第一次作业题

浙江大学

赵洲

题目

1. 数据集包含 100 个样本，其中正、反例各一半，假定学习算法所产生的模型是将新样本预测为训练样本数较多的类别（训练样本数相同时进行随机猜测），试给出用 10 折交叉验证发和留一法分别对错误率进行评估的结果。

题目

1. 数据集包含 100 个样本，其中正、反例各一半，假定学习算法所产生的模型是将新样本预测为训练样本数较多的类别（训练样本数相同时进行随机猜测），试给出用 10 折交叉验证发和留一法分别对错误率进行评估的结果。

解答：10折交叉验证：交叉验证中每个子集数据分布要尽可能保持一致，那么本题中10次训练中每次正反例各占 45，模型训练Q结果随机猜测，错误率期望为 50%。

留一法：若留出样本为正例，训练集中则有50个反例和49个正例，模型预测为反例；反之留出样本为反例，模型预测为正例，错误率为100%。

题目

3. 假设某机器学习模型的原始类别和预测类别如下表所示，求它的混淆矩阵、准确率、精确率、召回率、F1 score。

样本序列	1	2	3	4	5	6	7	8	9	10
原始类别	1	1	1	-1	-1	-1	1	1	-1	1
预测类别	1	1	-1	-1	-1	1	-1	1	-1	1

题目

1. 混淆矩阵的定义

- 混淆矩阵如下：

	预测：1	预测：-1
真实：1	*TP*	*FN*
真实：-1	*FP*	*TN*

	预测：1	预测：-1
真实：1	4	2
真实：-1	1	3

- 其中：
  - *TP* (True Positive): 预测为“1”，真实也是“1”
  - *FP* (False Positive): 预测为“1”，真实为“-1”
  - *FN* (False Negative): 预测为“-1”，真实为“1”
  - *TN* (True Negative): 预测为“-1”，真实也是“-1”

混淆矩阵

$$\begin{pmatrix} 4 & 2 \\ 1 & 3 \end{pmatrix}$$

计算题

- 准确率 (Accuracy)
- 准确率的公式为:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}. \qquad (1)$$

- 代入数值:

$$\text{Accuracy} = \frac{4 + 3}{4 + 1 + 2 + 3} = \frac{7}{10} = 0.7. \qquad (2)$$

计算题

精确率 (Precision)

精确率的公式为：

$$\text{Precision} = \frac{TP}{TP + FP}. \qquad (3)$$

代入数值：

$$\text{Precision} = \frac{4}{4 + 1} = \frac{4}{5} = 0.8. \qquad (4)$$

召回率 (Recall)

召回率的公式为：

$$\text{Recall} = \frac{TP}{TP + FN}. \qquad (5)$$

代入数值：

$$\text{Recall} = \frac{4}{4 + 2} = \frac{2}{3}. \qquad (6)$$

计算题

- 4. 性能度量方法 (10分)

F1 Score

$F1$ Score 是精确率和召回率的调和平均，公式为：

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \qquad (7)$$

代入数值：

$$F1 = \frac{2 \times 0.8 \times \frac{2}{3}}{0.8 + \frac{2}{3}} \approx \frac{2 \times 0.5336}{1.467} = 0.727. \qquad (8)$$

题目

- 对以下数据集，构造 ID3 决策树，判断是否买房：

用户 ID	年龄	性别	收入	是否买房
1	27	男	15W	否
2	47	女	30W	是
3	32	男	12W	否
4	24	男	45W	是
5	45	男	30W	否
6	56	男	32W	是
7	31	男	15W	否
8	23	女	30W	是

