模式识别与计算机视觉：第三次作业

April 26, 2023

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 2023.4.26，交作业时间：2023 年 5 月 11 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，只要答案，不要抄写题目。
4. 手写答案的可以拍照、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。请提交后重新检查，确认已经提交成功。
8. 本次提交的作业应当包含作业文件和代码文件，代码文件命名为 `Problem1.py`，`Problem6.py`，并将所有文件打包成单个 ZIP 文件上传。

1 习题一（10 分 =5+5）

在教材的第 5.3 与 5.4 节中，我们已经对 PCA 的推导过程进行了详细的阐述。在第 5.4 节，我们省略了对 PCA 投影到更多维度时其余投影向量的推导。现对其进行分析。

沿用正文中的符号，将从训练集合 X 中计算得到的 x 的协方差矩阵表示为

$$\mathrm{Cov}(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T \,.$$

假设对 $\mathrm{Cov}(x)$ 进行特征分解之后得到的特征向量为 $\xi_1, \xi_2, ..., \xi_D$，对应的特征值为 $\lambda_1, \lambda_2, ..., \lambda_D$ 并有 $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D \geq 0$。通过教材中的公式 (5.24) 可知，在一维 PCA 降维中，任意的原始输入 x_i 可被表示为

$$x_i \approx \bar{x} + \xi_1^T(x_i - \bar{x})\xi_1 \,.$$

在得到了一维 PCA 的降维表示后，现在对二维 PCA 的另一个降维表示 $\xi_2^T(x_i - \bar{x})\xi_2$ 进行推导。

(a) 令 $y_i = x_i - \xi_1^T(x_i - \bar{x})\xi_1$，已知 $\mathrm{Cov}(x) = \sum_{i=1}^{D}\lambda_i\xi_i\xi_i^T$ (教材中公式 (5.25))，请证明：

$$\mathrm{Cov}(y) = \sum_{i=2}^{D}\lambda_i\xi_i\xi_i^T \,.$$

(b) 也就是说，如果我们希望根据算法 5.1 将 y 降低到 1 维空间，得到的 $\mathrm{Cov}(y)$ 最大的特征值应该为 λ_2，其对应的特征向量应该为 ξ_2，这实际上就是原始 x 的二维 PCA 的另一个表示。现在我们希望同学能针对如下的一个输入矩阵 X 对该结论进行验证，并提交相关代码 (请将代码文件命名为 `Problem1.py`)。

$$X = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 0 & 2 \\ 1 & 2 & 9 \\ 3 & 5 & 2 \end{bmatrix} .$$

(注：X 包含 4 个样本，每个样本的输入维度为 3)

2 习题二（15 分 =5+5+5）

在教材第二章的习题 2.8 中，我们已经研究了瑞利商。广义瑞利熵可以看做是瑞利熵的扩展，在这里，我们将进一步探究广义瑞利商的的一系列性质。

给定 S_B 与 S_w 为两个 $n \times n$ 的对称实矩阵，那么若存在 λ 使得方程 $S_B w = \lambda S_w w$，则称 λ 为 S_B 相对于 S_w 的广义特征值，w 为对应的广义特征向量。

现在假设 S_w 正定，那么有排序后的广义特征值为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$，对应的广义特征向量为 $w_1, w_2, \cdots, w_n$。

(a) 求证广义特征向量之间带权正交，即当 $i = j$ 时，$w_i^T S_w w_j = 1$，否则为 0（提示：可以对 S_w 做 Cholesky 分解）。

(b) 求广义瑞利商 $J(w) = \dfrac{w^T S_B w}{w^T S_w w}$ 的最大值和最小值。

(c) 给定 W 为一个 $n \times d$ 的矩阵，其各列向量对应于前面所述的 d 个特征向量 $w_1, w_2, \cdots, w_d$，求 $J = \dfrac{|W^T S_B W|}{|W^T S_w W|}$ 的值。

3 习题三（18 分 =3+3+3+3+3+3）

在 SVM 中，核方法 (kernel method) 使得我们能将数据隐式地映射到一个新的特征空间中，从而把一个非线性分类问题转化为一个等价的线性分类问题。使用核技巧 (kernel trick)，SVM 模型进行预测的过程可以被表示为

$$\begin{aligned} f(\boldsymbol{x}) &= \boldsymbol{w}^T \phi(\boldsymbol{x}) + b && (1) \\ &= \sum_{i=1}^{n} \alpha_i y_i \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}) + b && (2) \\ &= \sum_{i=1}^{n} \alpha_i y_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}) + b \,, && (3) \end{aligned}$$

