# 人工智能与机器学习基础 2025-HW2

TA: 杨睿卿

October 2025
截止日期:2024/11/23 23:59

## 1 主成分分析：PCA

主成分分析（PCA）是一种广泛使用的数据降维技术，它能帮助数据科学家从具有多个变量的复杂数据集中提取关键信息。数据中心化和方差最大化是理解 PCA 核心原理的关键步骤。

假设我们得到一组点 $\{x^{(1)}, ..., x^{(m)}\}$。假设我们像往常一样对数据进行了预处理，使每个坐标的均值为零，方差为单位，我们令处理后的向量为 $\mathbf{X} = (x^{(1)}, ..., x^{(m)}) \in \mathbf{R}^{n\times m}$。。这里，我们希望把数据从 n 维降低到 d 维。

**(a)(10 分)** 在对高维数据降维之前应先进行“中心化”，这里我们是使得每个坐标的均值为零，方差为单位。试推导其矩阵形式的表达式，分析其效果。

**(b)(15 分)** 在课堂上，我们展示了 PCA 可以找到将数据投影到的“方差最大化”方向。在这个问题中，我们发现了 PCA 的另一种解释: 最小化投影点和原始点之间的均方误差。

给定单位向量 $u \in \mathbb{R}^d$，$\|u\| = 1$，令

$$f_u(x) = \arg\min_{\alpha\in\mathbb{R}} \|x - \alpha u\|^2$$

即把 $x$ 投影到方向 $u$ 上。

我们要证明, 最小化投影点和原始点之间的均方误差的单位长度向量 u 对应于数据的第一个主成分。即证明

$$\arg\min_{u:u^Tu=1}\sum_{i=1}^{m}\left\|x^{(i)}-f_u(x^{(i)})\right\|^2$$

为第一个主成分方向（也就是协方差矩阵最大特征值对应的特征向量。）

# 2 聚类算法

## 一、K-means++

在机器学习领域，聚类是一种常见的无监督学习方法，用于对未标记数据进行分组。K-means 算法是最简单、最广泛使用的聚类算法之一，其主要目标是最小化每个类内数据点和其对应中心点之间的距离之和。

然而，K-means 算法的效果很大程度上依赖于初始中心点的选择，这可能导致算法收敛至局部最优解。为了改进这一点，K-means++ 算法被提出，通过一种特定的概率方法选择初始中心，以期实现更优的聚类效果。

其中，K-means++ 算法实现如下:

**算法 1** K-means++ 初始中心选择

**输入：** 数据集 $\mathcal{X}=\{x^{(1)},\ldots,x^{(m)}\}\subset\mathbb{R}^d$，聚类数 $k$

**输出：** 初始中心集合 $\mathcal{C}=\{c_1,\ldots,c_k\}$

```
从 X 中均匀随机选取一个点作为第一个中心 c_1
C ← {c_1}
for j = 2 to k do
    for i = 1 to m do                          ▷ 计算每个点到已选中心的最小平方距离
        D(x^(i)) ← min_{c∈C} ||x^(i) − c||^2
    end for
    构造概率分布 P(x^(i)) = D(x^(i)) / Σ_{p=1}^{m} D(x^(p))
    按 P 从 X 中抽样得到下一个中心 c_j
    C ← C ∪ {c_j}
end for
return C
```

假设我们有一个二维数据集

$$X = \{(1,1),(1,2),(2,1),(2,2),(10,10),(10,11),(11,10),(11,11)\},$$

我们希望将其分为 $k=2$ 类，默认采用欧氏距离。

**(a)(5 分)** 如果初始类中心选择为 $c_1=(1,1)$，$c_2=(10,10)$，请执行 $k$-means 算法，给出迭代过程和最终的类中心。

**(b)(5 分)** 如果初始类中心选择为 $c_1=(1,1)$，$c_2=(2,2)$，请执行 $k$-means 算法，给出迭代过程和最终的类中心。

**(c)(5 分)** 比较两种初始类中心选择，可以感受到不同类中心选择对算法执行的影响, 不同的初始中心，可能导致截然不同的聚类结果。如何解决初始中心选择的问题？请提出至少三条解决措施。

**(d)(5 分)** 现在使用 $k$-means++ 算法，并希望将数据集分为 $k=3$ 类。假设第一个类中心 $c_1$ 已经被选择为 $(1,1)$，请计算每个点被选为第二个类中心 $c_2$ 的概率。

**(e)(5 分)** 如果第二个类中心 $c_2$ 被选择为 $(10,10)$，请计算每个点被选为第三个类中心 $c_3$ 的概率。

## 二、核 K 均值聚类

核 K 均值聚类（Kernel K-means clustering）是一种非线性化的 K 均值聚类算法。它通过使用核函数将数据映射到高维空间，在这个空间中执行传统的 K 均值聚类算法，从而实现非线性聚类。