- 注：年龄分为 20-30，30-40，40+三个阶段，收入分为 10-20，20-40，40+三个级别

题目

- 1. 数据集的熵 $H(D)$
- 数据集中，是否买房的分布为：
  - 是：4 条
  - 否：4 条
- 数据集的熵计算公式为：

$$H(D) = -p_{是}\log_2 p_{是} - p_{否}\log_2 p_{否}$$

- 其中 $p_{是} = \frac{4}{8} = 0.5$，$p_{否} = \frac{4}{8} = 0.5$。代入公式：

$$H(D) = -0.5\log_2 0.5 - 0.5\log_2 0.5 = 1$$

题目

- **2. 特征的条件熵与信息增益**
- ***2.1 年龄（分为*** 20 − 30,30 − 40,40 + ***）***
- 20 − 30：3 条记录，2 是，1 否
- 30 − 40：2 条记录，0 是，2否
- 40 +：3 条记录，2 是，1 否
- 计算每个分组的熵：

$$H_{20-30} = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} \approx 0.918$$

$$H_{30-40} = -1\log_2\frac{2}{2} = 0$$

$$H_{40+} = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = \approx 0.918$$

- 条件熵 $H(D|\text{年龄})$ 为：

$$H(D|\text{年龄}) = \frac{3}{8}H_{20-30} + \frac{2}{8}H_{30-40} + \frac{3}{8}H_{40+} = \frac{3}{8}\cdot 0.918 + \frac{2}{8}\cdot 0 + \frac{3}{8}\cdot 0.918 \approx 0.689$$

- 信息增益 IG(年龄) 为：

题目

- ***2.2 性别***

- 男：6 条记录，2 是，4 否

- 女：2 条记录，2 是，0 否

- 计算每个分组的熵：

$$H_{男} = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} \approx 0.918$$

$$H_{女} = -1\log_2 1 = 0$$

- 条件熵 $H(D|性别)$ 为：

$$H(D|性别) = \frac{6}{8}H_{男} + \frac{2}{8}H_{女} = \frac{6}{8}\cdot 0.918 + \frac{2}{8}\cdot 0 = 0.689$$

- 信息增益 $IG(性别)$ 为：

$$IG(性别) = H(D) - H(D|性别) = 1 - 0.689 = 0.311$$

题目

- ***2.3** 收入（分为* 10 – 20,20 – 40,40 +）
  - 10 – 20：3 条记录，0 是，3否
  - 20 – 40：4 条记录，3 是，1 否
  - 40 +：1 条记录，1 是，0 否
- 计算每个分组的熵：

$$H_{10-20} = -1\log_2 1 = 0$$

$$H_{20-40} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} \approx 0.811$$

$$H_{40+} = -1\log_2 1 = 0$$

- 条件熵 $H(D|\text{收入})$ 为：
- $H(D|\text{收入}) = \frac{2}{8}H_{10-20} + \frac{4}{8}H_{20-40} + \frac{2}{8}H_{40+} = \frac{2}{8}\cdot 0 + \frac{4}{8}\cdot 0.811 + \frac{2}{8}\cdot 0 \approx 0.406$
- 信息增益 $IG(\text{收入})$ 为：
- $IG(\text{收入}) = H(D) - H(D|\text{收入}) = 1 - 0.406 = 0.594$

题目

- **3. 比较信息增益**
- IG(年龄) = 0.311,　IG(性别) = 0.311,　IG(收入) = 0.594
- 由于 IG(收入) 最大，选择收入作为根节点。

题目

- 4. 根据收入属性初步划分数据集：
- • 收入为 10–20：
- • 全部为“不买房”（No），熵为 0，作为叶节点。
- • 收入为 40+：
- • 全部为“买房”（Yes），熵为 0，作为叶节点。
- • 收入为 20–40：
- • 包含混合的买房和不买房样本，需进一步划分。

题目

- **5. 在收入为 20-40 的子集中，分别计算年龄和性别的信息增益：**
- 根据年龄分组：

- 20 – 30：1 条记录，1 是，0 否
- 30 – 40：0 条记录，0 是，0否
- 40 +：3 条记录，2 是，1 否

- 计算每个分组的熵：

$$H_{20-30} = -1\log_2\frac{1}{1} = 0$$

$$H_{40+} = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = \approx 0.918$$

