深度学习与自然语言处理第二次作业

李家哲 ZY2203105
lijiazhebuaa@buaa.edu.cn

作业要求

请使用上面链接中的代码身高数据，需要使用 EM 算法来估计高斯混合模型的参数，并使用这些参数来进行预测。你需要对模型进行评估，并解释模型的性能。

主要方法

M1：高斯混合模型

高斯混合模型（Gaussian Mixture Model）可以看作是由 K 个单高斯模型组合而成的模型，这 K 个子模型是混合模型的隐变量（Hidden variable）。一般来说，一个混合模型可以使用任何概率分布，这里使用高斯混合模型是因为高斯分布具备很好的数学性质以及良好的计算性能。

高斯混合模型可对样本数据进行分类，下图所示

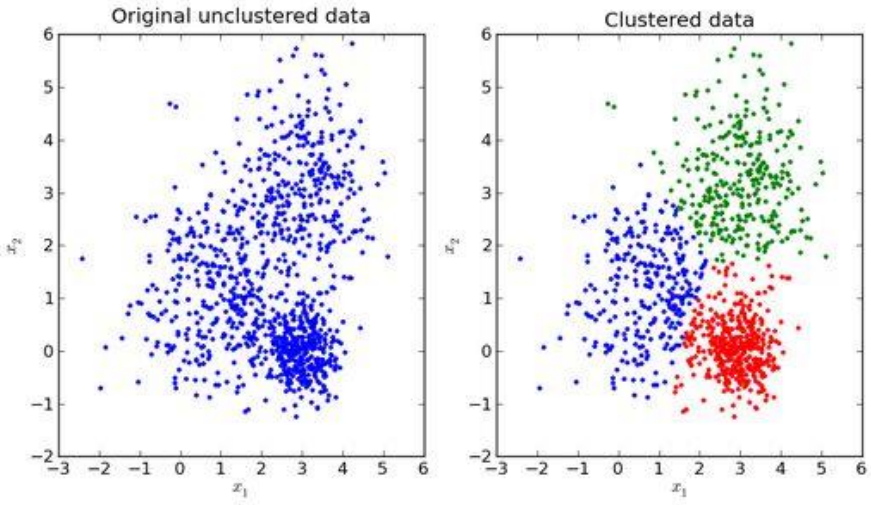


图 1 高斯混合模型中各子模型

定义如下信息：

- x_j 表示第 j 个观测数据，$j=1,2,...,N$
- k 是混合模型中子高斯模型的数量，$k=1,2,...,K$
- α_k 是观测数据属于第 k 个子模型的概率，$\alpha_k \geq 0, \sum_{k=1}^{K} \alpha_k = 1$

- $\phi\left(x|\theta_k\right)$ 是第 k 个子模型的高斯分布密度函数，$\theta_k=\left(\mu_k,\sigma_k{}^2\right)$。其展开形式为

$$\phi\left(x|\theta_k\right)=\frac{1}{\sqrt{2\pi\sigma_k{}^2}}\exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k{}^2}\right)$$

- γ_{jk} 表示第 j 个观测数据属于第 k 个子模型的概率

高斯混合模型的概率分布为：

$$P(x|\theta_k)=\sum_{k=1}^{K}\alpha_k\phi\left(x|\theta_k\right)$$

对于这个模型而言，参数 $\theta_k=\left(\tilde{\mu}_k,\tilde{\sigma}_k,\tilde{\alpha}_k\right)$，也就是每个子模型的期望、方差（或协方差）、在混合模型中发生的概率。

M2: EM 算法在 GMM 模型中应用

EM 算法是一种迭代优化策略，由于它的计算方法中每一次迭代都分两步，其中一个为期望步(E 步)，另一个为极大步(M 步)，所以算法被称为 EM 算法(Expectation-Maximization Algorithm)。EM 算法受到缺失思想影响，最初是为了解决数据缺失情况下的参数估计问题，其基本思想是：首先根据己经给出的观测数据，估计出模型参数的值；然后再依据上一步估计出的参数值估计缺失数据的值，再根据估计出的缺失数据加上之前己经观测到的数据重新再对参数值进行估计，然后反复迭代，直至最后收敛，迭代结束。

