机器学习导论
习题六

学号, 作者姓名, 邮箱

2020 年 6 月 18 日

1 [25pts] Bayesian Network

贝叶斯网(Bayesian Network)是一种经典的概率图模型，请学习书本7.5节内容回答下面的问题：

(1) **[5pts]** 请画出下面的联合概率分布的分解式对应的贝叶斯网结构：

$$\Pr(A,B,C,D,E,F,G) = \Pr(A)\Pr(B)\Pr(C)\Pr(D|A)\Pr(E|A)\Pr(F|B,D)\Pr(G|D,E)$$

(2) **[5pts]** 请写出图1中贝叶斯网结构的联合概率分布的分解表达式。

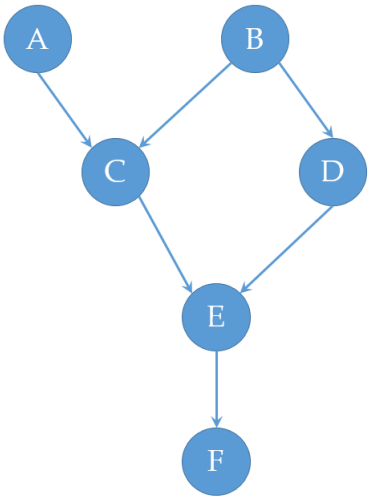
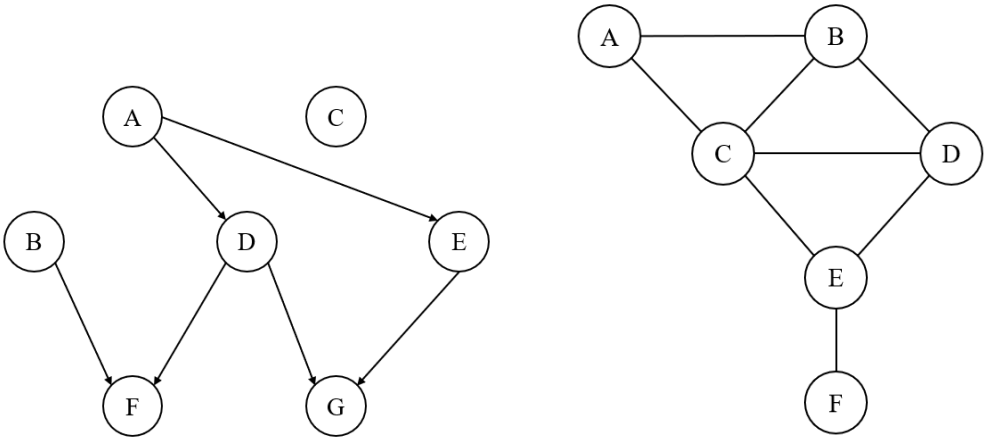


图 1: 题目1-(2)有向图

(3) **[15pts]** 基于第(2)问中的图1, 请判断表格1中的论断是否正确。首先需要作出对应的道德图，并将下面的表格填完整。

Solution.

(1) 对应贝叶斯网络结构如图2(a)所示



(a) 题目1-(1)贝叶斯网络结构 (b) 题目1-(3)道德图

图 2: Solution

(2) $\Pr(A,B,C,D,E,F) = \Pr(A)\Pr(B)\Pr(C|A,B)\Pr(D|B)\Pr(E|C,D)\Pr(F|E)$

(3) 对应的道德图如图2(b)所示

表 1: 判断表格中的论断是否正确

序号	关系	True/False	序号	关系	True/False
1	$A \perp\!\!\!\perp B$	True	7	$F \perp B \mid C$	False
2	$A \perp B \mid C$	False	8	$F \perp B \mid C, D$	True
3	$C \perp\!\!\!\perp D$	False	9	$F \perp B \mid E$	True
4	$C \perp D \mid E$	False	10	$A \perp\!\!\!\perp F$	False
5	$C \perp D \mid B, F$	False	11	$A \perp F \mid C$	False
6	$F \perp\!\!\!\perp B$	False	12	$A \perp F \mid D$	False

助教注：有部分同学没有注意第三小问中要求画出道德图；另外还有很多同学选择画图拍照答题，这些做法均被扣除了一定的分数。

2 [35+10pts] Theoretical Analysis of k-means Algorithm

给定样本集$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, k-means聚类算法希望获得簇划分$\mathcal{C} = \{C_1, C_2, \cdots, C_k\}$使得最小化欧式距离

$$J(\gamma, \mu_1, \ldots, \mu_k) = \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij} \left\|\mathbf{x}_i - \mu_j\right\|^2 \tag{1}$$