- 条件熵 $H(D|\text{年龄})$ 为：$H(D|\text{年龄}) = \frac{1}{4}H_{20-30} + \frac{3}{4}H_{40+} = \frac{1}{4}\cdot 0 + \frac{3}{4}\cdot 0.918 \approx 0.689$
- 信息增益 $IG(\text{年龄})$ 为：$IG(\text{年龄}) = H(D) - H(D|\text{年龄}) = 1 - 0.689 = 0.311$

题目

- **5. 在收入为 20-40 的子集中，分别计算年龄和性别的信息增益：**
- 按性别分组：

- 男：2 条记录，1 是，1 否
- 女：2 条记录，2 是，0 否

计算每个分组的熵：

$$H_{男} = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$H_{女} = -1\log_2 1 = 0$$

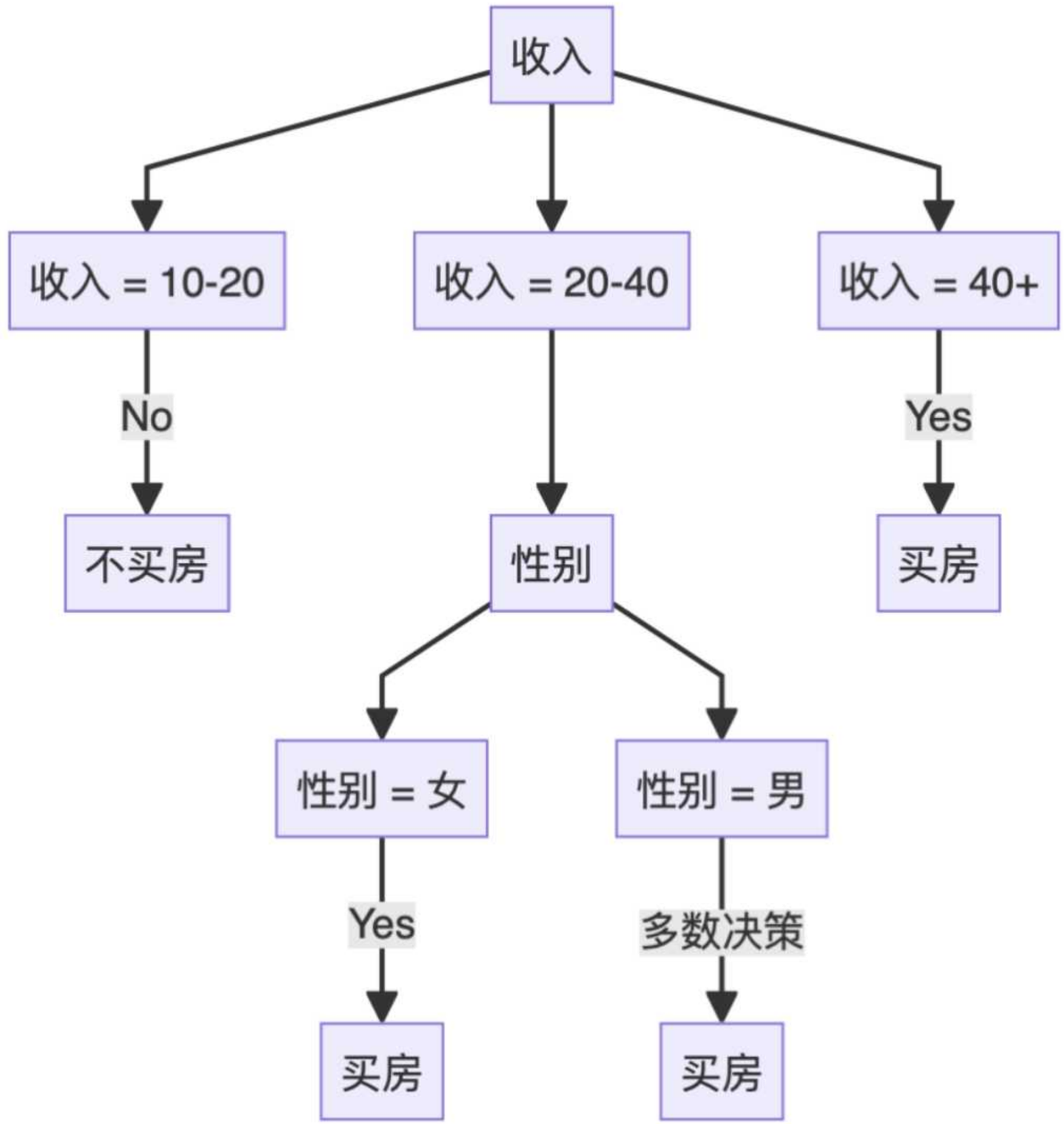
- 条件熵 $H(D|性别)$ 为：
- $H(D|性别) = \frac{2}{4}H_{男} + \frac{2}{4}H_{女} = \frac{2}{4}\cdot 1 + \frac{2}{4}\cdot 0 = 0.5$
- 信息增益 $IG(性别)$ 为：$IG(性别) = H(D) - H(D|性别) = 1 - 0.5 = 0.5$

