hw1

谭昊童 231300039@smail.nju.edu.cn

2025 年 12 月 1 日

1 (20 points) 聚类

1. 谈谈对聚类的理解，包括其应用场景和难点。（5 分）
2. 针对西瓜数据集 3.0（表 1）, 使用密度和含糖率这两个属性, 计算样本 1 和样本 2 之间的闵可夫斯基距离, 以及样本 1 和样本 3 之间的闵可夫斯基距离. 本题中, 闵可夫斯基距离 (Minkowskidistance) 的参数 p 取值为 1, 2, 3. 观察计算结果, 在不同的 p 的取值下, 这三个样本的相似性关系是否一致?（5 分）
3. 针对表 1, 计算属性“色泽”上三个离散值“青绿”,”乌黑”和“浅白”两两之间的 VDM(Value Difference Metric) 距离,VDM 距离中参数 p 取值为 1, 类别按照数据集的真实类别判定. 验证 VDM 距离满足直递性. （5 分）
4. 按照自己的理解介绍 k 均值聚类算法，并分析其适用场景与存在的不足。（5 分）

Solution. 1. 聚类是一种**无监督学习**任务。聚类算法的目标是将一个数据集中没有标签的样本，按照它们内在的相似性或距离，自动地划分成若干个簇。

理想的聚类结果应该满足两个特性：簇内相似性高; 簇间相似性低。

应用场景：图像分割、异常检测等。

难点：距离度量的选择、簇数的选择、维度诅咒等。

p	1 和 2	1 和 3
1	0.1610	0.2590
2	0.1140	0.2059
3	0.1016	0.1981

2. 结论: 在不同的 p 取值下, 相似性关系 (谁更近) 保持一致。

3.

$$
\begin{aligned}
VDM(\text{青绿, 乌黑}) &= |P(\text{好瓜}|\text{青绿}) - P(\text{好瓜}|\text{乌黑})|^1 + |P(\text{坏瓜}|\text{青绿}) - P(\text{坏瓜}|\text{乌黑})|^1 \\
&= |\frac{3}{6} - \frac{4}{6}| + |\frac{3}{6} - \frac{2}{6}| \\
&= \frac{1}{3} \\
VDM(\text{青绿, 浅白}) &= |P(\text{好瓜}|\text{青绿}) - P(\text{好瓜}|\text{浅白})|^1 + |P(\text{坏瓜}|\text{青绿}) - P(\text{坏瓜}|\text{浅白})|^1 \\
&= |\frac{3}{6} - \frac{1}{5}| + |\frac{3}{6} - \frac{4}{5}| \\
&= \frac{3}{5} \\
VDM(\text{乌黑, 浅白}) &= |P(\text{好瓜}|\text{乌黑}) - P(\text{好瓜}|\text{浅白})|^1 + |P(\text{坏瓜}|\text{乌黑}) - P(\text{坏瓜}|\text{浅白})|^1 \\
&= |\frac{4}{6} - \frac{1}{5}| + |\frac{2}{6} - \frac{4}{5}| \\
&= \frac{14}{15}
\end{aligned}
$$

验证直递性:

$$
\begin{aligned}
VDM(\text{青绿, 乌黑}) + VDM(\text{乌黑, 浅白}) &= \frac{1}{3} + \frac{14}{15} = \frac{19}{15} > \frac{3}{5} = VDM(\text{青绿, 浅白}) \\
VDM(\text{青绿, 浅白}) + VDM(\text{乌黑, 浅白}) &= \frac{3}{5} + \frac{14}{15} = \frac{23}{15} > \frac{1}{3} = VDM(\text{青绿, 乌黑}) \\
VDM(\text{青绿, 乌黑}) + VDM(\text{青绿, 浅白}) &= \frac{1}{3} + \frac{3}{5} = \frac{14}{15} \geq \frac{14}{15} = VDM(\text{乌黑, 浅白})
\end{aligned}
$$

满足直递性。

4. K-Means 是最著名、最简单的聚类算法之一。它的核心思想是“最小化簇内误差平方和 (SSE)”, 它试图找到 K 个中心点（质心），并使得数据集中每个点到其所属簇的质心的距离平方和最小。

   K-Means 的迭代步骤如下：

   (1) 初始化：首先，预先指定要分成的簇数 K。然后，从数据集中随机选择 K 个样本点作为初始的簇质心。

   (2) 分配：对于数据集中的每一个样本点，分别计算它到 K 个质心的距离, 将该样本分配给距离它最近的那个质心所代表的簇。

   (3) 簇质心更新：当所有样本都分配完毕后，取该簇中所有样本点的均值作为新的质心。

   (4) 迭代：重复执行第 2 步和第 3 步，直到满足某个停止条件。

   适用场景:

