高级机器学习
作业二

周韧哲 181220076

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. .

令

$$\mathcal{H}_{ERM} = \arg\min_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} I(h(x_i) \neq y_i)$$

$$\mathcal{H}' = \{h \in \mathcal{H}, E(h) > \epsilon\}$$

注意到假设空间可分，所以也有 $\mathcal{H}_{ERM} = \{h \in \mathcal{H}, \hat{E}_D(h) = 0\}$。对于 $\mathcal{H}_{ERM}$ 中的假设，我们计算其泛化误差大于 ϵ 的假设子集 $\mathcal{H}'_{ERM} = \{h \in \mathcal{H}, \hat{E}_D(h) = 0, E(h) > \epsilon\}$，则有

$$\begin{aligned} P(\mathcal{H}'_{ERM}) &= P(\bigcup_{h\in\mathcal{H}'} \hat{E}_D(h) = 0) \\ &\leq \sum_{h\in\mathcal{H}'} P(\hat{E}_D(h) = 0) \end{aligned}$$

对于分布 D 上随机采样的任何样例 (x, y), 有 $P(h(x) = y) = 1 - P(h(x) \neq y) = 1 - E(h) \leq 1 - \epsilon$。因为 D 是从 $\mathcal{D}$ 中独立同分布采样而来的, 所以 $P(\hat{E}_D(h) = 0) = (1 - P(h(x) \neq y))^m \leq (1-\epsilon)^m$。所以

$$P(\mathcal{H}'_{ERM}) = P(E(h_D) > \epsilon) \leq \sum_{h\in\mathcal{H}'} (1-\epsilon)^m = |\mathcal{H}'|(1-\epsilon)^m \leq |\mathcal{H}|(1-\epsilon)^m \leq |\mathcal{H}|e^{-\epsilon m}$$

当 $m \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})$ 时，则有

$$P(E(h_D) > \epsilon) \leq \delta$$

从而,

$$P\left(E(h_D) \leq \epsilon\right) \geq 1 - \delta$$

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)。见 [Ng]

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

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. 对于正则化框架的左侧容易得到：

$$\begin{aligned}\mathcal{Q}_l(F) &= \frac{1}{2}\left(\sum_{i,j=1}^{n} W_{ij}\left\|\frac{1}{\sqrt{d_i}}F_i - \frac{1}{\sqrt{d_j}}F_j\right\|^2\right)\\ &= \frac{1}{2}\sum_{i,j=1}^{n} W_{ij}\sum_{k=1}^{|\mathcal{Y}|}(\frac{1}{\sqrt{d_i}}F_{ik} - \frac{1}{\sqrt{d_j}}F_{jk})^2\\ &= \frac{1}{2}\sum_{k=1}^{|\mathcal{Y}|}(\sum_{i=1}^{n}\frac{\sum_{j=1}^{n}W_{ij}}{d_i}F_{ik}^2 + \sum_{j=1}^{n}\frac{\sum_{i=1}^{n}W_{ij}}{d_j}F_{jk}^2 - 2\sum_{i,j=1}^{n}\frac{W_{ij}}{\sqrt{d_i d_j}}F_{ik}F_{jk})\end{aligned}$$

定义 $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$，$f_k$ 为 $\mathbf{F}$ 的第 k 列，由于 $d_i = \sum_{j=1}^{n} W_{ij} = \sum_{i=1}^{n} W_{ji}$，所以有

$$\begin{aligned}\mathcal{Q}_l(F) &= \frac{1}{2}\sum_{k=1}^{|\mathcal{Y}|}(\sum_{i=1}^{n}F_{ik}^2 + \sum_{j=1}^{n}F_{jk}^2 - 2\sum_{i,j=1}^{n}\frac{W_{ij}}{\sqrt{d_i d_j}}F_{ik}F_{jk})\\ &= \sum_{k=1}^{|\mathcal{Y}|}(\sum_{i=1}^{n}F_{ik}^2 - \sum_{i,j=1}^{n}\frac{W_{ij}}{\sqrt{d_i d_j}}F_{ik}F_{jk})\\ &= \sum_{k=1}^{|\mathcal{Y}|}(f_k^T\mathbf{I}f_k - f_k^T\mathbf{S}f_k)\\ &= tr(\mathbf{F}^T(\mathbf{I}-\mathbf{S})F)\end{aligned}$$

对于正则化框架的右侧容易得到：

$$\begin{aligned}\mathcal{Q}_r(F) &= \mu\sum_{i=1}^{n}\|F_i - Y_i\|^2\\ &= \mu\sum_{i,j=1}^{n}(F_{ij} - Y_{ij})^2\\ &= \mu\|\mathbf{F}-\mathbf{Y}\|_F^2\\ &= tr(\mu(\mathbf{F}-\mathbf{Y})^T(\mathbf{F}-\mathbf{Y}))\end{aligned}$$