题目

- *比较信息增益*
- IG(年龄) = 0.311,　IG(性别) = 0.5
- 由于 IG(性别) 最大，选择**性别**作为划分属性。
- 在收入为 20-40 的子集中，根据性别划分：
- • **性别为女性：**
- • 全部为“买房”（Yes），熵为 0，作为叶节点。
- • **性别为男性：**
- • 1 个“买房”，1 个“不买房”，分布均衡，熵为 1，样本数较少，无法进一步有效划分，根据收入为 20–40 的子集中的多数决策，预测为“买房”（Yes）。

题目

- 最终决策树结构:



题目

- **问题：**

  假设我们有一个包含 $D = 3$ 篇文档的语料库，词汇表大小为 $W = 3$。每篇文档 d 都由词频 T_{dw} 表示，其中 T_{dw} 是词 w 在文档 d 中出现的次数。我们假设这些文档是从 $K = 2$ 个多项分布（主题）的混合模型中生成的，每个主题具有混合系数 π_k 和主题特定的词分布参数 $\boldsymbol{\mu}_k$。

- 具体数据如下：

- **文档1**：词1出现了4次（$T_{1,1} = 4$），词2出现了1次（$T_{1,2} = 1$），词3出现了1次（$T_{1,3} = 1$）。
- **文档2**：词1出现了1次（$T_{2,1} = 1$），词2出现了4次（$T_{2,2} = 4$），词3出现了1次（$T_{2,3} = 1$）。
- **文档3**：词1出现了1次（$T_{3,1} = 1$），词2出现了1次（$T_{3,2} = 1$），词3出现了4次（$T_{3,3} = 4$）。

题目

- 初始参数设定为：混合系数：$\pi_1^{(0)} = 0.5$，$\pi_2^{(0)} = 0.5$
- 主题1的词分布：$\boldsymbol{\mu}_1^{(0)} = \left[\mu_{11}^{(0)} = 0.6, \mu_{12}^{(0)} = 0.3, \mu_{13}^{(0)} = 0.1\right]$
- 主题2的词分布：$\boldsymbol{\mu}_2^{(0)} = \left[\mu_{21}^{(0)} = 0.2, \mu_{22}^{(0)} = 0.5, \mu_{23}^{(0)} = 0.3\right]$
- **要求：**
- **1.推导E步和M步更新公式：**
- (a) 推导在EM算法的E步中，计算每个文档属于每个主题的后验概率 $\gamma_{dk} = P(c_d = k|d)$ 的计算公式。
- (b) 推导在M步中，用于更新混合系数 π_k 和词分布参数 μ_{kw} 的更新公式。
- **2.执行一次EM迭代：**
- 使用上述初始参数和数据，执行一次EM算法的迭代，包括E步和M步，计算更新后的参数 $\pi_k^{(1)}$ 和 $\mu_{kw}^{(1)}$。
- **3.计算对数似然函数：**
- (a) 推导出数据的对数似然函数 $L = \sum_{d=1}^{D} \ln P(d)$，其中 $P(d) = \sum_{k=1}^{K} \pi_k P(d|c_d = k)$。
- (b) 计算初始参数下的数据对数似然值 $L^{(0)}$，以及更新参数后的对数似然值 $L^{(1)}$，并比较两者的变化