其中$\mu_1, \ldots, \mu_k$为k个簇的中心(means), $\gamma \in \mathbb{R}^{n\times k}$为指示矩阵(indicator matrix)定义如下：若$\mathbf{x}_i$属于第j个簇, 则$\gamma_{ij} = 1$, 否则为0，则最经典的k-means聚类算法流程如算法1中所示

(1) **[5pts]** 试证明, 在算法1中, Step 1和Step 2都会使目标函数J的值降低.

Algorithm 1 $k-$means Algorithm

1: Initialize $\mu_1, \ldots, \mu_k$;

2: **repeat**

3: **Step 1:** Decide the class memberships of $\{\mathbf{x}_i\}_{i=1}^n$ by assigning each of them to its nearest cluster center.

$$\gamma_{ij} = \begin{cases} 1, & \|\mathbf{x}_i - \mu_j\|^2 \leq \|\mathbf{x}_i - \mu_{j'}\|^2, \forall j' \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

4: **Step 2:** For each $j \in \{1, \cdots, k\}$, recompute μ_j using the updated γ to be the center of mass of all points in C_j :

$$\mu_j = \frac{\sum_{i=1}^n \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^n \gamma_{ij}} \tag{3}$$

5: **until** the objective function J no longer changes;

(2) **[5pts]** 试证明, 算法1会在有限步内停止。

(3) **[10pts]** 试证明, 目标函数J的最小值是关于k的非增函数, 其中k是聚类簇的数目。

(4) **[15pts]** 记$\hat{\mathbf{x}}$为n个样本的中心点, 定义如下变量,

$$\begin{aligned} T(X) &= \textstyle\sum_{i=1}^n \|\mathbf{x}_i - \hat{\mathbf{x}}\|^2 / n \\ W_j(X) &= \textstyle\sum_{i=1}^n \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2 / \sum_{i=1}^n \gamma_{ij} \\ B(X) &= \textstyle\sum_{j=1}^k \frac{\sum_{i=1}^n \gamma_{ij}}{n} \|\mu_j - \hat{\mathbf{x}}\|^2 \end{aligned}$$

试探究以上三个变量之间有什么样的等式关系？基于此请证明, k-means聚类算法可以认为是在最小化$W_j(X)$的加权平均, 同时最大化$B(X)$.

(5) **[Bonus 10pts]**在公式1中, 我们使用ℓ_2-范数来度量距离(即欧式距离), 下面我们考虑使用ℓ_1-范数来度量距离

$$J'(\gamma, \mu_1, \ldots, \mu_k) = \sum_{i=1}^n \sum_{j=1}^k \gamma_{ij} \|\mathbf{x}_i - \mu_j\|_1 \tag{4}$$

- 请仿效算法1,给出新的算法(命名为$k-$means-ℓ_1算法)以优化公式4中的目标函数J'.
- 当样本集中存在少量异常点(outliers)时, 上述的$k-$means-ℓ_2和$k-$means-ℓ_1算法, 我们应该采用哪种算法？即哪个算法具有更好的鲁棒性？请说明理由。

Solution.

(1) Step1的过程是指派每个点隶属于其最近的簇中心，在这个过程中不影响其他样本的从属关系，由于$\|\mathbf{x}_i - \mu_j\|^2 \leq \|\mathbf{x}_i - \mu_{j'}\|^2, \forall j'$，对应项J中的值将减小。

Step2的过程是根据当前样本的从属关系，计算每个簇的簇心。由于J是凸函数，通过对μ_j求导可知，能够最小化J的簇心μ_j即为当前簇样本的均值

$$\frac{\partial J}{\partial \mu_j} = 2\sum_{i=1}^{n} \gamma_{ij}(\mathbf{x}_i - \mu_j)$$

令其为0，得到

$$\mu_j^* = \frac{\sum_{i=1}^{n} \gamma_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} \gamma_{ij}}$$

(2) 由上一问可知，k-means算法随着迭代过程将不断降低目标函数J的值，这意味着算法不会重复得到同样的γ矩阵，由于γ矩阵是具有有限大小(nk)且每个元素γ_{ij}的取值范围均为0或1，这意味着它具有有限多的状态，因此在有限步内算法会停止。

(3) 考虑算法在给定簇数目为k的情况下已经收敛，此时得到的目标函数最小值为$J(k)$，此时在保留原有簇心的情况下引入一个新的簇心μ_{k+1}，由于尚未指派旧样本到新的簇心的所属关系，当前$\gamma_{i,k+1}$均为0。这种情况可以被视为聚类数为k+1的一种特殊初始化，且当前的目标函数$J'(k+1) = J(k)$。

由于之前已经讨论证明算法将会随着运行过程不断减小目标函数的值，因此在有限步迭代后，算法将对$k+1$个簇心收敛，且此时的$J(k+1) \leq J'(k+1) = J(k)$.

