清华大学电子工程系

媒体与认知 课堂 2

2022-2023 学年春季学期

作业 3

2023 年 4 月 19 日

本次作业包含特征提取与降维、基于贝叶斯决策的统计模式识别、KNN 和 KMeans 算法、支持向量机等相关内容，通过编程实现加深对支持向量机的理解。理论部分包含第 1，2 题，所有同学均需完成；编程部分为 3、4、5、6 题，已确认自选课题的同学只需完成第 7 题即可。

1. 单选题（15 分）
2. 计算题（15 分）
3. 完成支持向量机的程序代码（30 分）
4. 训练/测试/可视化/比较（25 分）
5. 使用 SVM 算法在 MNIST 数据集上进行分类（5 分）
6. 撰写作业报告（10 分）
7. 汇报自选课题进度 *（70 分）

理论部分

1 单选题（15 分）

1.1 设 $\phi(t) \in L^2(\mathbb{R})(L^2(\mathbb{R})$ 表示实数域上的平方可积函数空间，即能量有限信号空间)，其对应的傅里叶变换为 $\psi(\omega)$，如果满足 $C_\phi = \int_{\mathbb{R}} \frac{|\psi(\omega)|^2}{|\omega|} d\omega < \infty$，则称 $\phi(t)$ 为一个小波母函数。对小波母函数进行尺度变换和平移以得到一组小波基函数。如果变换后的小波基函数为 $|a|\phi(\frac{a^2}{b}(bt+1))$，则尺度因子和平移因子为：

(A) a，b

(B) $\frac{1}{a}$，b

(C) $\frac{1}{a^2}$，$-\frac{1}{b}$

(D) $\frac{b}{a^2}$，$\frac{1}{b}$

1.2 关于主成分分析 PCA 和线性判别分析 LDA，以下说法正确的是：

(A) PCA 是有监督的，LDA 是无监督的

(B) PCA 是最小均方误差准则下区分多类数据，LDA 是最小均方误差准则下保留原始数据信息

(C) PCA 和 LDA 都是基于高斯假设的非线性特征变换法

(D) PCA 取数据投影方差最大的方向，LDA 取分类性能最好的投影方向

1.3 以下说法正确的是：

(A) 贝叶斯分类器是鉴别式模型

(B) 支持向量机是生成式模型

(C) XOR 问题是线性可分的

(D) 基于核函数的 SVM 可以用于求解 XOR 问题

1.4 对于正态分布的贝叶斯决策，下面哪个条件会使得两类问题的分类界面退化成线性超平面，且分类超平面与两类中心点的连线垂直：

(A) $\Sigma_1 = \Sigma_2$

(B) Σ_1 和 Σ_2 是对角阵

(C) 每一类先验概率相等

(D) $\Sigma_i = \sigma^2 I$

1.5 以下说法正确是：

(A) KNN 算法不能用于解决回归问题

(B) 对于某一样本集，多次使用 KMeans 算法得到的结果相同

(C) EM 算法可以得到样本概率分布等模型参数的估计

(D) KMeans 算法和 EM 算法必须满足高斯分布的假设

2 计算题（15 分）

2.1 距离地球很远有一个双星系统，其中有两个星球 A 和 B，从地球观测时，星球 A 和 B 的位置重叠。已知星球 A 有 60% 的部分是海洋，其余是陆地，而星球 B 则全是陆地。某一时刻，观测到星球 A 或 B 的概率相同，假设此时观测到该星球上的陆地，计算该星球是星球 A 的概率。
(提示：全概率公式 $P(Y)=\sum_{i=1}^{N}P(Y|X_i)P(X_i)$**)**

2.2 给定 5 维空间中的 6 个样本点，写成如下的 6x5 维矩阵 X**，其中每一行代表一个样本点**

$$X=\begin{bmatrix}1&1&1&0&0\\-3&-3&-3&0&0\\2&2&2&0&0\\0&0&0&-1&-1\\0&0&0&2&2\\0&0&0&-1&-1\end{bmatrix}$$

(i) 计算该数据集的均值;