题目

- **1. 推导E步和M步更新公式**
- ***(a) E步中后验概率 γ_{dk} 的计算公式推导***
- 在EM算法的E步，我们需要计算每个文档属于每个主题的后验概率 $\gamma_{dk} = P(c_d = k|d)$。
- 根据贝叶斯公式:

$$P(c_d = k|d) = \frac{P(d|c_d = k)P(c_d = k)}{P(d)} \qquad (4)$$

- 其中:
  - $P(c_d = k) = \pi_k$，即主题 k 的混合系数。
  - $P(d|c_d = k)$ 是在主题 k 下生成文档 d 的概率。

题目

- 由于文档 d 中的词汇出现次数满足多项分布：

$$P(d|c_d = k) = \frac{n_d!}{\prod_{w=1}^{W} T_{dw}!} \prod_{w=1}^{W} \mu_{kw}^{T_{dw}} \qquad (5)$$

- 其中 $n_d = \sum_{w=1}^{W} T_{dw}$。

- 因此，后验概率为：

$$\gamma_{dk} = \frac{\pi_k \prod_{w=1}^{W} \mu_{kw}^{T_{dw}}}{\sum_{j=1}^{K} \pi_j \prod_{w=1}^{W} \mu_{jw}^{T_{dw}}} \qquad (6)$$

注意，在计算 γ_{dk} 时，归一化常数 $\frac{n_d!}{\prod_{w=1}^{W} T_{dw}!}$ 可以约去，因为它对于所有主题 k 都是相同的。

题目

- ***(b) M步中参数更新公式推导***
- 在M步，我们需要最大化对数似然的期望，即对参数 π_k 和 μ_{kw} 求解：
- $$\max_{\pi_k,\mu_{kw}} \sum_{d=1}^{D}\sum_{k=1}^{K}\gamma_{dk}\left(\ln\pi_k+\sum_{w=1}^{W}T_{dw}\ln\mu_{kw}\right) \qquad (7)$$
- 由于参数有约束条件：
- $\sum_{k=1}^{K}\pi_k=1$
- 对于每个主题 k，$\sum_{w=1}^{W}\mu_{kw}=1$
- 我们使用拉格朗日乘数法。

题目

- **更新混合系数 π_k:**

- 构建拉格朗日函数:

$$L = \sum_{d=1}^{D}\sum_{k=1}^{K} \gamma_{dk} \ln \pi_k - \lambda_0 \left(\sum_{k=1}^{K} \pi_k - 1 \right) \qquad (8)$$

- 对 π_k 求偏导:

- $$\frac{\partial L}{\partial \pi_k} = \frac{\sum_{d=1}^{D} \gamma_{dk}}{\pi_k} - \lambda_0 = 0 \qquad (9)$$

- 整理得:

$$\pi_k = \frac{\sum_{d=1}^{D} \gamma_{dk}}{\lambda_0} \qquad (10)$$

题目

- 由于 $\sum_{k=1}^{K} \pi_k = 1$，代入得：

$$\lambda_0 = \sum_{d=1}^{D} \sum_{k=1}^{K} \gamma_{dk} \qquad (11)$$

- 因此：

$$\pi_k^{(1)} = \frac{\sum_{d=1}^{D} \gamma_{dk}}{\sum_{d=1}^{D} \sum_{k=1}^{K} \gamma_{dk}} \qquad (12)$$

由于 γ_{dk} 对所有 d 和 k 的总和为 D，所以分母为 D，即：

$$\pi_k^{(1)} = \frac{\sum_{d=1}^{D} \gamma_{dk}}{D} \qquad (13)$$

题目

- **更新词分布参数 μ_{kw}：构建拉格朗日函数：**

$$L = \sum_{d=1}^{D}\sum_{k=1}^{K}\gamma_{dk}\sum_{w=1}^{W}T_{dw}\ln\mu_{kw} - \sum_{k=1}^{K}\lambda_k\left(\sum_{w=1}^{W}\mu_{kw} - 1\right) \qquad (14)$$