令 $K(\boldsymbol{x},\boldsymbol{z})$ 为一个核函数，其具有隐式特征映射 $\boldsymbol{\phi}(\mathbf{x})$。我们规定 $K(\boldsymbol{x},\boldsymbol{z}) = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{z})\rangle$, 其中 $\langle\cdot,\cdot\rangle$ 为内积操作。

我们希望在更高维的特征空间 $\boldsymbol{\phi}(\mathbf{x})$ 中进行聚类，而无需显式地计算 $\boldsymbol{\phi}(\mathbf{x})$。

假设我们已经将一组数据 $\{x^{(1)},...,x^{(m)}\}$，通过核函数 $\boldsymbol{\phi}(\mathbf{x})$ 映射到了一个高维特征空间，并打算在这个空间中执行 K 均值聚类。

假设我们设定簇的数量 k=c。记簇中心集合为 $\{\mu_i\}, (i = 1, 2, \ldots, c)$，属于这个簇的数据点集合为 $C_i, (i = 1, 2, \ldots, c)$。

**(a)(5 分)** 请给出在特征空间中，簇中心的求解公式。

**(b)(10 分)** 请给出任意一点 $\phi(\mathbf{x})$, 簇分类更新公式。(注意，在最终算法中，我们特别不希望有单独的 $\boldsymbol{\phi}(\mathbf{x})$ 项或 $\boldsymbol{\mu_i}$ 项，因为它们可能代表一个无限维的对象，因此无法计算。)

# 3 期望最大化：EM 算法

隐马尔科夫模型（Hidden Markov Model，简称 HMM）是比较经典的机器学习模型，它在语言识别，自然语言处理，模式识别等领域得到广泛的应用。

隐马尔可夫模型是基于序列的、包含隐藏状态和观察状态的统计模型，其中隐藏状态无法直接观察，但可以通过观察序列间接地进行推断。

**模型定义**：

- 状态集 $S = \{s_1, s_2, \ldots, s_N\}$ 共 $N$ 个状态。
- 观察集 $V = \{v_1, v_2, \ldots, v_M\}$ 共 $M$ 个可能观察。
- 状态转移概率矩阵 $A = [a_{ij}]$，其中 $a_{ij}$ 表示从状态 $s_i$ 转移到状态 $s_j$ 的概率。
- 观测概率矩阵 $B = [b_j(k)]$，其中 $b_j(k)$ 表示在状态 $s_j$ 下观测到 $v_k$ 的概率。
- 初始状态概率向量 $\pi = [\pi_i]$，其中 $\pi_i$ 表示模型在时间 $t = 1$ 处于状态 $s_i$ 的概率。

在隐马尔可夫模型中，我们的目标是给定观测序列 $O = (o_1, \ldots, o_m)$，最大化模型参数 $\theta = \{A, B, \Pi\}$ 的观察概率 $P(O \mid \theta)$。但是，隐藏状态序列 $Q = (q_1, \ldots, q_n)$ 是不可见的，因此，HMM 的最大似然学习问题正好符合 EM 的使用场景。

假设我们有一个隐马尔可夫模型 (Hidden Markov Model, HMM)，其观察值序列由气温读数组成：高温，低温，高温，低温；模型的隐藏状态为 *Sunny* 和 *Rainy*。模型的参数设置如下：

**初始状态概率：** $\pi(\text{Sunny}) = 0.4, \pi(\text{Rainy}) = 0.6$

**状态转移概率：**

| | *To Sunny* | *To Rainy* |
|---|---|---|
| *Sunny* | 0.6 | 0.4 |
| *Rainy* | 0.3 | 0.7 |

表 1: 晴天和雨天状态的转移概率

**观察概率：**

| | *High Temperature* | *Low Temperature* |
|---|---|---|
| *Sunny* | 0.8 | 0.2 |
| *Rainy* | 0.3 | 0.7 |

表 2: 晴天与雨天状态下高温、低温两种观测结果的观察概率

**(a)(10 分)** 请使用 EM 算法在第一次迭代中针对上述模型参数进行估计。详细描述 EM 算法中的 E 步骤（期望步骤）和 M 步骤（最大化步骤），并进行相应计算。

**(b)(20 分)** 对于模型参数的估计，EM 算法扮演了关键的角色。请证明 EM 算法在 HMM 中的收敛性 (请用上述**模型定义**中提到的符号，一些符号可能需要你自己定义)。

(提示: 我们 EM 算法依旧通过最大似然估计法来计算概率模型的参数，即我们的目标是证明对数似然函数 $l(\theta) = \log P(O \mid \theta)$ 的收敛性, 即证明其单调且有上界)

**(c)(5 分)** 我们已经证明了 HMM 问题上 EM 算法的收敛性，请解释为什么它在某些情况下只能收敛到局部最优，而不是全局最优？并且给出你解决方案的建议（一条即可）