姓名：毛九弢　学号：221900175　2024 年 11 月 12 日

一. (30 分) 软间隔 SVM

教材 6.4 节介绍了软间隔概念，用来解决线性不可分情况下的 SVM 问题，同时也可缓解 SVM 训练的过拟合问题。定义松弛变量 $\xi = \{\xi_i\}_{i=1}^m$，其中 $\xi_i > 0$ 表示样本 x_i 对应的间隔约束不满足的程度。软间隔 SVM 问题可以表示为：

$$\max_{w,b} \rho$$

$$\text{s.t. } \frac{y_i(w^\top x_i + b)}{\|w\|_2} \geq \rho, \quad \forall i \in [m].$$

该式显式地表示了分类器的间隔 ρ。基于这种约束形式的表示，可以定义两种形式的软间隔。

- 第一种是绝对软间隔：

$$\frac{y_i({w_i}^\top x_i + b)}{\|w\|_2} \geq \rho - \xi_i.$$

- 第二种是相对软间隔：

$$\frac{y_i({w_i}^\top x_i + b)}{\|w\|_2} \geq \rho(1 - \xi_i).$$

这两种软间隔分别使用 ξ_i 和 $\rho\xi_i$ 衡量错分样本在间隔上的违背程度。在优化问题中加入惩罚项 $C\sum_{i=0}^m \xi_i^p$（其中 $C > 0, p \geq 1$），使得不满足约束的样本数量尽量小（即让 $\xi_i \to 0$）。

问题：

1. **（10 分）**软间隔 SVM 通常采用相对软间隔，写出其原问题的形式（要求不包含 ρ）。
2. **（10 分）**写出采用绝对软间隔的 SVM 原问题（不包含 ρ），并说明为什么一般不使用绝对软间隔来构建 SVM 问题。
3. **（10 分）**写出 $p = 1$ 情况下软间隔 SVM 的对偶问题。

解：

1. **（10 分）**软间隔 SVM 通常采用相对软间隔，写出其原问题的形式（要求不包含 ρ）。

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^n \xi_i$$

$$\text{s.t. } \begin{cases} y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, \end{cases} \quad , i = 1, 2, \ldots, m$$

2. **（10 分）**写出采用绝对软间隔的 SVM 原问题（不包含 ρ），并说明为什么一般不使用绝对软间隔来构建 SVM 问题。

$$\min_{\mathbf{w},b,\xi_i} \quad \min_{i\in[m]} \left(\frac{1}{2}||\mathbf{w}||^2 + C\xi_i\right)$$

$$\text{s.t.} \quad \begin{cases} y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, \end{cases} \quad , i = 1, 2, \ldots, m$$

为什么一般不使用绝对软间隔 SVM：

很明显，绝对软间隔 SVM 的计算量远大于相对软间隔 SVM。虽然绝对软间隔 SVM 在理论上可以提供稀疏解，有助于特征选择，但由于优化难度大、实现复杂、计算代价高，以及在实际应用中未表现出明显优势，因此一般不使用绝对软间隔 SVM。相反，相对软间隔 SVM 具有良好的数学性质，优化问题为光滑的凸函数，易于求解，且在实践中表现出良好的泛化能力和稳定性。因此，L2 正则化的软间隔 SVM 更广泛地应用于实际问题中。

3.（**10 分**）写出 $p = 1$ 情况下软间隔 SVM 的对偶问题。

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

$$\text{s.t.} \quad \begin{cases} \sum_{i=1}^{m} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C, \end{cases} \quad i = 1, 2, \ldots, m.$$

二. (20 分) SVM 编程

设想你正在进行一项客户数据的分类任务，目标是通过支持向量机（SVM）构建一个模型，准确地区分两类客户。以下是你的任务要求：

已知数据集

我们有一个二维数据集，其中包含两个类别的点，数据如下：

数据点编号	x_1	x_2	类别
1	1.0	2.0	1
2	2.0	3.5	1
3	1.5	1.0	1
4	3.0	3.0	1
5	2.0	1.5	1
6	8.0	8.5	2
7	9.0	10.0	2
8	8.5	9.5	2
9	7.0	7.5	2
10	6.5	9.0	2

