高级机器学习
作业二

张逸凯 171840708

2020 年 12 月 25 日

学术诚信

本课程非常重视学术诚信规范，助教老师和助教同学将不遗余力地维护作业中的学术诚信规范的建立。希望所有选课学生能够对此予以重视。[1]

(1) 允许同学之间的相互讨论，但是署你名字的工作必须由你完成，不允许直接照搬任何已有的材料，必须独立完成作业的书写过程;

(2) 在完成作业过程中，对他人工作（出版物、互联网资料）中文本的直接照搬（包括原文的直接复制粘贴及语句的简单修改等）都将视为剽窃，剽窃者成绩将被取消。对于完成作业中有关键作用的公开资料，应予以明显引用；

(3) 如果发现作业之间高度相似将被判定为互相抄袭行为，抄袭和被抄袭双方的成绩都将被取消。因此请主动防止自己的作业被他人抄袭。

作业提交注意事项

(1) 请在LaTeX模板中第一页填写个人的姓名、学号信息；

(2) 本次作业需提交该pdf文件、 问题4可直接运行的源码， 将以上几个文件压缩成zip文件后上传。 zip文件格式为学号.zip， 例如170000001.zip； pdf文件格式为学号_姓名.pdf，例如170000001_张三.pdf。

(3) 未按照要求提交作业，或提交作业格式不正确，将会被扣除部分作业分数；

(4) 本次作业提交截止时间为12月25日23:59:59。除非有特殊情况（如因病缓交），否则截止时间后不接收作业，本次作业记零分。

[1] 参考尹一通老师高级算法课程中对学术诚信的说明。

1 [20pts] PAC Learning for Finite Hypothesis Sets

对于可分的有限假设空间，简单的 ERM 算法也可以导出 PAC 可学习性。请证明：

令 $\mathcal{H}$ 为可分的有限假设空间, D 为包含 m 个从 $\mathcal{D}$ 独立同分布采样所得的样本构成的训练集, 学习算法 $\mathfrak{L}$ 基于训练集 D 返回与训练集一致的假设 h_D, 对于任意 $c \in \mathcal{H}$, $0 < \epsilon, \delta < 1$, 如果有 $m \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})$, 则

$$P\left(E(h_D) \leq \epsilon\right) \geq 1 - \delta, \tag{1.1}$$

即$E(h) \leq \epsilon$ 以至少 $1-\delta$ 的概率成立.

提示：注意到 h_D 必然满足 $\widehat{E}_D(h_D) = 0$.

Solution. 由题中条件, 我们有:

$$\begin{aligned}
&\because m \geqslant \frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right) \\
&\Rightarrow e^{-m\epsilon} \leqslant \frac{\delta}{|\mathcal{H}|} \\
&\therefore |\mathcal{H}|e^{-m\epsilon} \leqslant \delta \;\Rightarrow\; 1-\delta \leqslant 1-|\mathcal{H}|e^{-m\epsilon}
\end{aligned}$$

考虑在分布 $\mathcal{D}$ 上随机采样而得到的任何一个样例 $(\boldsymbol{x}, y)$, 且假定 h 的泛化误差$\geqslant \epsilon$, 我们有:

$$\begin{aligned}
P(h(\boldsymbol{x}) = y) &= 1 - P(h(\boldsymbol{x}) \neq y) \\
&= 1 - E(h) \\
&\leqslant 1 - \epsilon
\end{aligned}$$

考虑 m个样本独立同分布采样, 自然地有:

$$\begin{aligned}
P\left(\left(h\left(\boldsymbol{x}_1\right) = y_1\right) \wedge \ldots \wedge \left(h\left(\boldsymbol{x}_m\right) = y_m\right)\right) &= \left(1 - P(h(\boldsymbol{x}) \neq y)\right)^m \\
&\leqslant (1-\epsilon)^m
\end{aligned}$$

因为我们事先不知道学习算法 $\mathfrak{L}$ 会输出 $\mathcal{H}$ 中的哪个假设, 即对同一个观察的数据集D的输出也可能是不确定的, 考虑 $\mathcal{H}$ 中所有泛化误差大于 ϵ 且经验误差为 0 的假设(即在训练集上表现完美的假设):

$$\begin{aligned}
&P\left(E(h_D) > \epsilon\right) \\
&= P(h \in \mathcal{H} : E(h) > \epsilon \wedge \widehat{E}(h) = 0) \\
&= \sum_{i}^{|\mathcal{H}|} P\left(E\left(h_i\right) > \epsilon \wedge \widehat{E}\left(h_i\right) = 0\right) \\
&\leqslant |\mathcal{H}|(1-\epsilon)^m
\end{aligned}$$

当 $\epsilon \in (0, 1]$, m 是正整数时, $(1-\epsilon)^m \leqslant e^{-m\epsilon}$, 求导易证. 所以 $|\mathcal{H}|(1-\epsilon)^m \leqslant |\mathcal{H}|e^{-m\epsilon}$.