每次迭代包含两个步骤：

1. E-step：求期望 $E\left(\gamma_{jk}|X,\theta\right)$ 对于 $j=1,2,...,N$

2. M-step：求极大，计算新一轮迭代的模型参数

对于高斯混合模型来说，我们有样本数据 $x_1,x_2,...,x_N$ 和一个有 K 个子模型的高斯混合模型，想要推算出这个高斯混合模型的最佳参数。其具体流程为：

1. 首先初始化参数

2. E-step：依据当前参数，计算每个数据 j 来自子模型 k 的可能性

$$\gamma_{jk}=\frac{\alpha_k\phi\left(x_j|\ \theta_k\right)}{\sum_{k=1}^{K}\alpha_k\phi\left(x_j|\ \theta_k\right)},\ j=1,2,\ldots,N;k=1,2,\ldots,K$$

3. M-step：计算新一轮迭代的模型参数

$$\mu_k=\frac{\sum_{j}^{N}\left(\gamma_{jk}x_j\right)}{\sum_{j}^{N}\gamma_{jk}},k=1,2,\ldots,K$$

$$\sigma_k = \frac{\sum_j^N \gamma_{jk}\left(x_j - \mu_k\right)^2}{\sum_j^N \gamma_{jk}}, k = 1,2,\ldots,K$$

$$\alpha_k = \frac{\sum_{j=1}^N \gamma_{jk}}{N}, k = 1,2,\ldots,K$$

4. 重复计算 E-step 和 M-step 直至收敛，即$\left\|\log L(\theta_{i+1}) - \log L(\theta_i)\right\| < \varepsilon$，$\varepsilon$是一个很小的正数，表示经过一次迭代之后参数变化非常小)。高斯混合模型的对数似然函数为

$$\log L(\theta) = \sum_{j=1}^N \log P\left(x_j \middle| \theta\right) = \sum_{j=1}^N \log\left(\sum_{k=1}^K \alpha_k \phi\left(x \middle| \theta_k\right)\right)$$

至此，我们就找到了高斯混合模型的参数。需要注意的是，EM 算法具备收敛性，但并不保证找到全局最大值，有可能找到局部最大值。解决方法是初始化几次不同的参数进行迭代，取结果最好的那次。

M3：算法流程

step1：读取样本数据

所用的样本数据为男女生的 2000 个身高数值，首先读取“height_data.csv”文件中的样本数据。

```
def data_processing_csv(file_path):
"""
获取待处理的文件数据
:param file_path: csv 文件名对应路径
:return data: 样本数据
"""
data = []
with open(file_path, 'r') as f:
    for i in f.readlines():
        if i != "height\n":
            data.append(eval(i.strip()))
return data
```

---

step2：运行 em 算法求解高斯混合模型参数

由男女两种性别，设定子模型个数为 2，并给$\left(\mu_k, \sigma_k, \alpha_k\right)$随机赋初值，其初值与算法

运行效果密切相关，具体分析见“实验结果”一节。

根据 M2 中的迭代步骤，迭代进行求解，设定最大迭代次数为 1000，求解精度为 $\varepsilon = 10^{-6}$ 。

```
    def gmm_em(data, k, alpha, mu, sigma, tol, max_step):
    """
    用 em 算法求解高斯混合模型参数
    :param data: 样本数据
    :param k: 混合模型个数
    :param alpha: 第 k 个模型在混合模型中发生的概率(k 维)
    :param mu: 第 k 个模型的期望(k 维)
    :param sigma: 第 k 个模型的标准差(k 维)
    :param tol: 参数变化小于该值时代表收敛，算法结束
    :param max_step: 最大循环求解次数
    :return [alpha, mu, sigma]: 高斯混合模型参数[每个子模型的期望,标准差,在混合模型中发生的概率]
    """
    num = len(data)
    likelihood_old = 0
    data_2d = np.zeros((num, k))
    for i in range(num):
        data_2d[i, :] = [data[i]] * k
    for i in range(max_step):
        # E 步
        gamma = np.zeros((num, k))  # 每个数据来自各个子模型的概率
        for j in range(k):
            gamma[:, j] = alpha[j] * norm.pdf(data, mu[j], sigma[j])
        gamma = gamma / np.sum(gamma, axis=1, keepdims=True)  # 求解每个数据来自各个子模型的概率

        # M 步
        mu = np.sum(gamma * data_2d, axis=0) / np.sum(gamma, axis=0)  # 更新期望
        sigma = np.sqrt(np.sum(gamma * (data_2d - mu) ** 2, axis=0) / np.sum(gamma, axis=0))  # 更新标准差
        alpha = np.sum(gamma, axis=0) / num  # 更新子模型在混合模型中发生的概率

        # 计算对数似然
        likelihood_new = sum(np.log(sum(alpha[j] * norm.pdf(data, mu[j], sigma[j]) for j in range(k))))

        # 检查收敛性
```

```
        if np.abs(likelihood_new - likelihood_old) < tol:
            break
        likelihood_old = likelihood_new

    return [alpha, mu, sigma]
```

