

机器学习

作业三

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：

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

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 $C \sum_{i=1}^m \xi_i^p$ （其中 $C > 0, p \geq 1$ ），使得不满足约束的样本数量尽量小（即让 $\xi_i \rightarrow 0$ ）。

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。

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撑向量。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。

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”。

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。

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 $\theta_B^{(1)}$ 。