所以

$$\mathcal{Q}(\mathbf{F}) = tr(\mathbf{F}^T(\mathbf{I}-\mathbf{S})F + \mu(\mathbf{F}-\mathbf{Y})^T(\mathbf{F}-\mathbf{Y}))$$

对其求导可得

$$\frac{\partial\mathcal{Q}(\mathbf{F})}{\partial\mathbf{F}} = 2(\mathbf{I}-\mathbf{S})\mathbf{F} + 2\mu(\mathbf{F}-\mathbf{Y})$$

令导数为 $\mathbf{0}$，从而

$$\mathbf{F}^* = (1-\frac{1}{1+\mu})(\mathbf{I}-\frac{1}{1+\mu}\mathbf{S})\mathbf{Y}$$

2. 容易看出，当 $\mu = \frac{1}{1+\alpha}$ 时，多分类标记传播算法就是该正则化框架的迭代解，通过构造迭代式 $\mathbf{F}(t+1) = \alpha\mathbf{S}\mathbf{F}(t) + (1-\alpha)\mathbf{Y}$ 来迭代地求解该正则化框架的最优解。折中参数 α 越大，则迭代过程中结果偏离初始化项 $\mathbf{Y}$ 越大，对应的正则化参数 μ 越小，即限制 $\sum_{i=1}^{n}\|F_i - Y_i\|^2$ 更宽松，$\mathbf{F}$ 更偏离 $\mathbf{Y}$。

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)

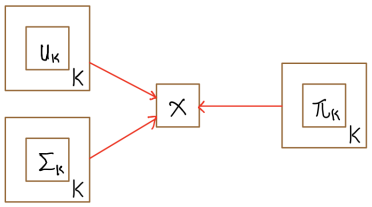

图 1: 盘式计法

Solution. .

1. 见图*1*。

2. 在 E 步，已知参数 $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ 推断隐变量分布，首先定义:

$$\begin{aligned} w_k^i &:= p(z=k|\boldsymbol{x}_i) \\ &= \frac{p(\boldsymbol{x}_i|z=k)p(z=k)}{p(\boldsymbol{x}_i)} \\ &= \frac{p(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})\boldsymbol{\pi}_k}{\sum_{k=1}^{K} \boldsymbol{\pi}_k p(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})} \end{aligned}$$

然后写出似然函数：

$$\begin{aligned}
L(\boldsymbol{X}) &= \sum_{i=1}^{N} \ln p(\boldsymbol{x}_i, z|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \\
&= \sum_{i=1}^{N} \ln \sum_{k=1}^{K} p(\boldsymbol{x}_i|z=k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) p(z=k) \\
&= \sum_{i=1}^{N} \ln \sum_{k=1}^{K} p(z=k|\boldsymbol{x}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \frac{p(\boldsymbol{x}_i|z=k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) p(z=k)}{p(z=k|\boldsymbol{x}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma})}
\end{aligned}$$

由 *Jensen* 不等式容易得到

$$\begin{aligned}
L(\boldsymbol{X}) &\geq \sum_{i=1}^{N} \sum_{k=1}^{K} p(z=k|\boldsymbol{x}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \ln \frac{p(\boldsymbol{x}_i|z=k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) p(z=k)}{p(z=k|\boldsymbol{x}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma})} \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} w_k^i \ln \frac{\frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp(-\frac{1}{2}(\boldsymbol{x_i} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x_i} - \boldsymbol{\mu}_k)) \boldsymbol{\pi}_k}{w_k^i}
\end{aligned}$$

不等号右边即是 $\mathcal{Q}(\boldsymbol{\pi}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$，最大化 $\mathcal{Q}(\boldsymbol{\pi}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ 即最大化 $L(\boldsymbol{X})$。
在 M 步，最大化 $\mathcal{Q}(\boldsymbol{\pi}_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$：

$$\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\mu}_k} &= \sum_{i=1}^{N} w_k^i \frac{\partial(-\frac{1}{2}(\boldsymbol{x_i} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x_i} - \boldsymbol{\mu}_k))}{\partial \boldsymbol{\mu}_k} \\
&= \sum_{i=1}^{N} w_k^i \boldsymbol{\Sigma}^{-1} (\boldsymbol{x_i} - \boldsymbol{\mu}_k) \\
&= 0 \\
\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\Sigma}} &= -\frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{K} w_k^i [\frac{\partial((\boldsymbol{x_i} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x_i} - \boldsymbol{\mu}_k))}{\partial \boldsymbol{\Sigma}} + \frac{\partial \ln |\boldsymbol{\Sigma}|}{\partial \boldsymbol{\Sigma}}] \\
&= -\frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{K} w_k^i [-\boldsymbol{\Sigma}^{-1^T} (\boldsymbol{x_i} - \boldsymbol{\mu}_k)(\boldsymbol{x_i} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1^T} + \boldsymbol{\Sigma}^{-1^T}] \\
&= 0
\end{aligned}$$