---

step3：预测与评估

根据迭代求解的高斯混合模型参数，对样本中每一个数据进行分类，辨别其性别为男或女，求解得到男女人数、性别比例等结果，并与原始数据进行比较，评估模型准确率。

```
    # 预测每个数据点的聚类(0 为女生，1 为男生)
result = np.argmax(np.array([alpha[j] * norm.pdf(data, mu[j],
sigma[j]) for j in range(k)]).T, axis=1)
    # 评估模型准确率
gender = [0] * 500 + [1] * 1500
accuracy = np.mean(result == gender)
```

实验结果

选取不同的初始参数，对模型进行迭代求解，以测试算法对初始参数的敏感性。基于对样本数据的分析，可将身高以性别为标签分为两类。以普遍规律来看，将身高均值较小的子模型看作女生模型，将身高均值较大的子模型看作男生模型。

考虑如表 1 所示的六种情况。case1 为正常情况，即在求解前对样本数据有较好的总体认知；case2 和 case3 测试在标准差取较大或较小值时，模型的拟合结果；case4 测试模型比例对模型拟合的影响；case5 和 case6 测试样本均值对求解的影响。

在不同初始值下利用 em 算法对高斯混合模型的参数进行迭代求解，得到估计的模型参数$\left(\tilde{\mu}_k,\tilde{\sigma}_k,\tilde{\alpha}_k\right)$后，对样本中的每个数据求解其来自男/女模型的概率$\gamma_{jk}$，将概率最大的子模型作为分类结果。分类完成后，将预测分类结果与真实值进行对比，求解模型准确率。

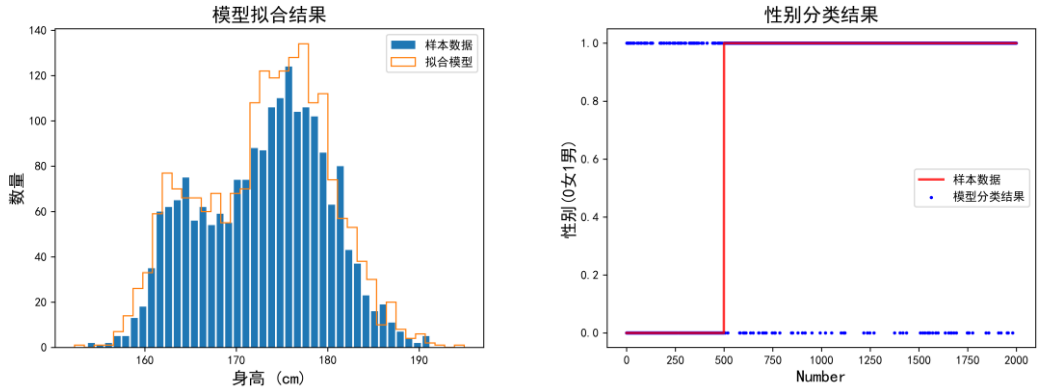
不同初始值下的迭代求解结果如表 2 所示，不同情况下的模型拟合结果和性别分类结果如图 2 所示。对于 case1，em 算法可以迅速收敛到真实值，高斯混合模型的拟合参数与真实值之间误差很小，其预测的性别分类也较为准确，准确率高达 92.95%，具有良好的性能。对于 case2 和 case3，在初始标准差取很大时，算法也可能时拟合参数收敛到真实值，如图 2(b)，但当标准差取很小时，算法失效。对于 case4，在子模型比例相差很大时，算法依旧可以收敛。对于 case5 和 case6，若是对样本数据总体认知较差，选取的样本均值与真实值相差较大，则拟合的参数或是收敛到错误的结果上，或是算法失效，如图 2(e)和(f)所示。

