**] 基学习器数量即决策树的数量由1逐个递增到60(因为实验中发现60已足够大)，运用5折交叉验证测得不同基学习器数量对应的平均AUC值T_{best}，然后把基学习器数量设为T_{best}运用全部训练数据训练，在测试集上测试作为最终结果。

4.2.2 RandomForest

1. [**算法框架**] 在教材8.3.1节中图8.5所示的Bagging算法伪代码的基础上进一步在决策树的训练过程中引入了随机属性选择，随机性参数k取值按照教材的推荐值 .
2. [**基学习器**] 选用决策树，调用sklearn中的DecisionTreeClassifier实现，最大深度设为14，其他值为default。
3. [**自助采样**] 调用sklearn中util模块中的resample函数，每次采样的数目和本轮训练样本数目一致。
4. [**参数分析**] 基学习器数量即决策树的数量由1逐个递增到60(因为实验中发现60已足够大)，运用5折交叉验证测得不同基学习器数量对应的平均AUC值T_{best}，然后把基学习器数量设为T_{best}运用全部训练数据训练，在测试集上测试作为最终结果。

4.3 实验结果

4.3.1 AUC性能

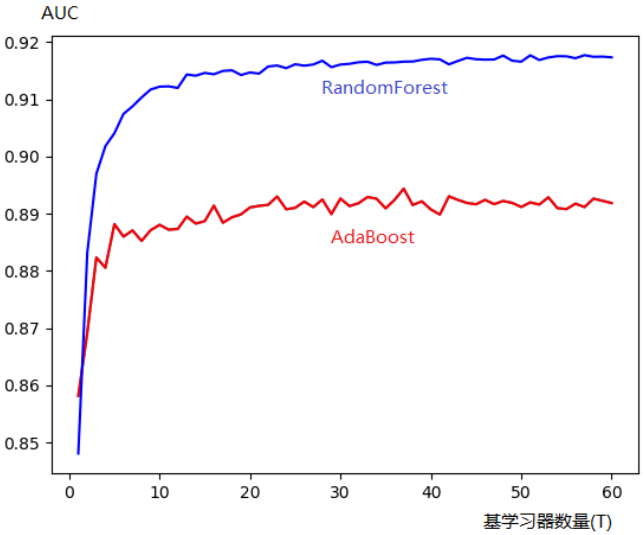


图 2: AUC折线图

1. 如图2所示，可见当基学习器数量大于40时交叉验证的效果基本稳定。
2. 容易看出，本实验中RandomForest的性能优于AdaBoost。
3. 两个模型的最优学习器数量及其测试结果表2所示。

表 2: 测试结果

	T_{best}	AUC_{test}
AdaBoost	36	0.894
RandomForest	56	0.916

4.3.2 实验分析

1. 实验中发现决策树的最大深度会对实验结果造成较大影响。如果深度过深，则很容易过拟合，如果深度限制得过低，则对于Adaboost的顺序较为靠后的学习器无法完成对训练样本的有效决策划分，即训练误差大于0.5，从而导致早听。因此本实验设置了一个较为适中的决策树深度(即14).
2. 训练过程中AdaBoost的训练误差可以迅速收敛到0(比直接调用sklearn中的AdaBoost还快)，但是RandomForest的训练误差无法收敛到0.这与两个模型的理论上的表现是一致的。
3. 实验的结果具有一定的随机性，但是多次实验的结果相差不大所以直接取了其中的一次实验结果。