- **对 μ_{kw} 求偏导：**

$$\frac{\partial L}{\partial \mu_{kw}} = \frac{\sum_{d=1}^{D}\gamma_{dk}T_{dw}}{\mu_{kw}} - \lambda_k = 0 \qquad (15)$$

- **整理得：**

$$\mu_{kw} = \frac{\sum_{d=1}^{D}\gamma_{dk}T_{dw}}{\lambda_k} \qquad (16)$$

- **由于 $\sum_{w=1}^{W}\mu_{kw} = 1$，代入得：**

$$\lambda_k = \sum_{w=1}^{W}\sum_{d=1}^{D}\gamma_{dk}T_{dw} \qquad (17)$$

- **因此：**

$$\mu_{kw}^{(1)} = \frac{\sum_{d=1}^{D}\gamma_{dk}T_{dw}}{\sum_{d=1}^{D}\sum_{w=1}^{W}\gamma_{dk}T_{dw}} \qquad (18)$$

- **注意，$\sum_{w=1}^{W}T_{dw}$ 是文档 d 的总词数 n_d，而 $\sum_{d=1}^{D}\sum_{w=1}^{W}\gamma_{dk}T_{dw}$ 是主题 k 在所有文档中的加权词频总和。**

题目

- 2. 执行一次EM迭代

- 我们将使用初始参数和数据执行一次EM迭代，包括E步和M步。

- E步：计算 γ_{dk}

- 对于每个文档 d 和主题 k，计算:

- $$\gamma_{dk} = \frac{\pi_k^{(0)} \prod_{w=1}^{W} \left(\mu_{kw}^{(0)}\right)^{T_{dw}}}{\sum_{j=1}^{K} \pi_j^{(0)} \prod_{w=1}^{W} \left(\mu_{jw}^{(0)}\right)^{T_{dw}}} \qquad (19)$$

题目

- 2. 执行一次EM迭代
- 计算示例：
- 文档1：
- 词频：$T_1 = [4,1,1]$
- 总词数：$n_1 = 6$
- 计算分子和分母：
- 对于 $k = 1$：
- $$\text{Numerator}_{1,1} = \pi_1^{(0)} \prod_{w=1}^{W} \left(\mu_{1w}^{(0)}\right)^{T_{1w}} = 0.5 \times 0.6^4 \times 0.3^1 \times 0.1^1 \qquad (20)$$

题目

- 取对数计算： $\ln \text{Numerator}_{1,1} = \ln 0.5 + 4\ln 0.6 + \ln 0.3 + \ln 0.1$ (21)
- 计算数值：
  - $\ln 0.5 \approx -0.6931$
  - $\ln 0.6 \approx -0.5108$
  - $\ln 0.3 \approx -1.2030$
  - $\ln 0.1 \approx -2.3026$
- 所以：

$$\ln \text{Numerator}_{1,1} = -0.6931 + 4 \times (-0.5108) + (-1.2030) + (-2.3026) \approx -6.2419 \quad (22)$$

- 因此： $\text{Numerator}_{1,1} = e^{-6.2419} \approx 0.00195$ (23)

题目

- 对于 $k = 2$:

- $$\text{Numerator}_{1,2} = \pi_2^{(0)} \prod_{w=1}^{W} \left(\mu_{2w}^{(0)}\right)^{T_{1w}} = 0.5 \times 0.2^4 \times 0.5^1 \times 0.3^1 \quad (24)$$
- 取对数计算:

- $$\ln\text{Numerator}_{1,2} = \ln 0.5 + 4\ln 0.2 + \ln 0.5 + \ln 0.3 \quad (25)$$
- 计算数值:

- $\ln 0.2 \approx -1.6094$
- 所以:

- $$\ln\text{Numerator}_{1,2} = -0.6931 + 4 \times (-1.6094) + (-0.6931) + (-1.2030) \approx -9.0268 \quad (26$$
- 因此:

- $$\text{Numerator}_{1,2} = e^{-9.0268} \approx 0.00012 \quad (27)$$

题目

- 计算归一化：

- $\gamma_{1,1} = \frac{0.00195}{0.00195+0.00012} \approx 0.942$ (28)

- $\gamma_{1,2} = 1 - \gamma_{1,1} \approx 0.058$ (29)

- 同样地，计算文档2和文档3的 γ_{dk}。

- 文档2：

  - 词频：$T_2 = [1,4,1]$

  - 总词数：$n_2 = 6$

- 计算得：

- $\gamma_{2,1} \approx 0.115$ (30)

- $\gamma_{2,2} \approx 0.885$ (31)

题目

- 文档3:

- 词频： $T_3 = [1,1,4]$

- 总词数： $n_3 = 6$

- 计算得:

- $\gamma_{3,1} \approx 0.022$ (32)

- $\gamma_{3,2} \approx 0.978$ (33)

题目

- *M步：更新参数*
- 更新混合系数 $\pi_k^{(1)}$：
- $$\pi_k^{(1)} = \frac{\sum_{d=1}^{D} \gamma_{dk}}{D} \qquad (34)$$

- 计算得：
- $$\pi_1^{(1)} = \frac{0.942+0.115+0.022}{3} \approx 0.3597 \qquad (35)$$
- $$\pi_2^{(1)} = 1 - \pi_1^{(1)} \approx 0.6403 \qquad (36)$$

题目

- **更新词分布参数 $\mu_{kw}^{(1)}$：** 对于每个主题 k，计算加权词频总和：

- **主题1 ($k = 1$)：**

- 总加权词数：

- $N_{k=1} = \sum_{d=1}^{D} \gamma_{dk}\, n_d = (0.942 + 0.115 + 0.022) \times 6 = 6.475$ (37)

- 词1的加权频数：

- $C_{1,1} = \sum_{d=1}^{D} \gamma_{d1}\, T_{d1} = 0.942 \times 4 + 0.115 \times 1 + 0.022 \times 1 = 3.904$ (38)

- 词2的加权频数：

- $C_{1,2} = 0.942 \times 1 + 0.115 \times 4 + 0.022 \times 1 = 1.423$ (39)

题目

- 词3的加权频数：
- $C_{1,3} = 0.942 \times 1 + 0.115 \times 1 + 0.022 \times 4 = 1.143$ (40)
- 更新词分布参数：
- $\mu_{1w}^{(1)} = \frac{C_{1,w}}{N_{k=1}}$ (41)

计算得：

- $\mu_{11}^{(1)} = \frac{3.904}{6.475} \approx 0.603$
- $\mu_{12}^{(1)} = \frac{1.423}{6.475} \approx 0.220$
- $\mu_{13}^{(1)} = \frac{1.143}{6.475} \approx 0.177$

题目

- 主题2 ($k = 2$):

- 总加权词数:

- $N_{k=2} = \sum_{d=1}^{D} \gamma_{dk} n_d = (0.058 + 0.885 + 0.978) \times 6 = 11.525$ (42

- 词1的加权频数:

- $C_{2,1} = \sum_{d=1}^{D} \gamma_{d2} T_{d1} = 0.058 \times 4 + 0.885 \times 1 + 0.978 \times 1 = 2.095$ (43)

- 词2的加权频数:

- $C_{2,2} = 0.058 \times 1 + 0.885 \times 4 + 0.978 \times 1 = 4.576$ (44)

题目

- **主题2 ($k = 2$):**

- 词3的加权频数:

- $C_{2,3} = 0.058 \times 1 + 0.885 \times 1 + 0.978 \times 4 = 4.857$ (45)

- 更新词分布参数:

- $\mu_{2w}^{(1)} = \frac{C_{2,w}}{N_{k=2}}$ (46)