(ii) 求解矩阵 X 的 SVD 分解;（提示：矩阵 X 的 SVD 分解的形式为 $\begin{bmatrix}a&0\\-3a&0\\2a&0\\0&b\\0&-2b\\0&b\end{bmatrix}\times\begin{bmatrix}\sigma_1&0\\0&\sigma_2\end{bmatrix}\times\begin{bmatrix}c&c&c&0&0\\0&0&0&d&d\end{bmatrix}$）

(iii) 计算该数据集的第一主成分（即最大特征值对应的归一化特征向量）。

2.3 设有两类正态分布的样本集，第一类均值为 $\mu_1 = [1, 2]^T$，第二类均值为 $\mu_2 = [-1, -2]^T$。两类样本集的协方差矩阵相等 $\Sigma_1 = \Sigma_2 = \Sigma = \begin{bmatrix} 3.0 & 1.0 \\ 1.0 & 2.0 \end{bmatrix}$，先验概率分别为 $p(\omega_1) = 0.6$，$p(\omega_2) = 0.4$。试计算分类界面，并对样本 $x = [0, 0]^T$ 分类。

2.4 使用 KMeans 算法对 2 维空间中 6 个点 $(0, 2)$，$(2, 0)$，$(2, 3)$，$(3, 2)$，$(4, 0)$，$(5, 4)$ 进行聚类，距离函数选择哈密顿距离 $d = |x_1 - x_2| + |y_1 - y_2|$。

(i) 起始聚类中心选择 $(2, 3)$ 和 $(4, 0)$，计算聚类中心；

(ii) 起始聚类中心选择 $(2, 0)$ 和 $(5, 4)$，计算聚类中心。

2.5 试用 Lagrange 乘子法计算 $x^2 + 2y^2$ 在 $x^2 + y^2 \leq 1$ 区域内的极值。

编程部分

编程部分包括第 3，4，5，6 题，选择自选课题的同学请完成第 7 题。

3 实现 hinge loss 模拟支持向量机并运行自动评判程序（25 分）

在本任务中，实现 hinge loss 模拟支持向量机的代码。
在开始前，请先安装 libsvm 库，在 anaconda 命令行终端中可执行下述指令以安装最新版的 libsvm 库：pip install -U libsvm-official
程序清单如下：

文件或目录	说明	注意事项
hw3.zip	作业 3 程序压缩包	解压可以得到下列文件
classify_hw.py	线性分类程序	需要完成代码
svm_hw.py	线性层 +hinge loss 模拟 SVM 程序	需要完成代码
check.py	自动评判程序	请勿修改
process_mnist.py	提取 MNIST 数据集中两类数字特征程序	请勿修改
\data	存放本次作业所用数据集	请勿修改
\MNIST	存放 MNIST 手写数字图片数据集	请勿修改

请在程序"???" 提示处补全代码，程序中每处需要补全代码的地方均有注释提示，请注意阅读。需要补全代码的清单如下：

序号	内容	程序	补全行号	说明
1	class Linear	svm_hw.py	29	实现线性层的前向计算过程
2	class Linear	svm_hw.py	46，47	实现线性层的反向传播过程
3	class Hinge	svm_hw.py	63	实现 hinge loss
4	class Hinge	svm_hw.py	75，76	实现 loss 层的反向传播过程
5	class SVM_HINGE	svm_hw.py	90，91	定义线性层的参数 W，b

在补全代码之后，可以运行自动评判程序检验 svm_hw.py 代码实现效果：
运行命令：python check.py
若代码正确则可以进行后续任务了。本任务测试成功的截图需要附在作业报告中。

4 训练/验证/可视化/比较（30 分）

使用支持向量机完成线性分类任务：字符/背景图片特征分类，对比 libsvm 库的分类结果和使用线性层 +hinge loss 的模拟结果。

4.1 Hinge loss 模拟 SVM 的训练及验证

该部分需要大家补全 classify_hw.py 中的代码，程序中每处需要补全代码的地方均有注释提示，请注意阅读。需要补全代码的清单如下：

序号	内容	程序	补全行号	说明
1	class FeatureDataset	classify_hw.py	所有的???	实现图像特征的数据类
2	def train_val_hinge	classify_hw.py	所有的???	实现 hinge loss 模拟 SVM 的训练，验证代码