所以

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} w_k^i \boldsymbol{x_i}}{\sum_{i=1}^{N} w_k^i}$$

$$\boldsymbol{\Sigma} = \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} w_k^i (\boldsymbol{x_i} - \boldsymbol{\mu}_k)(\boldsymbol{x_i} - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^{N} \sum_{k=1}^{K} w_k^i}$$

$\boldsymbol{\pi}_k$ 还需满足 $\sum_{k=1}^{K} \boldsymbol{\pi}_k = 1$，所以加上拉格朗日项：

$$\mathcal{Q}' := \mathcal{Q} + \lambda(\sum_{k=1}^{K} \boldsymbol{\pi}_k - 1)$$

求导得:

$$\frac{\partial \mathcal{Q}'}{\partial \boldsymbol{\pi}_k} = \sum_{i=1}^{N} w_k^i \frac{1}{\boldsymbol{\pi}_k} + \lambda = 0$$
$$\frac{\partial \mathcal{Q}'}{\partial \lambda} = \sum_{k=1}^{K} \boldsymbol{\pi}_k - 1 = 0$$

得到 $\boldsymbol{\pi}_k = -\frac{\sum_{i=1}^{N} w_k^i}{\lambda}$，又有 $1 = \sum_{k=1}^{K} \boldsymbol{\pi}_k = -\frac{\sum_{i=1}^{N}\sum_{k=1}^{K} w_k^i}{\lambda} - \frac{N}{\lambda}$，所以

$$\lambda = -N, \boldsymbol{\pi}_k = \frac{\sum_{i=1}^{N} w_k^i}{N}$$

以上部分参考自 *[Ng]*。

3. 由贝叶斯定理和全概率公式易知:

$$\begin{aligned} p(\boldsymbol{x}_a|\boldsymbol{x}_b) &= \frac{p(\boldsymbol{x}_a, \boldsymbol{x}_b)}{p(\boldsymbol{x}_b)} \\ &= \frac{p(\boldsymbol{x}_a, \boldsymbol{x}_b)}{\int p(\boldsymbol{x}_a, \boldsymbol{x}_b) d\boldsymbol{x}_a} \\ &= \frac{\sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_a, \boldsymbol{x}_b|k)}{\int \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_a, \boldsymbol{x}_b|k) d\boldsymbol{x}_a} \\ &= \frac{\sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_a, \boldsymbol{x}_b|k)}{\sum_{k=1}^{K} \pi_k \int p(\boldsymbol{x}_a, \boldsymbol{x}_b|k) d\boldsymbol{x}_a} \end{aligned}$$

给定 $\boldsymbol{x}_b$ 时，$\int p(\boldsymbol{x}_a, \boldsymbol{x}_b|k) d\boldsymbol{x}_a$ 为定值，因此分母也为定值，又因为 $\sum_{k=1}^{K} \pi_k = 1$，从而分布

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

为混合概率分布，其第 k 个混合系数为 π_k，对应的分量概率密度为 $\frac{p(\boldsymbol{x}_a, \boldsymbol{x}_b|k)}{\sum_{k=1}^{K} \pi_k \int p(\boldsymbol{x}_a, \boldsymbol{x}_b|k) d\boldsymbol{x}_a}$。

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. 我的所有代码均在 *manjaro* 系统中实现，使用的 *python* 版本为 *python3.8*。

1. 我使用开源的 *sklearn* 的 *LDA* 模型，其实现位于 *lda_use.py* 中，可选命令行参数有 *topic_nums*, *max_iter_nums* 等。命令行输入 ***python lda_use.py —topic_nums 5*** 可运行并获得 5 个话题下概率最大的 10 个词及其概率，结果保存在*./assets/*下。最终我迭代了 1000 次，结果文件为 ***./assets/results_use_topick_iter1000.txt(k=5,10,20)***。

2. 由于数据预处理需要较长时间，我将预处理的模型保存在*./assets/lda.b* 中，命令行参数 *load* 默认为 *True*，会加载该模型；否则会重新处理数据。可选命令行参数有 *topic_nums*, *alpha*, *beta*, *max_iter_nums* 等。命令行输入 ***python lda.py —topic_nums 5*** 可运行并获得 5 个话题下概率最大的 10 个词及其概率，保存在*./assets/*下。最终我迭代了 200 次，结果文件分别为 ***./assets/results_topick_iter200.txt(k=5,10,20)***。

由于结果数量太多，不方便写在 *tex* 文件中，因此请查看*./assets/*下的结果文件。由于模型复杂度高，运行时间长，可以减小迭代次数以快速验证模型。

参考文献

A. Ng. Cs229 lecture notes, the em algorithm. https://see.stanford.edu/materials/aimlcs229/cs229-notes8.pdf.

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.