清华大学电子工程系

媒体与认知

2023-2024 学年春季学期

作业 3

2024 年 4 月 29 日

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 $a\phi\left(a(at+b)\right)(a>0)$，则尺度因子和平移因子分别为：

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

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

(C) $\frac{1}{a^2}, -b$

(D) $\frac{1}{a^2}, -\frac{b}{a}$

1.2 关于 PCA 和 LDA，以下哪一个说法是正确的：

(A) PCA 是有监督的降维方法，LDA 是无监督的降维方法

(B) PCA 选择分类性能最好的投影方向，LDA 选取数据投影方差最大的方向

(C) PCA 和 LDA 都是基于高斯假设的线性特征变换法

(D) PCA 降维最多降到 C-1 的维数 (C 为类别数)，LDA 则没有限制

1.3 以下说法正确的是：( )

(A) 贝叶斯分类器选择先验概率最大的类别，可实现最小错误率的判决

(B) 假设先验概率相等，正态分布下的贝叶斯决策将成为线性分类器

(C) K-means 计算相邻的 K 个节点的状态来决定样本状态

(D) 对于不同的初值，K-means 的结果可能不同

1.4 考虑一个线性可分问题，使用支持向量机对一组样本点 $X = \{x_i\}_{i=1}^N$ 做二分类，$Y = \{y_i\}_{i=1}^N$ 为对应样本的类别标签，$y_i \in \{-1, +1\}$。假定求解出的支持向量机的决策函数为 $w^T x + b$，那么样本 x^* 是支持向量的必要条件为：( )

(A) $w^T x^* + b = 0$

(B) $w^T x^* + b > 1$

(C) $y^*(w^T x^* + b) = 0$

(D) $y^*(w^T x^* + b) \leq 1$

1.5 对于一个核函数 K，其在训练集上的矩阵形式为 $\mathbf{K}$。K 是有效核函数的充要条件为：( )

(A) $K = K^T$

(B) $K = K^T$, 且对于 $\forall \mathbf{x}$, 有 $\mathbf{x}^T K \mathbf{x} \geq 0$

(C) $K = K^T$, 且对于 $\forall \mathbf{x}$, 有 $\mathbf{x}^T K \mathbf{x} > 0$

(D) 对于 $\forall \mathbf{x}$, 有 $\mathbf{x}^T K \mathbf{x} \geq 0$

2 计算题（15 分）

2.1 给定两个类别的样本分别为:

$$\omega_1 : \{(3,1),(2,2),(4,3),(3,2)\}$$
$$\omega_2 : \{(1,3),(1,2),(-1,1),(-1,2)\}$$

试利用 LDA，将样本特征维数压缩为一维。

2.2 模型训练通常需要大量的数据，假设某采集的数据集包含 80% 的有效数据和 20% 的无效数据。采用一种算法判断数据是否有效，其中无效数据被成功判别为无效数据的概率为 90%，而有效数据被误判为无效数据的概率为 5%。如果某条数据经过该算法被判别为无效数据，则根据贝叶斯定理，这条数据是无效数据的概率是多少？
(提示：全概率公式 $P(Y) = \sum_{i=1}^{N} P(Y|X_i)P(X_i)$**)**

2.3 设有两类正态分布的样本集，第一类均值为 $\mu_1 = [2,-1]^T$**，第二类均值为** $\mu_2 = [1,1]^T$**。两类样本集的协方差矩阵和出现的先验概率都相等：** $\Sigma_1 = \Sigma_2 = \Sigma = \begin{bmatrix} 4 & 2 \\ 2 & \frac{4}{3} \end{bmatrix}$，
$p(\omega_1) = p(\omega_2)$**。试计算分类界面，并对特征向量** $x = [6,2]^T$ **分类。**