   - 数据量大且维度较低：K-Means 算法复杂度相对较低，计算效率高，适合处理大规模数据集。
   - 簇的形态为凸形（球状）：K-Means 倾向于发现大小相似、形状近乎球形（或凸形）的簇，并且簇与簇之间有较好的分离度。

- 数值型数据：算法依赖于计算“均值”和“欧氏距离”，因此它天然适用于连续的数值型特征。
- 快速得到初步结果：作为一种简单高效的基准模型，K-Means 常常用于对数据进行快速的初步探索性分析。

存在的不足

- K 值需要预先指定：K-Means 无法自动确定最佳簇数 K，K 值的选择非常依赖经验或辅助方法。
- 对初始质心敏感：随机选择初始质心可能导致算法陷入局部最优解，而不是全局最优解。不同的初始点可能导致完全不同的聚类结果。
- 对异常值（Outliers）敏感：因为 K-Means 使用“均值”来更新质心，少数几个极端异常值就可能严重影响聚类结果。
- 无法处理非球状簇：它无法很好地识别非凸形状的簇（如月牙形、环形）或密度差异很大的簇。

□

2 (20 points) 降维

1. 谈谈对维度灾难的理解，包括其原因以及危害，还有可以从哪些方面缓解维度灾难。(5 分)
2. 本题考察 PCA 相关的线性代数基础知识以及基本操作。给定 d 维空间中 m 个样本构成的矩阵

$$X = [x_1^\top; \ldots; x_m^\top] \in \mathbb{R}^{m \times d},$$

$\hat{X} \in \mathbb{R}^{m \times d}$ 为 X 中心化后得到的矩阵。严格的协方差矩阵具有 $\frac{1}{m-1}$ 因子，由于常数对本题分析结果无影响，所以在本题的讨论中忽略该常数因子。

(1) $\hat{X}^\top \hat{X}$ 和 $\hat{X}\hat{X}^\top$ 为什么是半正定矩阵？二者的特征值有什么联系？受此启发, 请思考当特征维度远大于样本个数时 ($d \gg m$), 使用特征值分解求解 PCA 应如何执行将更加高效? (8 分)

(2) 针对以下样本矩阵 (包含 5 个示例, 每个示例 2 维), 请对其进行主成分分析, 将样本降至一维并写出详细计算过程。(7 分)

$$X^\top = \begin{pmatrix} 3 & 4 & 4 & 6 & 3 \\ 2 & 3 & 2 & 3 & 0 \end{pmatrix}$$

Solution. 1. 维度灾难是指在数据挖掘和机器学习中，当数据的特征维度（即 d）急剧增加时，一系列问题随之产生，导致算法性能急剧下降、计算成本激增的现象。

根本原因：随着维度的增加，数据空间的体积会呈指数级增长（例如 $V \propto L^d$）。为了保持相同的数据密度，我们所需要的样本量 m 也必须呈指数级增长。然而在现实中，样本量（m）的增长速度远远跟不上维度（d）的增长速度。

主要危害：

- 数据稀疏性与近邻失效：在高维空间中，样本点会变得极其稀疏。任意两个样本点之间的距离都倾向于变得很大。这使得基于“邻域”的算法（如 KNN）和基于密度的算法（如 DBSCAN）几乎失效，因为“局部”的概念失去了意义。
- 距离度量失去意义：在高维空间中，任意一点到其他所有点的“最近距离”和“最远距离”之间的相对差距会趋向于 0，所有基于距离度量的算法（如 K-Means 聚类、KNN）的性能都会严重退化。
- 计算复杂度剧增：算法的计算和存储成本通常与维度 d 相关（例如 $O(d), O(d^2)$）。高维度会使算法变得极其缓慢，甚至在计算上不可行。
- 过拟合风险：当特征维度 d 远远大于样本数 m 时（$d \gg m$），模型有太多的“自由度”。这使得模型很容易背下训练数据中的噪声，而不是学习到底层的数据规律，导致模型在训练集上表现很好，但在测试集上表现很差。

