清华大学电子工程系

媒体与认知 课堂 2

2021-2022 学年春季学期

作业 4

2022 年 4 月 14 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、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 题，已确认自选课题的同学只需完成第 6 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（30 分）
5. 撰写作业报告（10 分）
6. 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。如果尺度因子为 a(a>0)，平移因子为 b($b \in \mathbb{R}$)，则得到的一个小波基函数为：

(A) $\frac{1}{a}\phi(t-b)$

(B) $\frac{1}{a}\phi(\frac{t}{a}-b)$

(C) $\frac{1}{\sqrt{a}}\phi(\frac{t}{a}-b)$

(D) $\frac{1}{\sqrt{a}}\phi(\frac{t-b}{a})$

1.2 关于主成分分析 PCA 和线性判别分析 LDA，以下哪一个说法是正确的：

(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 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

(D) PCA 和 LDA 都是基于高斯假设的非线性特征变换法

1.3 贝叶斯分类器是_____ 模型，支持向量机是_____ 模型。

(A) 生成式; 生成式

(B) 生成式; 鉴别式

(C) 鉴别式; 生成式

(D) 鉴别式; 鉴别式

1.4 SVM 的松弛因子 ξ_i 和正则化系数 C 可以调整模型对训练集上错误的容许程度。C 过小可能导致_____。

(A) 模型欠拟合

(B) 模型过拟合

(C) 对模型无影响

1.5 对于一个核函数 K，其在训练集上的矩阵形式为 K。若 K 是有效核函数，则 K 一定是

(A) 可逆矩阵

(B) 正定矩阵

(C) 对称正定矩阵

(D) 对称半正定矩阵

2 计算题（15 分）

2.1 假设邮件粗略分为垃圾邮件和正常邮件，且存在一种垃圾邮件的检测方法，其中垃圾邮件被正确检测的概率为 a，正常邮件被误判为垃圾邮件的概率为 b。针对某一邮箱，所有邮件中垃圾邮件占的比例为 c，如果某封邮件被判定为垃圾邮件，根据贝叶斯定理，这封邮件是垃圾邮件的概率是多少？(提示：全概率公式 $P(Y) = \sum_{i=1}^{N} P(Y|X_i)P(X_i)$)

2.2 给定样本集合，其均值为 $\mu = [1, 2]^T$，样本协方差矩阵为 C，且已知 $CU = U\lambda$。

其中 $U = \begin{bmatrix} 0.5 & -0.4 \\ 0.5 & 0.4 \end{bmatrix}$，$\lambda = \begin{bmatrix} 10.7 & 0 \\ 0 & 0.4 \end{bmatrix}$。

试用主成分分析 PCA 将样本 $x = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$ 变换至一维。

(提示: 样本数据应减去均值；特征向量应归一化)

2.3 设有两类正态分布的样本集，第一类均值为 $\mu_1 = [1, 0]^T$，第二类均值为 $\mu_2 = [0, -1]^T$。两类样本集的协方差矩阵和出现的先验概率都相等： $\Sigma_1 = \Sigma_2 = \Sigma = \begin{bmatrix} 0.7 & 0.2 \\ 0.2 & 1.2 \end{bmatrix}$，$p(\omega_1) = p(\omega_2)$。**试计算分类界面，并对特征向量** $x = [0.2, 0.5]^T$ **分类。**

编程部分

编程部分包括第 3，4，5 题，选择自选课题的同学请完成第 6 题。

3 实现 hinge loss 模拟支持向量机并运行自动评判程序（30 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
hw4.zip	作业 4 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
\data	存放本次作业所用数据集	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 + L2 norm
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	控制代码运行的模式，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	epoch	训练的总轮数
6	valInterval	每几轮执行一次验证
7	lr	训练的学习率
8	C	引入松弛因子后添加的正则化系数
9	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模式的结果。
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 的不同取值下两种模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 撰写作业报告（10 分）

将 hw4 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 HW4-template.zip 完成作业报告。

6 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曾睿 (Email: zengr21@mails.tsinghua.edu.cn)。