任务要求

1.（**10 分**）用 Python 训练一个支持向量机分类模型，使用 `scikit-learn` 中的 `SVC` 来分类上表中的数据。要求：

(a) 训练一个非线性核（如 RBF 核）的支持向量机模型。

(b) 输出支持向量，并绘制分类边界。

请给出 SVM 模型训练过程的完整代码以及实验结果的截图。

2. **（10 分）**假设你希望提高模型的泛化能力，请完成以下任务：

通过网格搜索优化 SVM 的**惩罚参数** C 和**核系数** γ 。请尝试 C 取值 $[0.1, 1, 10, 100]$ 和 γ 取值 $[0.1, 1, 10]$，找出最佳参数组合，并在优化后输出训练准确率和支持向量。同时，总结惩罚参数 C 和核系数 γ 是如何影响分类效果和模型的泛化能力的。

提示：网格搜索是一种用于调优模型超参数的简单方法。它会在给定的参数范围内尝试所有可能的参数组合，选择效果最好的组合。

解：代码见 Prob2; 输出见 Prob2/Out

1. **SVM 模型训练过程的完整代码以及实验结果的截图**结果：输出支持向量，并绘制分类边界

支持向量：$[[2, 3.5], [1.5, 1.], [3., 3.], [9., 10.], [7., 7.5]]$

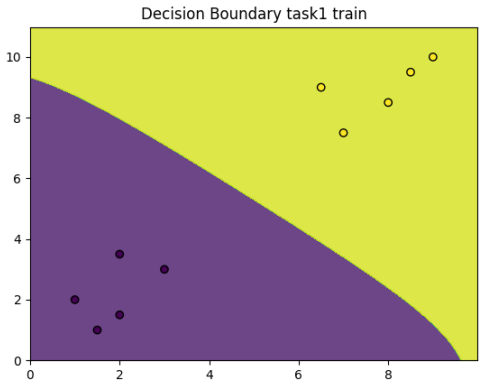

图 1: 分类边界线图

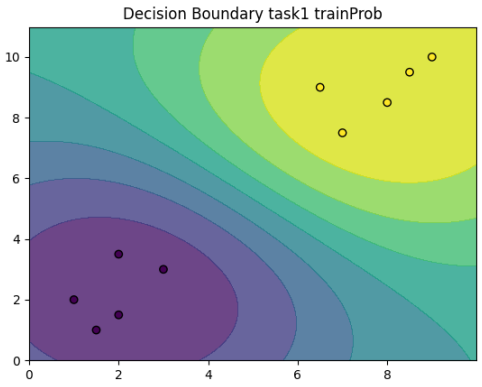

图 2: 概率等高线图

2. **通过网格搜索优化提高模型的泛化能力**我们通过网格搜索优化 SVM 的**惩罚参数** C 和**核系数** γ 。尝试 C 取值 $[0.1, 1, 10, 100]$ 和 γ 取值 $[0.1, 1, 10]$，找出最佳参数组合 $C = 0.1$ 和 $\gamma = 0.1$，输出了优化后训练准确率 1.0000 和支持向量索引 $[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]$。**总结如下：**
惩罚参数 C

- C 是 SVM 中的正则化参数，控制着训练模型对误分类的惩罚力度。
- 较大的 C 值，对误分类的惩罚更大，模型会尽量减少训练错误；模型更复杂，可能会拟合训练数据中的噪声，导致过拟合，泛化能力下降。
- 较小的 C 值，对误分类的容忍度更高，允许部分训练错误；模型更简化，可能欠拟合，但泛化能力可能提高。