综上所述,

$$\begin{aligned} P\left(E(h_D) \leq \epsilon\right) &= 1 - P\left(E(h_D) > \epsilon\right) \\ &= 1 - P(h \in \mathcal{H} : E(h) > \epsilon \wedge \widehat{E}(h) = 0) \\ &\geqslant 1 - |\mathcal{H}|(1-\epsilon)^m \\ &\geqslant 1 - |\mathcal{H}|e^m \\ &\geqslant 1 - \delta \end{aligned}$$

2 [20pts] semi-supervised learning

多标记图半监督学习算法 [Zhou et al., 2003]的正则化框架如下(另见西瓜书p303)。

$$\mathcal{Q}(F) = \frac{1}{2}\left(\sum_{i,j=1}^{n} W_{ij}\left\|\frac{1}{\sqrt{d_i}}F_i - \frac{1}{\sqrt{d_j}}F_j\right\|^2\right) + \mu\sum_{i=1}^{n}\|F_i - Y_i\|^2 \tag{2.1}$$

1. **[10pts]** 求正则化框架的最优解F^*。
2. **[10pts]**试说明该正则化框架与书中p303页多分类标记传播算法之间的关系。

Solution. .

- ***(1)*** 由题意, 转化为矩阵表达, 注意到 W 是对称阵(这在如下等式中做了相应的下标代换).

$$\begin{aligned} &\frac{1}{2}\left(\sum_{i,j=1}^{n} W_{ij}\left\|\frac{1}{\sqrt{d_i}}F_i - \frac{1}{\sqrt{d_j}}F_j\right\|^2\right) \\ =& \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\frac{1}{d_i}F_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\frac{1}{d_j}F_j^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\frac{1}{\sqrt{d_i d_j}}F_i \cdot F_j\right) \\ =& F^\top F - F^\top D^{-\frac{1}{2}} W D^{-\frac{1}{2}} F \end{aligned}$$

式 2.1 中 $\mathcal{Q}(F)$ 关于F 求偏导, 我们有:

$$\left.\frac{\partial \mathcal{Q}}{\partial F}\right|_{F=F^\star} = F^\star - SF^\star + \mu\left(F^\star - Y\right) = 0$$

其中 $S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$.

化简得:

$$F^\star - \frac{1}{1+\mu}SF^\star - \frac{\mu}{1+\mu}Y = 0$$

不妨令 $\alpha = \frac{1}{1+\mu}$, 我们有 $(I - \alpha S)F^\star = (1-\alpha)Y$

因为 $I - \alpha S$ 是可逆的, 可以得到:

$$F^\star = (1-\alpha)(I - \alpha S)^{-1}Y$$

Remark 本题部分参考自 *[Zhou et al., 2003]*

- ***(2)***

  式 2.1 中等号右边第二项是迫使学得结果在有标记样本上的预测与真实标记尽可能相同，而第一项则迫使相近样本具有相似的标记，这与多分类标记传播算法的半监督学习假设：相似的样本具有相似的输出 是相符合的. 而且上述正则化框架的最优解恰为标签传播算法的迭代收敛解 $\boldsymbol{F}^{\star}$.

  不同之处在于标签传播算法考虑输出连续值，而上述正则化框架考虑离散的类别标记.