缓解方法：

(a) 特征选择：通过选择最相关的特征子集，减少冗余和无关特征，从而降低维度。

(b) 特征提取/ 降维：通过将原始高维特征投影到一个新的低维空间中，生成全新的、数量更少的特征。

2. (1) **证明半正定矩阵：**设 $v \in \mathbb{R}^m$ 为任意非零向量，则

$$\begin{aligned} v^\top(\hat{X}^\top\hat{X})v &= (v^\top\hat{X}^\top)(\hat{X}v) \\ &= (\hat{X}v)^\top(\hat{X}v) \\ &= \sum_{i=1}^{d}(\hat{X}v)_i^2 \geq 0 \end{aligned}$$

因此 $\hat{X}^\top\hat{X}$ 是半正定矩阵。同理可证 $\hat{X}\hat{X}^\top$ 也是半正定矩阵。

二者有完全相同的非零特征值。设 $\lambda \neq 0$ 为 $\hat{X}^\top\hat{X}$ 的特征值，v 为任意非零向量，则有

$$(\hat{X}^\top\hat{X})v = \lambda v$$

两边左乘 $\hat{X}$ 得

$$\hat{X}(\hat{X}^\top\hat{X})v = \lambda\hat{X}v$$

即

$$(\hat{X}\hat{X}^\top)(\hat{X}v) = \lambda(\hat{X}v)$$

因为 $\lambda \neq 0, v \neq 0$，所以 $\hat{X}v \neq 0$，设 $u = \hat{X}v$，则 $u \neq 0$ 且

$$(\hat{X}\hat{X}^\top)u = \lambda u$$

因此 λ 也是 $\hat{X}\hat{X}^\top$ 的特征值。

如果要对 $d \times d$ 的协方差矩阵 $\hat{X}^\top\hat{X}$ 进行特征值分解，先对 $m \times m$ 的矩阵 $\hat{X}\hat{X}^\top$ 进行特征值分解，得到非零特征值 λ_i 和对应的特征向量 u_i.

设 $v_i = \hat{X}^\top u_i$，

$$(\hat{X}^\top\hat{X})v_i = \hat{X}^\top(\hat{X}\hat{X^\top}u_i) = \hat{X}^\top\lambda_i u_i = \lambda_i v_i$$

, 所以 v_i 是 $\hat{X}^\top\hat{X}$ 的特征向量。

(2) 对 X 进行中心化, 将 X 的每一行减去均值 $\mu^\top = (4, 2)$：

$$\hat{X} = X - \mu = \begin{pmatrix} 3-4 & 2-2 \\ 4-4 & 3-2 \\ 4-4 & 2-2 \\ 6-4 & 3-2 \\ 3-4 & 0-2 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 2 & 1 \\ -1 & -2 \end{pmatrix}$$

计算

$$\hat{X}^\top \hat{X} = \begin{pmatrix} -1 & 0 & 0 & 2 & -1 \\ 0 & 1 & 0 & 1 & -2 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 2 & 1 \\ -1 & -2 \end{pmatrix} = \begin{pmatrix} 6 & 4 \\ 4 & 6 \end{pmatrix}$$

解方程 $\det(\hat{X}^\top \hat{X} - \lambda I) = 0$，得到特征值 $\lambda_1 = 10, \lambda_2 = 2$.
因为题目要求降为一维，所以计算最大特征值 $\lambda_1 = 10$ 对应的特征向量 v_1.
解方程 $(\hat{X}^\top \hat{X} - 10I)v_1 = 0$，

$$\begin{pmatrix} -4 & 4 \\ 4 & -4 \end{pmatrix} \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix} = -4 \begin{pmatrix} v_{11} - v_{12} \\ -v_{11} + v_{12} \end{pmatrix} = 0$$

得到 $v_{11} = v_{12}$，单位化后，$v_1 = \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)^\top$.
将样本投影到主成分方向上，得到降维后的样本：

$$Y = \hat{X} v_1 = \begin{pmatrix} -1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 2 & 1 \\ -1 & -2 \end{pmatrix} \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix} = \begin{pmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \\ 0 \\ \frac{3\sqrt{2}}{2} \\ -\frac{3\sqrt{2}}{2} \end{pmatrix}$$

□

3 (20 points) 特征选择

1. 简述特征选择的目的，以及一些常用的特征选择方法。(5 分)
2. 当样本特征很多, 而样本数相对较少时, 模型很容易陷入过拟合, 为了缓解过拟合问题, 可引入正则化项。请通过公式说明 L1 范数和 L2 范数在正则化中的用法, 以及分析为什么它们可以缓解过拟合问题, 过程中结合画图分析 L1 与 L2 范数的差异。(10 分)
3. 字典学习与压缩感知都有对稀疏性的利用, 请你分析两者对稀疏性利用的异同点。(5 分)