补全好代码以后，即可执行 classify_hw.py 实现训练和验证过程，
classify_hw.py 可以调整的参数包括：

序号	名称	说明
1	mode	控制代码运行的方式，其中 hinge 表示使用 hinge loss 模拟 SVM，baseline 表示使用 libsvm 库实现分类
2	train_file_path	训练数据的文件路径
3	val_file_path	验证数据的文件路径
4	device	程序运行的设备，可以选择‘cpu’或’cuda’
5	feat_dim	特征的维数
6	epoch	训练的总轮数
7	valInterval	每几轮执行一次验证
8	lr	训练的学习率
9	C	引入松弛变量后添加的正则化系数
10	model_Path	模型的保存路径

使用 hinge loss 模拟 SVM，按缺省参数训练和验证，只需执行下述命令：
python classify_hw.py --mode hinge

4.2 可视化分类结果

在补全好的代码的基础上，使用 hinge loss 模拟 SVM 以及 libsvm 库对数据集进行分类，保存绘制的 loss 曲线以及特征点分布图（包含特征点、支持向量以及分类边界）。请在报告中比较两种方式的结果。
使用 libsvm 库实现分类的命令为: python classify_hw.py --mode baseline
运行时弹出的显示图片的窗口需要手动关闭，程序才会退出。

4.3 调整正则化系数 C，体会不同的 C 对分类效果的影响

分别设置不同的参数 C=0.0001,0.001,0.01,0.1,1,10，在报告中比较在 C 的不同取值下两种方式在验证集上的分类效果。
调整正则化系数 C 的值可以通过下述命令实现：
python classify_hw.py --mode hinge --C 1.0

5 使用 SVM 算法在 MNIST 数据集上进行分类（5 分）

该部分需要大家补全 process_mnist.py 中的代码，实现 PCA 降维，程序中需要补全代码的地方都有注释提示，请注意阅读。需要补全代码的清单如下：

序号	补全行号	说明
1	36	计算训练集均值
2	38	计算训练集协方差
3	40	SVD 分解协方差矩阵
4	42	压缩训练集维度
5	43	压缩测试集维度

补全好代码后，可以提取特征用于进一步分类，可调整的参数包括：

序号	名称	说明
1	class_0	选取的类别序号，范围 0 至 9
2	class_1	选取的类别序号，范围 0 至 9
3	feat_dim	降维后的特征维数

执行命令：
python process_mnist.py --class_0 0 --class_1 1 --feat_dim 10
可以从原始 MNIST 数据集中提取数字 0 和数字 1 的图片特征，并通过 PCA 算法降到 10 维，并保存在\MNIST 文件夹下。
数据提取成功后，执行命令：
python classify_hw.py --feat_dim 10 --train_file_path 'MNIST/train.npy' --val_file_path 'MNIST/val.npy'
在高维数据集上使用实现的 SVM 算法进行分类。
请在报告中汇报实验参数及结果。

6 撰写作业报告（10 分）

将 HW3 目录和作业报告打包为一个文件（例如 *.zip）提交到网络学堂。作业报告中包括选择题答案，计算题的解题步骤及答案，任务三、四、五运行结果及分析，本次作业遇到的问题及解决方法，对本次作业的意见及建议等。推荐同学们使用随作业发布的 LaTex 模板 HW3-template.zip 完成作业报告。

7 自选课题进度汇报（70 分）*

请已确认自选课题的同学，完成简短的自选课题工作进度汇报，例如，文献阅读、或者研究方案设计、或者原型系统搭建及实验结果等内容。

关于作业迟交的说明：由于平时作业计入总评成绩，希望同学们能按时提

交作业。若有特殊原因不能按时提交，请在提交截止时间之前给本次作业责任助教发 Email 说明情况并给出预计提交作业的时间。对于未能按时说明原因的迟交作业，将酌情扣分。

本次作业责任助教为唐沛 (Email: tp21@mails.tsinghua.edu.cn)。