其中 $\kappa(\cdot,\cdot)$ 就是核函数 (kernel function 的，其表示的实际上就是两个向量在新的特征空间中的内积，也即相似度。核函数最自然的构造方法是显式地定义出映射后的特征，然后根据新的特征反推对应的核函数。但是因为新的特征空间一般是高维的、甚至是无穷维的，因此更常见的情况是根据具体学习问题直接定义出核函数

的形式。然而，并非所有的函数都是合法的核函数，因为合法的核函数需要满足确实存在某一特征映射方式，使得该函数表示的是输入的向量在新的特征空间中的内积，而这可以通过 Mercer 条件进行判断。Mercer 条件告诉我们，一个对称的函数要是合法的核函数，需要满足对于任意的样本集合 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$，从该样本集合定义的矩阵 $K \in \mathbb{R}^n \times \mathbb{R}^n$，其中

$$[K]_{ij} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

且该矩阵 K 是半正定的。关于 Mercer 条件更加具体的描述可以参考教材中对应小节的内容。

在本题中，对于下面给定的不同的 κ，你需要判断它是否是一个合法的核函数，同时给出证明过程。

(a) $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \kappa_1(\boldsymbol{x}, \boldsymbol{y}) + \kappa_2(\boldsymbol{x}, \boldsymbol{y})$，其中 κ_1 和 κ_2 都是定义在 $\mathbb{R}^d \times \mathbb{R}^d$ 上合法的核函数。

(b) $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \kappa_1(\boldsymbol{x}, \boldsymbol{y}) - \kappa_2(\boldsymbol{x}, \boldsymbol{y})$，其中 κ_1 和 κ_2 都是定义在 $\mathbb{R}^d \times \mathbb{R}^d$ 上合法的核函数。

(c) $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \alpha\kappa_1(\boldsymbol{x}, \boldsymbol{y})$，其中 κ_1 是定义在 $\mathbb{R}^d \times \mathbb{R}^d$ 上合法的核函数，$\alpha \in \mathbb{R}^+$ 是一个正实数。

(d) $\kappa(\boldsymbol{x}, \boldsymbol{y}) = -\alpha\kappa_1(\boldsymbol{x}, \boldsymbol{y})$，其中 κ_1 是定义在 $\mathbb{R}^d \times \mathbb{R}^d$ 上合法的核函数，$\alpha \in \mathbb{R}^+$ 是一个正实数。

(e) $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \kappa_1(\boldsymbol{x}, \boldsymbol{y})\kappa_2(\boldsymbol{x}, \boldsymbol{y})$，其中 κ_1 和 κ_2 都是定义在 $\mathbb{R}^d \times \mathbb{R}^d$ 上合法的核函数。

(f) $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \kappa_3(\phi(\boldsymbol{x}), \phi(\boldsymbol{y}))$，其中 $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$，而 κ_3 是定义在 $\mathbb{R}^{d'} \times \mathbb{R}^{d'}$ 上合法的核函数。

4 习题四（12 分 =6+6）

在线性不可分问题下的 SVM（课本公式 7.46–48）当中，对于正样本和负样本，其在目标函数中分类错误或分对但置信度较低的代价是相同的。但是在很多不均衡分布下的应用场景中，比如负样本过少时，往往会出现负样本分类错误（即 false positive）的现象。

现在，针对课本公式 7.46–48，我们希望对负样本分类错误或分对但置信度较低的样本施加 $k > 0$ 倍于正样本中被分错的或者分对但置信度较低的样本的代价。此时：

(a) 请给出相应的 SVM 优化问题。

(b) 请推导出相应的对偶问题及 KKT 条件。

5 习题五（15 分 =5+10）

朴素贝叶斯是一种适用于分类的有监督学习算法。请查阅相关资料，并根据你的理解完成此题。

(a) 朴素贝叶斯所提出的基本假设是什么？这种假设带来了什么方便与局限？经典的朴素贝叶斯是参数化还是非参数化的？

(b) 高斯朴素贝叶斯算法是一种基于贝叶斯定理和特征条件独立性假设的分类方法。对于连续的数据，它假定每个类别的各个特征都服从高斯分布。给定以下三个类别和对应的训练样本：
类别 A：[(1,2),(2,3),(3,4),(4,5)]
类别 B：[(1,4),(2,5),(3,6),(4,7)]
类别 C：[(4,1),(5,2),(6,3),(7,4)]
请使用高斯朴素贝叶斯算法，对以下数据进行分类：(2,2), (6,1)，并写出详细过程。