Solution. 1. 目的：

- 缓解维度灾难与过拟合：当特征过多而样本过少时，模型容易过拟合。移除不相关的特征（噪声）有助于提升模型的泛化能力。

- 降低计算和存储成本：更少的特征意味着模型训练更快、推理更快、占用存储空间更少。
- 避免冗余信息：移除高度相关的（冗余的）特征，使模型更简洁高效。

常用方法：

- 过滤法：基于统计指标（如方差、相关系数、卡方检验等）评估每个特征与目标变量的关系，选择排名靠前的特征。
- 包装法：使用特定的机器学习算法作为“黑盒”，通过搜索不同的特征子集来评估模型性能，选择性能最好的子集。
- 嵌入法：在模型训练过程中自动进行特征选择，如决策树中的特征重要性、Lasso 回归中的 L1 正则化等。

2. 假设我们有一个损失函数 $L(\mathbf{w})$.

   L1 正则化目标函数：$J_{L1}(\mathbf{w}) = L(\mathbf{w}) + \lambda \sum_{j=1}^{d} |w_j|$

   L2 正则化目标函数：$J_{L2}(\mathbf{w}) = L(\mathbf{w}) + \lambda \sum_{j=1}^{d} w_j^2$

   为什么可以缓解过拟合：当 $d \gg m$ 时，模型过于复杂，它有足够的能力去“记住”训练数据中的噪声。这通常体现在模型参数（权重）$\mathbf{w}$ 非常大。添加正则项后，最小化目标函数要求模型在拟合数据（最小化 $L(\mathbf{w})$）的同时，也要保持权重的“规模”较小（最小化正则项）。这迫使模型学习到更简单、更平滑的函数，从而减少对训练数据噪声的敏感性，提升泛化能力。

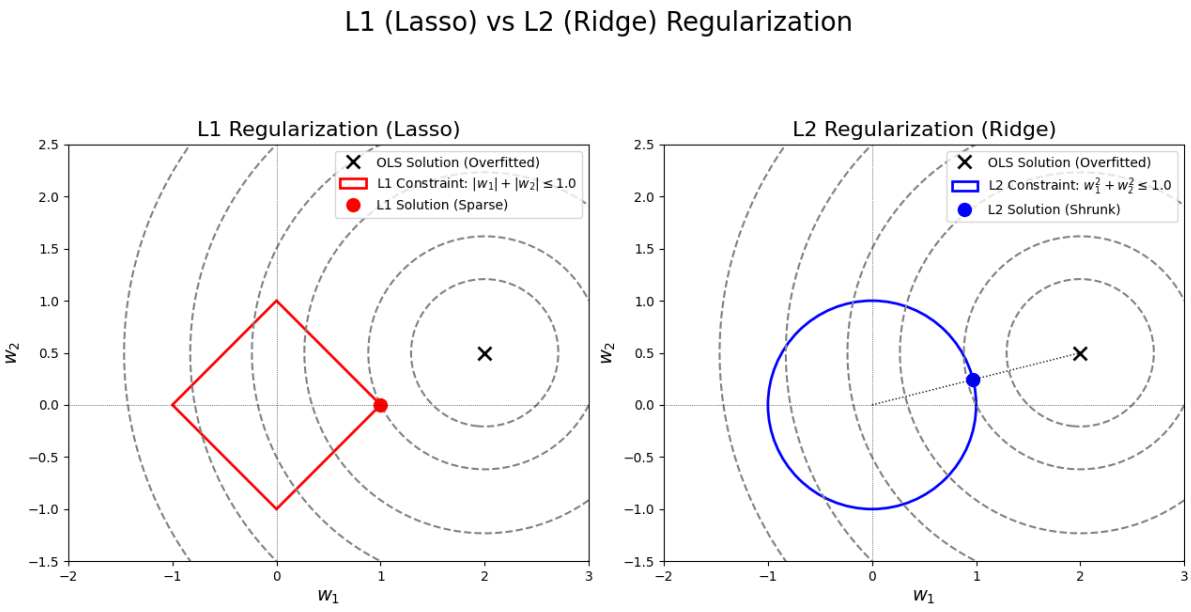

图 1: L1 与 L2 正则化的几何解释

如图，以 $d = 2$ 维特征 w_1, w_2 为例。

- $L(\mathbf{w})$ 等高线（椭圆）：图中的椭圆形等高线代表原始损失函数 $L(\mathbf{w})$。越靠近中心的椭圆，损失越小。中心点 $\hat{\mathbf{w}}_{OLS}$ 是没有正则化时的最优解（即过拟合的解）。
- L1 约束区域（菱形）：L1 范数的约束 $P_{L1}(\mathbf{w}) \leq C$（即 $|w_1| + |w_2| \leq C$）在二维平面上是一个菱形（或旋转的正方形）。