表 1 不同初始值

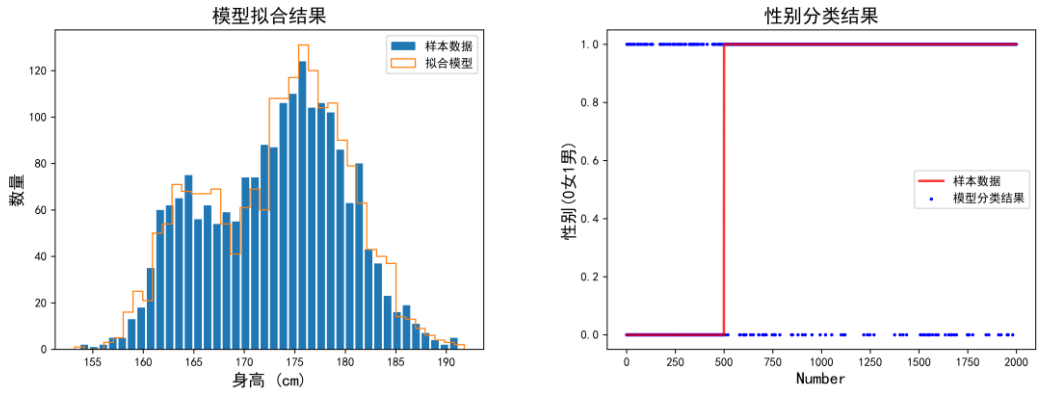
	女生模型			男生模型		
	α_1	μ_1	σ_1	α_2	μ_2	σ_2
case1	0.4	165	1	0.6	180	2
case2	0.5	155	20	0.5	185	10
case3	0.3	160	0.4	0.7	175	0.4
case4	0.05	160	2	0.95	180	2
case5	0.8	150	3	0.2	200	1
case6	0.5	50	1	0.5	100	1

表 2 不同初始值下的迭代求解结果

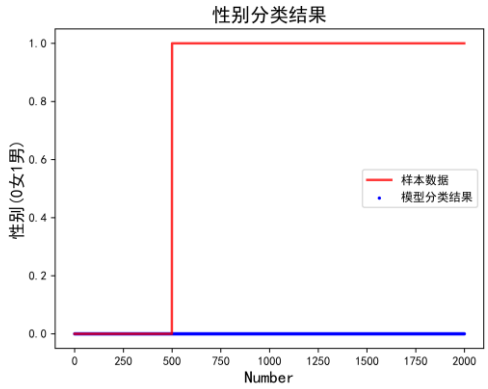
	女生模型			男生模型			准确率
	$\tilde{\alpha}_1$	$\tilde{\mu}_1$	$\tilde{\sigma}_1$	$\tilde{\alpha}_2$	$\tilde{\mu}_2$	$\tilde{\sigma}_2$	
真实值	0.25	164	3	0.75	176	5	—
case1	0.23	163.7	2.84	0.77	175.75	5.09	0.9295
case2	0.23	163.7	2.84	0.77	175.75	5.09	0.9295
case3	nan	nan	nan	nan	nan	nan	0.25
case4	0.23	163.7	2.84	0.77	175.75	5.09	0.9295
case5	0.98	172.77	6.86	0.02	181.21	0.37	0.25
case6	nan	nan	nan	nan	nan	nan	0.25