核系数 γ

- γ 控制 RBF（径向基函数）核函数的影响范围，决定单个训练样本对决策边界的影响。
- 较大的 γ 值，单个样本的影响范围小，决策边界更灵活；模型复杂度增加，可能紧贴训练数据，容易过拟合。

- 较小的 γ 值，单个样本的影响范围大，决策边界更平滑；模型复杂度降低，可能欠拟合。

综合来看：较小的 C 和 γ 值导致模型简单，决策边界平滑，泛化能力较好。在训练集上准确率高，但存在过拟合的可能性，应结合验证集或测试集评价模型性能。适当的参数选择使模型不过于复杂，避免过拟合，从而在未知数据上保持良好的性能。

三. (20 分) 朴素贝叶斯分类器

一家电商公司希望通过用户评论的关键词来预测评论的情感（**正面**或 **负面**）。假设已经收集了一小部分用户评论，并从中提取出以下五个关键词作为特征：**“good”**、**“bad”**、**“quality”**、**“price”** 和 **“recommend”**。每条评论可以被标记为“正面”或“负面”。

这里假设下方收集到的数据，每个评论仅有一个关键词

评论情感	good 出现次数	bad 出现次数	quality 出现次数	price 出现次数	recommend 出现次数
正面评论	50	5	45	20	60
负面评论	10	30	5	25	2

假设正面评论和负面评论的先验概率分别为 $P(\text{正面评论}) = 0.7$ 和 $P(\text{负面评论}) = 0.3$。

问题

1. **（8 分）**基于上述数据，使用**拉普拉斯修正**（$\alpha = 1$）计算以下条件概率：
   - $P(\text{good}|\text{正面评论})$
   - $P(\text{bad}|\text{正面评论})$
   - $P(\text{quality}|\text{正面评论})$
   - $P(\text{price}|\text{正面评论})$
   - $P(\text{recommend}|\text{正面评论})$

   同时，计算上述特征在负面评论下的条件概率。

2. **（12 分）**假设我们有一条新评论 $X = \{\text{good}, \text{quality}, \text{price}\}$，请使用朴素贝叶斯分类器计算该评论属于**正面评论**和**负面评论**的后验概率 $P(\text{正面评论}|X)$ 和 $P(\text{负面评论}|X)$，并根据计算结果确定该评论的情感类别。

提示：

- 本题的答案请以分式或者小数点后两位的形式给出，比如 P=0.67.
- 在计算条件概率时，请注意考虑拉普拉斯修正后的分母变化。
- 最终的后验概率可以使用贝叶斯公式结合条件独立性假设：

$$P(y|X) = \frac{P(y) \cdot P(X|y)}{P(X)}$$

因为 $P(X)$ 是相同的常数项，比较 $P(y) \cdot P(X|y)$ 即可。

解:

1.（**8 分**）基于上述数据，使用**拉普拉斯修正**（$\alpha = 1$）计算以下条件概率：

已知，正面评论和负面评论的总数分别为 $50 + 5 + 45 + 20 + 60 = 180, 10 + 30 + 5 + 25 + 2 = 72$
故而：

- $P(\text{good}|\text{正面评论}) = \frac{50+1}{180+2} \approx 0.28$
- $P(\text{bad}|\text{正面评论}) = \frac{5+1}{180+2} \approx 0.03$
- $P(\text{quality}|\text{正面评论}) = \frac{45+1}{180+2} \approx 0.25$
- $P(\text{price}|\text{正面评论}) = \frac{20+1}{180+2} \approx 0.12$
- $P(\text{recommend}|\text{正面评论}) = \frac{60+1}{180+2} \approx 0.34$
- $P(\text{good}|\text{负面评论}) == \frac{10+1}{72+2} \approx 0.15$
- $P(\text{bad}|\text{负面评论}) = \frac{30+1}{72+2} \approx 0.42$
- $P(\text{quality}|\text{负面评论}) = \frac{5+1}{72+2} \approx 0.08$
- $P(\text{price}|\text{负面评论}) = \frac{25+1}{72+2} \approx 0.35$
- $P(\text{recommend}|\text{负面评论}) = \frac{2+1}{72+2} \approx 0.04$