- L2 约束区域（圆形）：L2 范数的约束 $P_{L2}(\mathbf{w}) \leq C$（即 $w_1^2 + w_2^2 \leq C$）在二维平面上是一个圆形。

L1 正则化倾向于使最优解 $\mathbf{w}$ 落在坐标轴上，导致很多特征的权重 w_j 精确地等于 0。L2 正则化会使所有权重 w_j 趋近于 0，但通常不会等于 0。它倾向于将权重“平均”分配给相关的特征。

3. 相同点

   - 两者都建立在“信号是可稀疏表示的”这一核心假设之上。
   - 在数学上，两者都经常依赖 L0 范数（非零元素个数）来度量稀疏性，并由于 L0 的 NP-hard 性质，转而使用其凸松弛 L1 范数 ($\|\alpha\|_1$) 进行优化求解。

   不同点

   - 字典学习的目标是学习一个字典 D, 能让数据 X 得到最稀疏的表示.
   - 压缩感知的目标是重建一个信号 x。稀疏性在这里是一种先验知识。它是信号 x 固有的一个属性，是使得从欠定方程 $y = \Phi x$ 中解出 x 成为可能的前提。

□

4 四. (20 points) 半监督 SVM

考虑一个二分类问题，其中我们有一组标记数据集 $\mathcal{D}_L = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$ 和一组未标记数据集 $\mathcal{D}_U = \{x_{l+1}, x_{l+2}, \ldots, x_m\}$，其中 $y_i \in \{-1, +1\}$ 为标记数据的类别。假设数据集 $\{x_1, x_2, \ldots, x_m\}$ 来自于某个未知的分布，并且 TSVM 的目标是通过最小化带有约束的目标函数来学习分类器。在 TSVM 中，我们同时优化标记数据的分类误差和无标记数据的决策边界，使得无标记数据尽可能被正确分类。

1. 给出 TSVM 的目标函数并且给出 TSVM 的约束条件。(5 分)
2. 使用拉格朗日乘子法，推导 TSVM 的对偶问题。首先写出拉格朗日函数，并根据拉格朗日乘子法推导出对偶问题的目标函数。注意：在推导过程中，假设未标记样本的标签 y_u 暂时固定为常量。(9 分)

   提示：

   - 使用拉格朗日乘子法将约束条件引入目标函数。
   - 对每个约束条件引入拉格朗日乘子，并对拉格朗日函数进行偏导数计算，得到对偶问题。
   - 最终的对偶问题将是一个与原始空间（原始变量）相关的优化问题。
3. 解释为什么当未标记样本的标签也作为优化变量时, TSVM 的优化开销极大? 常见的优化策略是什么? (参考讲义第 13 章第 20 页 pseudocode 回答问题) (6 分)

Solution. 1. 原始目标函数是：

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C_L \sum_{i=1}^{l} \xi_i + C_U \sum_{u=l+1}^{m} \xi_u$$

约束条件 (Constraints) 是：

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l$$

$$y_u(\mathbf{w}^\top \mathbf{x}_u + b) \geq 1 - \xi_u, \quad u = l+1, \ldots, m$$

$$\xi_i \geq 0, \quad i = 1, \ldots, m$$

2. 拉格朗日函数为：

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{m} C_i\xi_i - \sum_{i=1}^{m} \alpha_i[y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^{m} \beta_i\xi_i$$

分别对 $\mathbf{w}, b, \xi_i$ 求偏导:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i = 0 \implies \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{m} \alpha_i y_i = 0 \implies \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C_i - \alpha_i - \beta_i = 0 \implies C_i = \alpha_i + \beta_i$$

代入拉格朗日函数得

$$\frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{m} \alpha_i y_i \mathbf{w}^\top \mathbf{x}_i = \frac{1}{2}\|\mathbf{w}\|^2 - \mathbf{w}^\top \left(\sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \right) = \frac{1}{2}\|\mathbf{w}\|^2 - \mathbf{w}^\top \mathbf{w} = -\frac{1}{2}\|\mathbf{w}\|^2$$

$$\mathcal{L} = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\|\mathbf{w}\|^2 = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^\top \mathbf{x}_j)$$