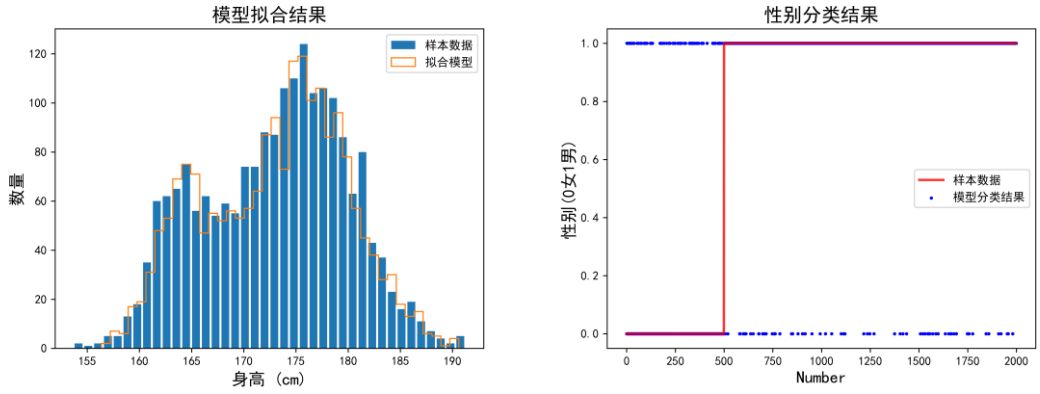
(a) case1



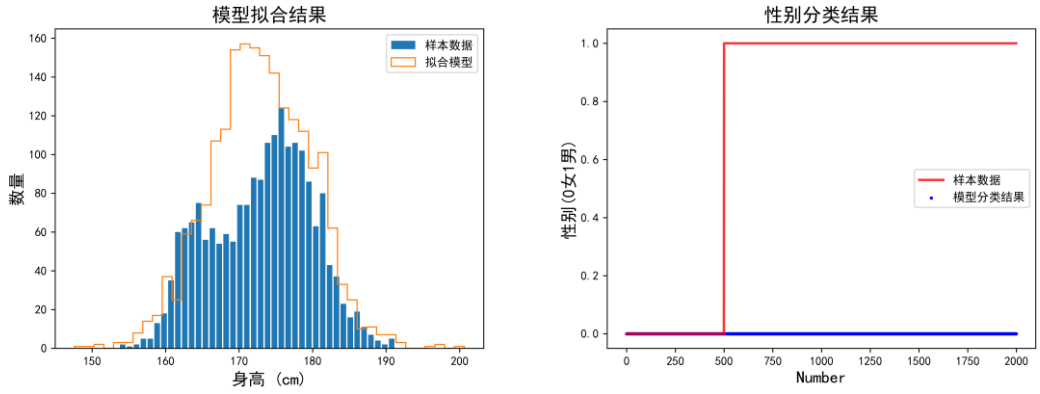
(b) case2



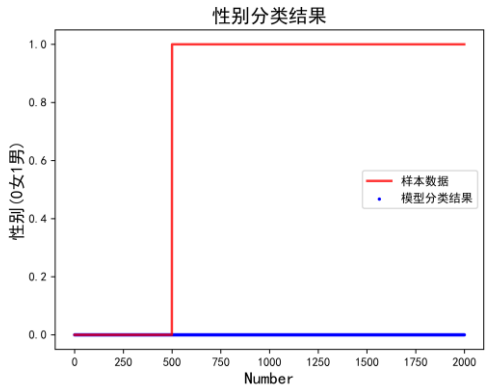
(c) case3



(d) case4



(e) case5



(f) case6

图 2 不同情况下的模型拟合结果和性别分类结果

结论

本文首先给出了作业要求，然后给出了主要理论知识和算法流程，针对身高样本数据，考虑六种不同的初始参数，编写了基于 em 算法的高速混合模型参数拟合算法，并利用拟合的参数对样本的性别进行了分类。

由实验结果可以看出，em 算法对初始值较为敏感。在高斯混合模型的应用中，对样本均值的敏感程度较高，在具体使用中，需要对样本数据有总体的认知，当选取的初始均值与真实值的差距在合理的范围内时，算法可以以较高的精度收敛；对于样本标准差而言，一般情况下可在 1~10 范围内进行选择；算法对于子模型比例较为不敏感。

参考文献

[1] 戴文亮，高斯混合模型（GMM），https://zhuanlan.zhihu.com/p/30483076