3 [30pts] Mixture Models

一个由K个组分(component)构成的多维高斯混合模型的概率密度函数如下:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} P(z=k)\, p(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3.1}$$

其中z是隐变量，$P(z)$表示K维离散分布，其参数为$\boldsymbol{\pi}$，即$p(z=k) = \pi_k$。$p(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$表示参数为$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$的多维高斯分布。

1. **[10pts]** 请使用盘式记法表示高斯混合模型。
2. **[10pts]** 考虑高斯混合模型的一个具体的情形，其中各个分量的协方差矩阵$\boldsymbol{\Sigma}_k$全部被限制为一个共同的值$\boldsymbol{\Sigma}$。求EM算法下参数$\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}$的更新公式。
3. **[10pts]** 考虑一个由下面的混合概率分布给出的概率密度模型:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|k) \tag{3.2}$$

并且假设我们将$\boldsymbol{x}$划分为两部分，即$\boldsymbol{x} = (\boldsymbol{x}_a, \boldsymbol{x}_b)$。证明条件概率分布$p(\boldsymbol{x}_a|\boldsymbol{x}_b)$本身是一个混合概率分布。求混合系数以及分量概率密度的表达式。(注意此题没有规定$p(\boldsymbol{x}|k)$的具体形式)

Solution. .

(1)

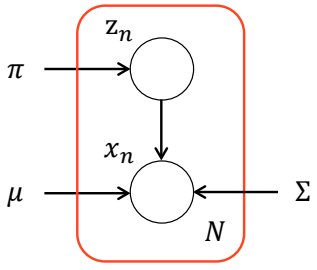

图 1: PPT作图!

其中N表示独立同分布数据点x_n的个数, 隐变量为z_n. 其中圆圈表示变量(最好涂上颜色).

Remark 这里的结果学习自*PRML9.2*节.

(2)

这里以π_k为例, **推导一遍**, 类似地有$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}$的更新公式(而且更简单, 因为这两个参数关于$Q$是无约束的).

*EM*算法的基本表达式为: $\theta^{t+1} = \underset{\theta}{argmax}\, \mathbb{E}_{z|x,\theta_t}[p(x,z|\theta)]$

对数据集来说, 在*GMM*中:

$$\begin{aligned} Q(\theta, \theta^t) &= \sum_z [\log \prod_{i=1}^N p(x_i, z_i|\theta)] \prod_{i=1}^N p(z_i|x_i, \theta^t) \\ &= \sum_z [\sum_{i=1}^N \log p(x_i, z_i|\theta)] \prod_{i=1}^N p(z_i|x_i, \theta^t) \end{aligned}$$

对中间的求和进行化简, 考虑其中一项的形式为:

$$\begin{aligned} \sum_z \log p(x_1, z_1|\theta) \prod_{i=1}^N p(z_i|x_i, \theta^t) &= \sum_z \log p(x_1, z_1|\theta) p(z_1|x_1, \theta^t) \prod_{i=2}^N p(z_i|x_i, \theta^t) \\ &= \sum_{z_1} \log p(x_1, z_1|\theta) p(z_1|x_1, \theta^t) \end{aligned}$$

所以有:

$$Q(\theta, \theta^t) = \sum_{i=1}^N \sum_{z_i} \log p(x_i, z_i|\theta) p(z_i|x_i, \theta^t)$$

其中

$$p(x,z|\theta) = p(z|\theta)p(x|z,\theta) = \pi_z \mathcal{N}(x|\mu_z, \Sigma), \;\; p(z|x,\theta^t) = \frac{p(x,z|\theta^t)}{p(x|\theta^t)} = \frac{\pi_z^t \mathcal{N}(x|\mu_z^t, \Sigma^t)}{\sum\limits_k \pi_k^t \mathcal{N}(x|\mu_k^t, \Sigma^t)}$$

整理一下, 我们需要对下述Q求最大值:

$$Q = \sum_{k=1}^K \sum_{i=1}^N (\log \pi_k + \log \mathcal{N}(x_i|\mu_k, \Sigma)) \frac{\pi_k^t \mathcal{N}(x_i|\mu_k^t, \Sigma^t)}{\sum\limits_k \pi_k^t \mathcal{N}(x_i|\mu_k^t, \Sigma^t)}$$

Remark 以上推导综合了西瓜书和*PRML*的知识.