对偶问题为：

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C_L, \quad i = 1, \ldots, l$$

$$0 \leq \alpha_i \leq C_U, \quad i = l+1, \ldots, m$$

3. 开销大的原因：优化目标函数在 $\mathbf{w}, b, \boldsymbol{\xi}$ 上是凸的（对于固定的 $\hat{\mathbf{y}}_U$），但在 $\hat{\mathbf{y}}_U$ 上是非凸的、离散的。要找到全局最优解，理论上需要尝试所有 2^k 种标签组合，为每种组合求解一个标准的凸 SVM 问题，然后比较 2^k 个结果的原始目标函数值。这种 $O(2^k)$ 的计算复杂度是指数级的，使得该问题成为 NP-Hard 问题。

   优化策略：如 13 章第 20 页的伪代码所示。先仅使用标记数据训练一个初始 SVM 分类器，然后使用该分类器对未标记数据进行预测，得到初始的伪标签。初始化折中参数 $C_u << C_l$. 基于标签和伪标签来优化 SVM，得到 $(\mathbf{w}, b), \xi$.

   对于相反的伪标签 $\hat{y}_i, \hat{y}_j$，如果 $\xi_i > 0$（i 进入间隔或者错分）且 $\xi_j > 0$（j 进入间隔或者错分）且 $\xi_i + \xi_j > 2$（至少有一个被严重错分），则交换它们的伪标签, 并重新求解 SVM 问题，得到新的 $(\mathbf{w}, b), \xi$. 重复上述过程，直到没有伪标签交换为止。

   然后将 C_u 增加一倍，重复上述过程，直到 $C_u \geq C_l$ 为止。

□

5 五. (20 points) EM 算法及其应用

1. 广义 EM 算法 (5 分)

   设 x 为观测数据，z 为潜在变量，θ 为参数。证明在广义 EM 算法中，不完全数据的对数似然函数 $l(\theta; x)$ 是非递减的。

2. 条件混合模型的构建与 EM 算法应用（15 分）

   最简单的构造条件混合模型的方法是直接将非条件混合模型（例如，混合高斯模型）中的密度替换为条件分布。在这个问题中，我们考虑一种简单的线性回归模型的混合模型。假设我们希望使用包含 c 个线性回归模型的混合模型来拟合数据集 $\{(x_i, y_i) : x_i \in \mathbb{R}^p, y_i \in \mathbb{R}, i = 1, \ldots, n\}$ 。每个回归模型有自己的回归系数 $w_k \in \mathbb{R}^p$ 和混合系数 π_k 。为了简化，假设所有回归模型共享方差 σ^2 。设 $\Theta = \{W, \pi, \sigma^2\}$ 为参数集，其中 $W = \{w_k\}, \pi = \{\pi_k\}$ 。类似于高斯混合模型（GMM），我们可以引入一个 c 维的二元潜在变量 z_i ，其中只有一个非零元素。响应变量 y_i 的生成过程可以写成：

   - 从多项分布 $Mult(\pi)$ 中抽取指示变量 z_i;
   - 从高斯分布 $N(w_{z_i}^\top x_i, \sigma^2)$ 中抽取响应变量 y_i，其中 w_{z_i} 表示由 z_i 选中的回归系数。

   接下来，我们将推导出一个 EM 算法来获得这个模型的最大似然估计（MLE）。具体来说

   (1) 推导出不完全数据对数似然函数的下界;

   (2) 推导 E 步, 使用 r_{ik} 作为“责任”变量; $r_{ik} = p(z_{ik} = 1 \mid x_i, y_i, \Theta)$

   (3) 推导 M 步, 给出参数 $\Theta = \{W, \pi, \sigma^2\}$ 的更新规则。

   提示：请参考讲义第 14 章第 46 页 GMM EM 推导的形式与步骤。类比 GMM 中的 latent variable z_i 和 soft assignment (responsibility)，我们将密度 $\mathcal{N}(x|\mu_k, \Sigma_k)$ 替换为条件分布 $\mathcal{N}(y_i \mid w_k^\top x_i, \sigma^2)$。

Solution. 1. 不完全数据的对数似然函数 $l(\theta; x) = \log p(x|\theta)$。引入潜在变量 z 后，我们有 $p(x|\theta) = \frac{p(x,z|\theta)}{p(z|x,\theta)}$。因此，$l(\theta; x) = \log p(x, z|\theta) - \log p(z|x, \theta)$。在给定当前参数 $\theta^{(t)}$ 和观测数据 x 的条件下，我们对上式两边取关于 z 的期望 $E_{z|x,\theta^{(t)}}[\cdot]$：

$$E_{z|x,\theta^{(t)}}[l(\theta; x)] = E_{z|x,\theta^{(t)}}[\log p(x, z|\theta)] - E_{z|x,\theta^{(t)}}[\log p(z|x, \theta)] \tag{1}$$