2.（**12 分**）假设我们有一条新评论 $X = \{\text{good}, \text{quality}, \text{price}\}$，请使用朴素贝叶斯分类器计算该评论属于**正面评论**和**负面评论**的后验概率 $P(\text{正面评论}|X)$ 和 $P(\text{负面评论}|X)$，并根据计算结果确定该评论的情感类别。

引入拉普拉斯修正后，类的先验概率为：

$$P(\text{good}) = \frac{50 + 10 + 1}{180 + 72 + 5} \approx 0.24$$

$$P(\text{bad}) = \frac{5 + 30 + 1}{180 + 72 + 5} \approx 0.14$$

$$P(\text{quality}) = \frac{45 + 5 + 1}{180 + 72 + 5} \approx 0.20$$

$$P(\text{price}) = \frac{20 + 25 + 1}{180 + 72 + 5} \approx 0.18$$

$$P(\text{recommend}) = \frac{60 + 2 + 1}{180 + 72 + 5} \approx 0.25$$

故而，$P(\text{正面评论}|X)$ 和 $P(\text{负面评论}|X)$ 为：

$$\begin{aligned} P(\text{正面评论}|X) &= \frac{P(\text{正面评论} \cdot P(X|\text{正面评论}))}{P(X)} \\ &= \frac{0.7 \cdot (\prod_{x \in X} P(x|\text{正面评论}))}{\prod_{x \in X} P(x)} \\ &= \frac{0.7 \cdot (0.28 \cdot 0.25 \cdot 0.12)}{0.24 \cdot 0.20 \cdot 0.18} \\ &\approx 0.68 \end{aligned}$$

$$
\begin{aligned}
P(\text{反面评论}|X) &= \frac{P(\text{反面评论} \cdot P(X|\text{反面评论}))}{P(X)} \\
&= \frac{0.3 \cdot (\prod_{x \in X} P(x|\text{反面评论}))}{\prod_{x \in X} P(x)} \\
&= \frac{0.3 \cdot (0.15 \cdot 0.08 \cdot 0.35)}{0.24 \cdot 0.20 \cdot 0.18} \\
&\approx 0.15
\end{aligned}
$$

最终，我们判断，该评论的情感类别为正面评论

四. (30 分) EM 算法

假设有一个包含 6 次硬币抛掷结果的数据集，记录了每次抛掷是否得到“正面”：

$$X = \{\text{正面, 正面, 反面, 正面, 反面, 反面}\}$$

假设这些结果是由两枚硬币 A 和 B 生成的，每次抛掷时选择使用硬币 A 或 B 的概率均为 0.5。然而，具体每次抛掷使用的是哪一枚硬币是未知的。硬币 A 和 B 的正面概率分别为 θ_A 和 θ_B。我们的目标是通过 EM 算法估计这两枚硬币的正面概率 θ_A 和 θ_B。

已知：1. 初始参数：硬币 A 的正面概率 $\theta_A^{(0)} = 0.6$ 和硬币 B 的正面概率 $\theta_B^{(0)} = 0.5$。2. 每次抛掷使用硬币 A 和硬币 B 的概率均为 0.5，即 $P(A) = 0.5$ 和 $P(B) = 0.5$。

请通过一轮 EM 算法的迭代步骤，估计硬币 A 和 B 的正面概率 θ_A 和 θ_B。本题的答案请以分式或者小数点后两位的形式给出，比如 P=0.67.