2.4 给定异或的样本集
$D = \left\{ \left((0,0)^T,-1\right), \left((0,1)^T,1\right), \left((1,0)^T,1\right), \left((1,1)^T,-1\right) \right\}$
该样本集是线性不可分的，可采用如下所示的多项式函数 $\phi(\mathbf{x})$ **将样本** $D = \{(\mathbf{x}_n, y_n)\}$ **映射为** $D_\phi = \{(\phi(\mathbf{x}_n), y_n)\}$**，其中** $\phi(\mathbf{x})$ **满足**

$$\phi_1(\mathbf{x}) = 2(x_1 - 0.5)$$
$$\phi_2(\mathbf{x}) = 4(x_1 - 0.5)(x_2 - 0.5)$$

(1) 给出映射后的样本集；

(2) 在映射后的样本集中，设计一个线性 SVM 分类器，给出支持向量及分类界面。

2.5 使用 KMeans 算法对 2 维空间中的 6 个点 $(0,2)$,$(2,0)$,$(2,3)$,$(3,2)$,$(4,0)$,$(5,4)$ 进行聚类，距离函数选择欧氏距离 $d=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$。

(1) 起始聚类中心选择 (0,0) 和 (4,3)，计算聚类中心；

(2) 起始聚类中心选择 (1,4) 和 (3,1)，计算聚类中心。

编程部分

编程部分包括第 3，4 题。在本次任务中，我们首先将使用第 2 次编程作业中的卷积神经网络对特定的 2 类交通标志图像进行特征提取，对提取的 2048 点特征采用 PCA 降维成 2 维样本。其次，基于 Hinge loss 实现 SVM 模型，并利用 SVM 对所降维后的特征进行分类。详细说明请参阅习题课课件。

3 代码补全（30 分）

在本任务中，实现 Hinge loss 模拟支持向量机的代码。

在开始本次作业前，请从第 2 次编程作业中拷贝如下文件/程序到当前目录：

- 第 2 次作业采用的数据集文件 (\data)
- 已经训练好的启用 batch normalization 的模型 (\checkponts\bn)
- 已完成的网络结构定义文件 networks.py

附：本次编程作业也为同学们提供了一个训练好的模型，同学们也可以直接下载: https://cloud.tsinghua.edu.cn/f/7e0cb4914bee40f9a3b5/

本次编程作业的程序清单如下：

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
\data	第 2 次编程作业所用数据集	**需要拷贝到当前目录**
\checkpoints\bn	第 2 次编程作业训练好的模型	**需要拷贝到当前目录**
networks.py	第 2 次编程作业网络结构定义	**需要拷贝到当前目录**
data_preprocess.py	本次作业中的数据预处理程序	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
train_svm.py	SVM 训练程序	需要完成代码
test_svm.py	SVM 测试程序	需要完成代码
datasets.py	自定义实现的 Traffic_Dataset，用于读入本次作业中预处理后的数据集	已完成代码
check.py	自动评判程序	已完成代码

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：

data_preprocess.py 文件中待完成内容：

序号	行号	内容	说明
TODO 1	22-54	完成卷积提取后特征的降维	利用计算好的训练集的均值、PCA 投影矩阵，对经过第 2 次作业中的卷积网络提取为 2048 点特征的训练集、验证集和测试集进行降维，降维后得到 2 维的样本
TODO 2	95-105	实现 PCA 算法	按照 PCA 算法的流程，计算数据样本均值、协方差矩阵，并利用 SVD 分解计算特征值最大的前 2 维特征矢量，组成 PCA 投影矩阵

svm_hw.py 文件中待完成内容：

序号	行号	内容	说明
TODO 1	16-66	完成 LinearFunction 结构	请参照之前的作业，完成线性层的前向计算和反向传播过程
TODO 2	69-107	完成 Hinge 结构	完成 Hinge loss 的前向计算和反向传播过程。关于 Hinge loss 请参照习题课课件
TODO 3	110-139	完成 SVM 模型的定义	完成 SVM 模型中线性层参数 W 和 b 的定义

train_svm.py 文件中待完成内容：

序号	行号	内容	说明
TODO 1	46-62	模型初始化	初始化数据载入器、模型及优化器
TODO 2	71-108	完成 SVM 训练过程	请参照每一步的提示，完成单个 epoch 下的 SVM 训练过程
TODO 3	117-140	完成 SVM 验证过程	请参照每一步的提示，完成单个 epoch 下的 SVM 验证过程
TODO 4	158-160	计算支持向量	请按照提示，计算训练集中支持向量的索引

test_svm.py 文件中待完成内容：

序号	行号	内容	说明
TODO 1	33-49	加载训练好的 SVM 模型	可参照 data_preprocess.py 中加载第 2 次作业训练好的卷积网络的方式，加载训练好的 SVM 模型，并完成数据载入器的初始化
TODO 2	51-74	完成 SVM 测试过程	与 train_svm.py 中的验证过程一致，请参照每一步的提示，完成单个 epoch 下的 SVM 测试过程