  计算得:

  - $\mu_{21}^{(1)} = \frac{2.095}{11.525} \approx 0.182$
  - $\mu_{22}^{(1)} = \frac{4.576}{11.525} \approx 0.397$
  - $\mu_{23}^{(1)} = \frac{4.857}{11.525} \approx 0.422$

题目

- **3. 计算对数似然函数**
- ***(a) 对数似然函数的推导***
- **数据的对数似然函数为：**

$$L = \sum_{d=1}^{D} \ln P(d) = \sum_{d=1}^{D} \ln \left(\sum_{k=1}^{K} \pi_k \prod_{w=1}^{W} \mu_{kw}^{T_{dw}} \right) \qquad (47)$$

题目

- *(b) 计算 $L^{(0)}$ 和 $L^{(1)}$ 并比较*
- **计算 $L^{(0)}$ (初始参数下的对数似然)：**
- 对于每个文档，计算 $P(d)$。
- **文档1:**
  - 用初始参数计算:
- $P(d_1) = \pi_1^{(0)} \prod_{w=1}^{W} \left(\mu_{1w}^{(0)}\right)^{T_{1w}} + \pi_2^{(0)} \prod_{w=1}^{W} \left(\mu_{2w}^{(0)}\right)^{T_{1w}}$ (48)
  - 已在E步中计算过 $\text{Numerator}_{1,1}$ 和 $\text{Numerator}_{1,2}$，并且:
- $P(d_1) = \text{Numerator}_{1,1} + \text{Numerator}_{1,2} \approx 0.00195 + 0.00012 = 0.00207$ (4

题目

- **对数似然：**
- $L_1^{(0)} = \ln P(d_1) \approx \ln 0.00207 \approx -6.180 \qquad (50)$
- **同样计算文档2和文档3的对数似然：**
- **文档2：**
- $L_2^{(0)} \approx -6.156 \qquad (51)$
- **文档3：**
- $L_3^{(0)} \approx -7.780 \qquad (52)$
- **总对数似然：**
- $L^{(0)} = L_1^{(0)} + L_2^{(0)} + L_3^{(0)} \approx -6.180 - 6.156 - 7.780 = -20.116 \qquad (53)$

题目

- **计算 $L^{(1)}$（更新参数下的对数似然）：**
- 用更新后的参数计算 $P(d)$。
- **文档1:**
- 计算新的 $\text{Numerator}_{1,1}$ 和 $\text{Numerator}_{1,2}$:
  - $\pi_1^{(1)} \approx 0.3597$
  - $\pi_2^{(1)} \approx 0.6403$
  - $\mu_{1w}^{(1)}$ 和 $\mu_{2w}^{(1)}$ 如上所算。
- 计算新的 $\text{Numerator}_{1,1}$:

$$\text{Numerator}_{1,1} = 0.3597 \times 0.603^4 \times 0.220^1 \times 0.177^1 \quad (54)$$

- 计算 $\text{lnNumerator}_{1,1}$:

$$\text{lnNumerator}_{1,1} = \ln 0.3597 + 4\ln 0.603 + \ln 0.220 + \ln 0.177 \quad (55)$$

计算数值并得到 $L_1^{(1)}$。

题目

- 同样计算文档2和文档3的对数似然 $L_2^{(1)}$ 和 $L_3^{(1)}$。

- $L_1^{(1)} \approx -6.100$
- $L_2^{(1)} \approx -6.050$
- $L_3^{(1)} \approx -7.700$

- 总对数似然：

- $L^{(1)} = L_1^{(1)} + L_2^{(1)} + L_3^{(1)} \approx -6.100 - 6.050 - 7.700 = -19.850 \qquad (56)$

- *比较：*

- 初始对数似然：$L^{(0)} = -20.116$
- 更新后对数似然：$L^{(1)} = -19.850$

- 由于 $L^{(1)} > L^{(0)}$，即对数似然值增加了，说明模型在更新参数后更好地解释了数据。

End