由于 $l(\theta; x)$ 是关于 z 的常数，其期望就是它自身。

$$l(\theta; x) = E_{z|x,\theta^{(t)}}[l(\theta; x)]$$

定义：

$$Q(\theta, \theta^{(t)}) = E_{z|x,\theta^{(t)}}[\log p(x, z|\theta)]$$
$$H(\theta, \theta^{(t)}) = E_{z|x,\theta^{(t)}}[\log p(z|x, \theta)]$$

式1可化为：

$$l(\theta; x) = Q(\theta, \theta^{(t)}) - H(\theta, \theta^{(t)})$$

要证明 l 非递减，即证明

$$l(\theta^{(t+1)}; x) - l(\theta^{(t)}; x) = [Q(\theta^{(t+1)}, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)})] - [H(\theta^{(t+1)}, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)})] > 0$$

考虑 Q 部分，M 步对 Q 进行最大化，所以 Q 必然是非递减的，

$$Q(\theta^{(t+1)}, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)}) \geq 0$$

考虑 H 部分：

$$H(\theta^{(t+1)}, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)}) = E_{z|x,\theta^{(t)}}[\log p(z|x, \theta^{(t+1)})] - E_{z|x,\theta^{(t)}}[\log p(z|x, \theta^{(t)})]$$

根据吉布斯不等式，对于任意两个概率分布 $P(z)$ 和 $q(z)$，$-KL(P \parallel q) \leq 0$，即 $E_P[\log q(z)] \leq E_P[\log P(z)]$。令 $P(z) = p(z|x, \theta^{(t)})$ 且 $q(z) = p(z|x, \theta^{(t+1)})$，我们得到：

$$E_{z|x,\theta^{(t)}}\left[\log p(z|x, \theta^{(t+1)})\right] \leq E_{z|x,\theta^{(t)}}\left[\log p(z|x, \theta^{(t)})\right]$$

因此，

$$[H(\theta^{(t+1)}, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)})] \leq 0$$

所以

$$l(\theta^{(t+1)}; x) - l(\theta^{(t)}; x) \geq 0$$

2. (1) 对于单个样本来说，似然函数

$$p(y_i|x_i, \Theta) = \sum_{k=1}^{c} p(z_{ik} = 1, y_i|x_i, \Theta) = \sum_{k=1}^{c} p(z_{ik} = 1|\Theta) p(y_i|x_i, z_{ik} = 1, \Theta)$$

根据模型定义，$p(z_{ik} = 1|\Theta) = \pi_k$ 且 $p(y_i|x_i, z_{ik} = 1, \Theta) = \mathcal{N}(y_i|w_k^\top x_i, \sigma^2)$。

$$p(y_i|x_i, \Theta) = \sum_{k=1}^{c} \pi_k \mathcal{N}(y_i|w_k^\top x_i, \sigma^2)$$

所以不完全数据对数似然函数 $l(\Theta)$ 为：

$$l(\Theta) = \sum_{i=1}^{n} \log p(y_i|x_i, \Theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{c} \pi_k \mathcal{N}(y_i|w_k^\top x_i, \sigma^2) \right)$$

完全数据对数似然 $l_c(\Theta) = \log p(Y, Z|X, \Theta) = \log \prod_{i=1}^{n} p(y_i, z_i|x_i, \Theta)$

$$p(y_i, z_i|x_i, \Theta) = \prod_{k=1}^{c} \left[p(z_{ik} = 1|\Theta) p(y_i|x_i, z_{ik} = 1, \Theta) \right]^{z_{ik}} = \prod_{k=1}^{c} \left[\pi_k \mathcal{N}(y_i|w_k^\top x_i, \sigma^2) \right]^{z_{ik}}$$

$$l_c(\Theta) = \sum_{i=1}^{n} \log \left(\prod_{k=1}^{c} \left[\pi_k \mathcal{N}(y_i|w_k^\top x_i, \sigma^2) \right]^{z_{ik}} \right) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \left(\log \pi_k + \log \mathcal{N}(y_i|w_k^\top x_i, \sigma^2) \right)$$