6 习题六（30 分 =5+5+5+5+5+5）

数据压缩是一种实用技术，常用于图像压缩、视频压缩、音频压缩中。数据压缩通常分为有损压缩和无损压缩。有损压缩指在压缩过程中存在信息损失，无法还原原始数据。因其有较高的压缩比，故有损压缩的应用更加广泛。下面是一个有损压缩模型。

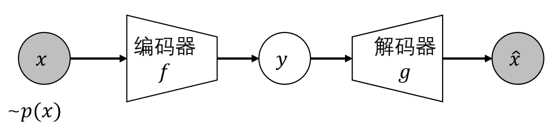

Figure 1: 有损压缩模型

通常假定原始数据分布为 $p(x)$，该分布是未知的。当我们从分布中采样一个 x，根据编码器 f 得到 y，再根据解码器 g 得到 $\hat{x}$。这里的 y 称为 x 的一个表示，$\hat{x}$ 称为 x 的重构。在有损压缩中，x 和 $\hat{x}$ 通常存在差异。在实际的压缩系统中，x 可能是实数、向量、矩阵等，y 可能是整数、二进制数等。这里我们首先考虑 x 是实数，y 是整数的情况。

(a) 首先我们举一个例子来帮助理解有损压缩模型。假设原始数据分布为一个离散分布：

$$P(x=1)=0.5, P(x=2)=0.25, P(x=3)=0.25.$$

编码器为：

$$f(x)=\begin{cases}0, & 若x=1,\\ 1, & 否则.\end{cases}$$

表示 y 的取值范围为 $\{0,1\}$，分布为：

$$P(y=0)=0.5, P(y=1)=0.5$$

解码器为：

$$g(y)=\begin{cases}1, 若y=0,\\ 2, 否则.\end{cases}$$

分别求 x，y，$\hat{x}$ 的信息熵，并证明该系统是有损的。试分析该系统什么情况下会导致信息损失。

(b) **压缩系统的性能**。对于有损压缩，需要考虑两方面的性能指标。其一是重构数据与原始数据的差距，简称为重构误差（D）；其二是表示 y 所产生的编码长短，简称为码率（R）。二者的权衡使用 λ 控制，即最终的性能指标为 $D+\lambda R$。本题中，重构误差用均方误差计算，编码长短用信息熵计算。试写出性能指标的完整公式。当我们希望码率最小时，应该如何设计编码器和解码器。试与问题（a）中的系统对比性能，分析它们分别在 $\lambda=0.1,1,10$ 情况下的优劣。

(c) y **的表达能力**。论证当 $y\in\{0,1\}$ 时，系统无法达到无损压缩。为了增加 y 的表达能力，我们可以让 $y\in\mathbb{Z}$ 或 $y\in\mathbb{R}$，试写出新的性能指标，并从表达能力、编码难度和编解码器设计难度三个方面比较两者的优劣。

(d) **基于机器学习的编解码器**。从上述题目中，我们的编解码器由手工设计。能否自动地学习出编解码器？试分析使用神经网络模型学习编解码器的可行性，注意这里使用拟合连续函数的神经网络，并且令 $y\in\mathbb{R}$。重点分析损失函数应该如何设计，码率应该如何计算和优化。

(e) **解决连续变量无法编码的问题**。连续变量难以编码，需将其转为离散变量。一个可能的解决方案如下图所示。

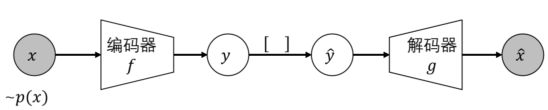

Figure 2: 有损压缩模型，量化表示 y

图中 [ ] 表示取整操作。然而取整操作难以产生有效梯度。加性噪声是一个可能的解决方案。在训练过程中，$\hat{y} = y + \epsilon$，其中 $\epsilon \sim U(-0.5, 0.5)$；在测试过程中，$\hat{y} = [y]$。试分析这样做的合理性（注：本题需要用到两个随机变量的函数分布，是基础概率统计的知识点）。

(f) **编程**。我们使用教材中公式（8.25）的数据，即

$$0.25N(x;0,1) + 0.75N(x;6,4)\,,$$

将其作为原始数据分布，从中抽样 10000 个样本点作为训练集，1000 样本点作为验证集，1000 样本点作为测试集。根据问题（e），尝试编程实现一个基于神经网络的数据压缩系统，完成对原始数据的压缩和重构。请在此处呈现你的实验结果和分析，并将代码文件命名为 `Problem6.py`。