- π_k^{t+1}:

  问题转化为:

$$\begin{aligned} &\pi_k^{t+1} = \underset{\pi_k}{argmax}\, Q \\ &s.t. \sum_{k=1}^K \pi_k = 1 \end{aligned}$$

拉格朗日函数:

$$\mathcal{L}(\pi_k, \lambda) = \sum_{k=1}^{K}\sum_{i=1}^{N} \log \pi_k p(z_i = k|x_i, \theta^t) - \lambda(1 - \sum_{k=1}^{K}\pi_k)$$

$$\begin{aligned} \frac{\partial}{\partial \pi_k}\mathcal{L} &= \sum_{i=1}^{N}\frac{1}{\pi_k}p(z_i = k|x_i, \theta^t) + \lambda = 0 \\ \Rightarrow \sum_{k}\sum_{i=1}^{N}\frac{1}{\pi_k}p(z_i = k|x_i, \theta^t) &+ \lambda\sum_{k}\pi_k = 0 \\ \Rightarrow \lambda &= -N \end{aligned}$$

所以 π_k 的更新公式为:

$$\pi_k^{t+1} = \frac{1}{N}\sum_{i=1}^{N} p(z_i = k|x_i, \theta^t)$$

$p(z_i = k|x_i, \theta^t)$ 如前所述为: $\frac{\pi_k^t \mathcal{N}(x_i|\mu_k^t, \Sigma^t)}{\sum_k \pi_k^t \mathcal{N}(x_i|\mu_k^t, \Sigma^t)}$

- μ_k^{t+1}:

$$\boldsymbol{\mu}_k^{t+1} = \frac{1}{N_k}\sum_{i=1}^{N}\frac{\pi_i \mathcal{N}\left(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k^t, \boldsymbol{\Sigma}\right)}{\sum_{k=1}^{K}\pi_k \mathcal{N}\left(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k^t, \boldsymbol{\Sigma}\right)}\boldsymbol{x}_i$$

- Σ^{t+1}:

$$\Sigma^{t+1} = \frac{1}{N_k}\sum_{i=1}^{N}\frac{\pi_i \mathcal{N}\left(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k^t, \boldsymbol{\Sigma}\right)}{\sum_{k=1}^{K}\pi_k \mathcal{N}\left(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k^t, \boldsymbol{\Sigma}\right)}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{t+1}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{t+1}\right)^\top$$

(3)

- 将$\boldsymbol{x}$以维度为划分, 成两部分.

  注意到此时由条件概率拆分:

$$p(\boldsymbol{x}_a \mid \boldsymbol{x}_b) = \frac{p(\boldsymbol{x}_a, \boldsymbol{x}_b)}{p(\boldsymbol{x}_b)} = \frac{\sum_{k=1}^{K}\pi_k p\left(\boldsymbol{x}|k\right)}{p(\boldsymbol{x}_b)}$$

  $\forall\, a \in \mathcal{A}$, 验证有满足归一化条件. 此时仍为混合分布, 混合系数为 $\frac{\pi_k}{p(\boldsymbol{x}_b)}$.

- 将$\boldsymbol{x}$以数据点为划分, 成两部分.

  注意到在条件概率的拆分式中分子分母都是混合概率分布, 当 $\boldsymbol{x}_a, \boldsymbol{x}_b$存在相交类别的元素时无法化简, 所以此时并不是题目考虑的情况.

4 [30pts]Latent Dirichlet Allocation

我们提供了一个包含8888条新闻的数据集news.txt.zip，该数据集中每一行是一条新闻。在该数据集上完成LDA模型的使用及实现。

数据预处理提示：你可能需要完成分词及去掉一些停用词等预处理工作。

在本题中需要完成:

1. **[10pts]**使用开源的LDA库（如scikit-learn），计算给出$K = \{5, 10, 20\}$个话题时，每个话题下概率最大的10个词及其概率。
2. **[20pts]**不借助开源库，手动实现LDA模型，计算给出$K = \{5, 10, 20\}$个话题时，每个话题下概率最大的10个词及其概率。

注：需要在报告中描述模型计算的结果，以及如何复现自己的结果，提交的作业中至少应该包含lda_use.py和lda.py两个文件，分别为使用和不使用第三方库的源码。

Solution. .

- 模型计算结果, 预处理文件news_preprocessing.txt:

  请从该链接中下载:

  https://drive.google.com/drive/folders/1Bd0kPnvwFkUkAymNQrp8QydVQUnYAFEk?usp=sharing

- 如何复现:

  注意: 请将 news.txt 和 news_preprocessing.txt 与代码文件放入同一路径中, 执行:

  ```
  python lda.py
  ```

  即可直接运行.

参考文献

D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16 (NIPS)*, pages 321–328, 2003.