4 训练/验证/可视化/比较（30 分）

4.1 程序验证

为了验证 svm_hw.py 下各模块前向传播和反向传播过程的正确性，在补全代码后，可以运行 check.py 进行检查。

运行命令：python check.py
若代码正确，则可以进行后续任务。
注意：本任务测试成功的截图需要附在作业报告中。

4.2 数据预处理

在进行 SVM 的训练之前，需要利用补全好的 data_preprocess.py，读取特定的 2 类交通标志数据集，利用第 2 次编程作业预训练好的卷积网络从图像中提取 2048 点特征，并采用 PCA 降至 2 维，最后保存为相应的数据集文件 (train.pt、val.pt、test.pt)。本次作业建议采用交通标志数据集中的 B 和 C 两类。
运行命令: python data_preprocess.py

当程序运行过程中，会可视化显示降维后的训练集、验证集和测试集的样本点，在图片显示窗口工具栏有保存图片的按钮，可手动保存图片。手动关闭图片窗口后，程序可以继续运行至结束。

注意：需要将降维后的 3 个数据集分别对应的可视化结果 (共 3 张图) 附在作业报告中。

4.3 训练、验证及测试

在完成 train_svm.py 和 test_svm.py 后，可以使用默认参数配置对模型进行训练和测试。在默认参数配置中，正则化系数 C 为 0.001。
训练模型:
(1) 利用 cpu 训练，则执行如下命令：
python train_svm.py --device cpu
(2) 利用 gpu 训练，则执行如下命令 (其中 'cuda:0' 对应第 1 张 GPU 显卡，'cuda:1' 对应第 2 张 GPU 显卡)：
python train_svm.py --device cuda:0
测试模型:
(1) 利用 cpu 测试，则执行如下命令：
python test_svm.py --device cpu
(2) 利用 gpu 测试，则执行如下命令：
python test_svm.py --device cuda:0
注意：请打开命令行终端，在终端中输入上述命令。

当程序运行训练 (train) 过程可视化显示训练阶段 loss 曲线、训练集和验证集上分类准确率、训练集样本和 SVM 决策面、验证集样本和 SVM 决策面

等图片时，在图片显示窗口工具栏有保存图片的按钮，可手动保存图片。手动关闭图片窗口后，程序可以继续运行至结束。

注意：需要将训练阶段 loss 曲线、分类准确率、训练集及验证集可视化图片（共 4 张图片），以及测试准确率的结果（截图或数值）附在作业报告中。

4.4 调整正则化系数 C，分析不同的 C 取值对分类效果的影响

分别设置不同的参数 C=1e-6,1e-3,1，对比不同 C 取值下的训练集及验证集上的准确率曲线、训练集及验证集的可视化图片（共 4 张图片），并记录在测试集上的分类准确率，在报告中分析 C 对分类效果的影响。
在训练过程中，调整正则化系数 C 的值可以通过如下命令实现：
python train_svm.py --C 1e-6

注意：需要将在 C 的不同取值下的各可视化图片（每种 C 取值下 4 张图片），以及测试准确率的记录 (截图或数值) 附在作业报告中，并需要包含对不同 C 取值所产生影响的分析（可以从过拟合/欠拟合的角度出发，例如，C 过大会导致什么？C 过小会导致什么?）。

5 撰写作业报告（10 分）

请同学们将**代码**和**作业报告**（请勿包括数据集文件夹\data 和模型保存文件夹\checkpoints）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 3、4 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。如果同学们使用大模型来完成作业，请在作业中说明，并指出使用大模型过程中出现的问题。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。

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)。