问题：

1. **E 步**（15 分）：对于每一次抛掷结果，使用当前的参数估计值（$\theta_A^{(0)}$ 和 $\theta_B^{(0)}$），计算该结果由硬币 A 和硬币 B 生成的后验概率，即每次抛掷属于硬币 A 和硬币 B 的“软分配”概率。

   请计算以下内容：

   - 在第 1 次到第 6 次抛掷中，每个结果（正面或反面）由硬币 A 生成的概率 $P(A|x_i)$。
   - 每个结果由硬币 B 生成的概率 $P(B|x_i)$。

2. **M 步**（15 分）：基于 E 步计算出的“软分配”概率，计算硬币 A 和 B 的正面和反面出现的期望次数，并更新硬币 A 和 B 的正面概率 θ_A 和 θ_B。

   请计算以下内容：

   - 硬币 A 的正面和反面期望出现次数，并据此更新硬币 A 的正面概率 $\theta_A^{(1)}$。
   - 硬币 B 的正面和反面期望出现次数，并据此更新硬币 B 的正面概率 $\theta_B^{(1)}$。

解：

根据结果，估计先验：

$$P(\text{正面}) = \frac{3}{6} = 0.5$$

$$P(\text{反面}) = \frac{3}{6} = 0.5$$

1. **E 步**（15 分）：请计算以下内容：在第 1 次到第 6 次抛掷中，1. 每个结果（正面或反面）由硬币 A 生成的概率 $P(A|x_i)$。2. 每个结果由硬币 B 生成的概率 $P(B|x_i)$。

$$P(A|\text{正面}) = \frac{P(\text{正面}|A) \cdot P(A)}{P(\text{正面})} = 0.6 \cdot 0.5/0.5 = 0.6$$

$$P(A|\text{反面}) = \frac{P(\text{反面}|A) \cdot P(A)}{P(\text{反面})} = 0.4 \cdot 0.5/0.5 = 0.4$$

$$P(B|\text{正面}) = \frac{P(\text{正面}|B) \cdot P(B)}{P(\text{正面})} = 0.5 \cdot 0.5/0.5 = 0.5$$

$$P(B|\text{反面}) = \frac{P(\text{反面}|B) \cdot P(B)}{P(\text{反面})} = 0.5 \cdot 0.5/0.5 = 0.5$$

归一化得：

$$P'(A|\text{正面}) = \frac{P(A|\text{正面})}{P(A|\text{正面}) + P(B|\text{正面})} = 0.6/(0.6+0.5) \approx 0.55$$

$$P'(B|\text{正面}) = \frac{P(B|\text{正面})}{P(A|\text{正面}) + P(B|\text{正面})} = 0.5/(0.6+0.5) \approx 0.45$$

$$P'(A|\text{反面}) = \frac{P(A|\text{反面})}{P(A|\text{反面}) + P(B|\text{反面})} = 0.4/(0.4+0.5) \approx 0.44$$

$$P'(B|\text{反面}) = \frac{P(B|\text{反面})}{P(A|\text{反面}) + P(B|\text{反面})} = 0.5/(0.4+0.5) \approx 0.56$$

2. **M 步**（15 分）：请计算以下内容：1. 硬币 A 的正面和反面期望出现次数，并据此更新硬币 A 的正面概率 $\theta_A^{(1)}$。2. 硬币 B 的正面和反面期望出现次数，并据此更新硬币 B 的正面概率 $\theta_B^{(1)}$。

我们使用 H 表示正面，T 表示反面

硬币 A 的正面和反面期望出现次数：

$$\mathbb{E}(AH) = 3 \cdot 0.55 = 1.65$$

$$\mathbb{E}(AT) = 3 \cdot 0.44 = 1.32$$

更新：

$$\theta_A^{(1)} = \frac{\mathbb{E}(AH)}{\mathbb{E}(AH) + \mathbb{E}(AT)} = 0.56$$

硬币 B 的正面和反面期望出现次数：

$$\mathbb{E}(BH) = 3 \cdot 0.45 = 1.35$$

$$\mathbb{E}(BT) = 3 \cdot 0.56 = 1.68$$

更新：

$$\theta_B^{(1)} = \frac{\mathbb{E}(BH)}{\mathbb{E}(BH) + \mathbf{E}(AT)} = 0.45$$