最后，考虑一种特殊情况，即簇心数目为n且每个簇心对应着一个样本点的情况，即$\mu_i = \mathbf{x}_i$，显然，此时的目标函数值为0，设计这一问是为了告诉大家在做$k-$means算法的时候，目标函数J的绝对值大小并不具有实际意义，但能够在一些算法中指导确定合适的聚类数目k。

(4) 首先观察到$\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}W_j(X) = J$，下面从这个关系出发构造$T(X)$和$B(X)$的等式关系。

$$\begin{aligned}
\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}W_j(X) + nB(X) &= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}\left\|\mathbf{x}_i - \mu_j\right\|^2 + \gamma_{ij}\left\|\mu_j - \hat{\mathbf{x}}\right\|^2 \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}\left(\left\|\mathbf{x}_i - \mu_j\right\|^2 + \left\|\mu_j - \hat{\mathbf{x}}\right\|^2\right) \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}\left(\mathbf{x}_i^2 + \hat{\mathbf{x}}^2 - 2\mathbf{x}_i\hat{\mathbf{x}} + 2x_i\hat{\mathbf{x}} + 2\mu_j^2 - 2\mathbf{x}_i\mu_j - 2\hat{\mathbf{x}}\mu_j\right) \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}\left(\left(\mathbf{x}_i^2 + \hat{\mathbf{x}}^2 - 2\mathbf{x}_i\hat{\mathbf{x}}\right) + \left(2x_i\hat{\mathbf{x}} + 2\mu_j^2 - 2\mathbf{x}_i\mu_j - 2\hat{\mathbf{x}}\mu_j\right)\right) \\
&= \sum_{j=1}^{k}\left(\sum_{i=1}^{n}\gamma_{ij}\left(\left\|\mathbf{x}_i - \hat{\mathbf{x}}\right\|^2\right)\right) + K
\end{aligned}$$

其中K是等式右半部分其余项的和，考虑将第一项对j展开

$$\begin{aligned}
\sum_{j=1}^{k}\left(\sum_{i=1}^{n}\gamma_{ij}\left(\left\|\mathbf{x}_i - \hat{\mathbf{x}}\right\|^2\right)\right) + K &= \sum_{i=1}^{n}\left(\left\|\mathbf{x}_i - \hat{\mathbf{x}}\right\|^2\right) + K \\
&= nT(X) + K
\end{aligned}$$

对于给定的数据集，$nT(X)$是一个常数，而k-means的优化过程会不断将J，即$W_j(X)$的加权平均进行最小化，此时也是在近似地最大化$B(X)$

注：本题题干部分遗漏了**近似地**最大化$B(X)$，在作业中指出题目问题的将得到额外2pts.

(5) 本题是开放性的，言之有理即可，下面提供一个参考解答，最终赋分并不需要与解答完全相同。

使用ℓ_1-范数来度量距离，在算法Step1中需要将样本指派到ℓ_1-距离最近的样本点，考虑到对此时的μ_j求导，对应项的值将为-1或1，能够使导数为0的簇心应该使簇心两侧-1和1的个数相等，因此在Step2中采用同簇样本的中位数作为簇心。

在面对异常点时，由于采用中位数来设计簇心，它并不关注异常点到簇心的距离；而原始的$k-$means-ℓ_2算法考虑了离群点到簇心的距离信息，通常这个距离属于干扰因素。综上所述，使用$k-$means-ℓ_1算法将对异常点更加鲁棒。

3 [40pts] Coding: Ensemble Methods

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，数据集文件夹为adult_dataset。

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，可调用sklearn算法包对AUC指标进行计算。

(1) 本次实验要求使用Python3编写，要求代码分布于两个文件中，BoostMain.py, RandomForestMain.py ，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，也可以手动实现，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，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) **[5pts]** 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.