模式识别与计算机视觉：第一次作业

2023 年 3 月 8 日

注意事项

1. 请务必认真阅读所有注意事项。
2. 本作业发布时间 2023.3.9，交作业时间：2023 年 3 月 23 日上午 9:00。此时间之后的提交不再接收，成绩以 0 分计。如确有特殊原因（例如因公出差），请提前向任课教师请假，提交相应证明材料后另行安排；如有紧急医疗需求等不可预知的特殊情况，需事后尽早提交正式医院证明等相关证明材料。
3. 请手写或通过 Word/LaTeX 等软件记录答案，回答尽量简洁，一般每次作业的答案（只要答案，不要抄写题目）不超过 3 页为佳。
4. 手写答案的同学可以用拍照、扫描等方式提交电子版，但应在保证内容清晰可见的前提下尽量减少文件大小。
5. 请在每次作业的开始部分写上姓名、学号、所属院系。缺少信息的，本次作业总分扣除 10 分。请注意：只有在正式选课名单上的同学，作业才会被批改并计算分数。
6. 建议作业完成后、交作业之前自行拍照或扫描并妥善保存，以备特殊情况时使用（例如认为自己已经交作业了，但系统中没有）。
7. 作业提交通过教学立方进行，请务必在教学立方中注册本课程。

1 习题一

教材第一章的习题 1.1。

2 习题二

若 $X \sim \mathcal{N}(0,1)$, 证明以下不等式:

(a). 对于任意 $\epsilon > 0$, 有

$$P(X \geq \epsilon) \leq \frac{1}{2} e^{-\epsilon^2/2}.$$

(b). 对于任意 $\epsilon > 0$, 有

$$P(|X| \geq \epsilon) \leq \min\left\{1, \sqrt{\frac{2}{\pi}} \frac{e^{-\epsilon^2/2}}{\epsilon}\right\}.$$

提示: 对于 $\mathcal{N}(0,1)$ 的概率密度函数 $f(x)$, 有 $f'(x) = -xf(x)$.

3 习题三

一个函数 f 的共轭函数 (conjugate function) 定义为

$$f^*(y) = \sup_{x \in \mathrm{dom}\, f} \left(y^T x - f(x)\right).$$

(a). 证明 $\inf_x f(x) = -f^*(0)$.

(b). 证明对任意 x, y, $f(x) + f^*(y) \geq x^T y$ (对于 $x \notin \mathrm{dom}(f)$ 令 $f(x) = \infty$).

(c). 证明对任意 x, $f^{**}(x) \leq f(x)$, 其中 $f^{**}(x)$ 为 f^* 的共轭函数.

4 习题四

在教材第三章中，我们了解到细节问题（p43）对设计一个模式识别系统的影响。现在我们将探讨如何解决以下细节问题（以教材中的人脸识别为例）

a) 假设存储在设备中的人脸图像是 100×100 的分辨率，即 $\boldsymbol{x} \in \mathbb{R}^{10000}$，而设备将你的照片拍成 400×400。请写出两种不同的预处理方式，使得你的照片能和设备中的照片正常匹配。

b) 我们假设一共有 n 张照片，且将每张存储的照片看作一个 100×100 的矩阵。已知两两不相交的 2×2 的像素格内都具有相似的像素值，如下矩阵示意：

$$\begin{bmatrix} 1 & 1 & 155 & 156 & ... \\ 1 & 1 & 154 & 155 & ... \\ 50 & 51 & 254 & 253 & ... \\ 49 & 50 & 255 & 255 & ... \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{1}$$

你有什么办法能降低存储照片的容量开销吗？存储开销能降低多少？

c) 教材中提到了不平衡二分类问题（p46）。我们假设训练集中：A 类有 9900 个样本，B 类有 100 个样本。测试集中：A 类有 5000 个样本，B 类有 5000 个样本。如果我们学习到一个映射 $f(\cdot)$，它将所有输入的样本都预测为 A 类，那么我们在训练集上的准确率 acc_{train} 是多少？在测试集上的准确率 acc_{test} 是多少？

d) 你可能已经知道计算准确率有两种不同的计算方法：micro 和 macro。请简要描述评价指标计算方法中 micro 和 macro 两种计算方式的区别？在 c) 中我们计算准确率用到的是 micro 还是 macro 的计算方式？如果不了解这两者的区别，请搜索网上资源，自行了解他们的区别。

e) 上述问题实际上描述的是一个长尾识别问题（long-tailed recognition problem）。在这种问题下，我们在训练集上应当采取哪种计算方式来评估准确率？请设计一种针对此问题的训练方法，使得训练集中样本量少的类别 B 能够在测试集上减少误判？此处只需描述主要思路即可，无需提供技术细节。

5 习题五

我们考虑近邻分类器问题。给定一个包含 8 个样本的训练集 $S = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_8\}$，其中 $\boldsymbol{x}_1 = (0, 0)$，$\boldsymbol{x}_2 = (0, 1)$，$\boldsymbol{x}_3 = (0, -1)$，$\boldsymbol{x}_4 = (-1, 0)$，$\boldsymbol{x}_5 = (1, 0)$，$\boldsymbol{x}_6 = (8, 0)$，$\boldsymbol{x}_7 = (8, 1)$，$\boldsymbol{x}_8 = (9, 0)$。它们的类别分别是 (A, A, A, A, A, B, A, B)

a) 对于两个测试样本 $\boldsymbol{z}_1 = (0, -2)$，$\boldsymbol{z}_2 = (8, 2)$，运用最近邻分类器（1-NN），得到这两个样本的分类结果是什么？

b) 同样的两个样本 $\boldsymbol{z}_1, \boldsymbol{z}_2$，运用近邻分类器 k-NN，取 k=3。得到的两个样本的分类结果是什么?

c) 分析两次结果不同的原因?

d) $\boldsymbol{x}_7$ 是否可能属于类别 B? 在此情况下 k-NN 相比 1-NN 的优势在何处?

6 习题六

本题为一道编程题：从零开始构建一个机器学习系统，请参见 'main.ipynb' 文件中的提示来完成相关的代码（请自行安装 Jupyter Notebook）。这份工程的功能包括：

1. 常见的机器学习数据集的读取过程（已提供）
2. 训练和验证集的划分（已提供）
3. 实现一个 KNN 分类器（需完成）
4. 实现评估指标-准确率的计算（需完成）
5. 根据验证集进行超参数选择（需完成）
6. 实现 5 折交叉验证并进行超参数选择（需完成）
7. 最终确定超参数之后，完成在测试集上的测试（需完成）
8. 针对不均衡数据集，实现 precision, recall 和 F1 score 的计算 (需完成)

在完成代码后，提交时需要 notebook 文件（包括代码和中间输出结果，notebook 可直接输出成 pdf 或 html），并谈谈你在这次编程的感想（可以包括你遇到的问题、收获等等）。