Q 函数是 $l_c(\Theta)$ 在给定 X, Y 和当前参数 $\Theta^{(t)}$ 下对 Z 的期望：

$$Q(\Theta, \Theta^{(t)}) = E_{Z|X,Y,\Theta^{(t)}}[l_c(\Theta)]$$

因为

$$E[z_{ik}] = p(z_{ik} = 1|x_i, y_i, \Theta^{(t)}) \triangleq r_{ik}$$

所以不完全数据对数似然函数的下界 (Q-函数) 为：

$$Q(\Theta, \Theta^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{c} r_{ik} \left(\log \pi_k + \log \mathcal{N}(y_i|w_k^\top x_i, \sigma^2) \right)$$

(2)

$$\begin{aligned} r_{ik} =& p(z_{ik}=1|x_i,y_i,\Theta^{(t)}) \\ =& \frac{p(z_{ik}=1,y_i|x_i,\Theta^{(t)})}{p(y_i|x_i,\Theta^{(t)})} \\ =& \frac{p(z_{ik}=1|\Theta^{(t)})p(y_i|x_i,z_{ik}=1,\Theta^{(t)})}{\sum_{j=1}^{c} p(z_{ij}=1|\Theta^{(t)})p(y_i|x_i,z_{ij}=1,\Theta^{(t)})} \\ =& \frac{\pi_k^{(t)}\mathcal{N}(y_i|(w_k^{(t)})^\top x_i,(\sigma^2)^{(t)})}{\sum_{j=1}^{c}\pi_j^{(t)}\mathcal{N}(y_i|(w_j^{(t)})^\top x_i,(\sigma^2)^{(t)})} \end{aligned}$$

(3)

$$\begin{aligned} Q(\Theta,\Theta^{(t)}) &= \sum_{i=1}^{n}\sum_{k=1}^{c} r_{ik}\log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{c} r_{ik}\log\mathcal{N}(y_i|w_k^\top x_i,\sigma^2) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{c} r_{ik}\log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{c} r_{ik}\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i-w_k^\top x_i)^2}{2\sigma^2}\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{c} r_{ik}\log\pi_k - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{k=1}^{c} r_{ik}(y_i-w_k^\top x_i)^2 \end{aligned}$$

更新 π_k:

$$\mathcal{L}(\pi,\lambda) = \sum_{i=1}^{n}\sum_{k=1}^{c} r_{ik}\log\pi_k + \lambda(1-\sum_{k=1}^{c}\pi_k)$$

对 π_k 求导并令其为 0:

$$\frac{\partial\mathcal{L}}{\partial\pi_k} = \sum_{i=1}^{n}\frac{r_{ik}}{\pi_k} - \lambda = 0 \implies \pi_k = \frac{\sum_{i=1}^{n} r_{ik}}{\lambda}$$

因为

$$\sum_{k=1}^{n}\pi_k = \frac{\sum_{k=1}^{n}\sum_{i=1}^{n} r_{ik}}{\lambda} = \frac{n}{\lambda} = 1$$

所以

$$\boxed{\pi_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} r_{ik}}$$

更新 w_k:

最大化 Q 中关于 w_k 的部分。这等价于最小化 Q 中关于 w_k 的负向部分：

$$\min_{w_k} \quad J(w_k) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{k=1}^{c} r_{ik}(y_i-w_k^\top x_i)^2$$

由于我们是分别为每个 k 更新 w_k，我们只需最小化：

$$\min_{w_k} \quad J(w_k) = \sum_{i=1}^{n} r_{ik}(y_i-w_k^\top x_i)^2$$

对其求梯度 $\nabla_{w_k}J(w_k)$ 并置零：

$$\nabla_{w_k}J(w_k) = \sum_{i=1}^{n} r_{ik}\cdot 2(y_i-w_k^\top x_i)\cdot(-x_i) = 0$$

$$\sum_{i=1}^{n} r_{ik}(w_k^\top x_i)x_i = \sum_{i=1}^{n} r_{ik} y_i x_i$$

$$\boxed{\left(\sum_{i=1}^{n} r_{ik} x_i x_i^\top\right) w_k^{(t+1)} = \sum_{i=1}^{n} r_{ik} x_i y_i}$$

更新 σ^2:

最大化 Q 中关于 σ^2 的部分。使用 M 步更新后的 $w_k^{(t+1)}$：

$$Q_{\sigma^2} = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{k=1}^{c} r_{ik}(y_i - (w_k^{(t+1)})^\top x_i)^2$$

对 σ^2 求导并置零：

$$\frac{\partial Q_{\sigma^2}}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} - (-1)\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\sum_{k=1}^{c} r_{ik}(y_i - (w_k^{(t+1)})^\top x_i)^2 = 0$$

$$\boxed{(\sigma^2)^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} r_{ik}(y_i - (w_k^{(t+1)})^\top x_i)^2}$$

□