多元统计分析大作业

姓名：陈思远

学号：162140117

目录

写在前面

在大模型泛滥时代，需首先声明，本文为独立撰写，文本参考源仅限《应用多元统计分析》，未参考任何关于此数据集的博客内容，未直接复制或间接复制修改大模型生成的文本；

代码撰写使用 jupyter notebook，在所有 python 库安装后应该可以直接运行，结果将保存在“./result”文件夹下，另外生成了 html 格式，通过浏览器打开即可快速查验。

1 数据集介绍

本次大作业从复习、实现、深入理解分析所学统计分析方法出发，并不考虑使用过于庞大复杂的数据集。因此选择 UCI 数据集网站中的威斯康辛州乳腺癌数据集（Breast Cancer Wisconsin (Diagnostic)-Donated on 10/31/1995）。该数据集包含 **569 个样本**，每一个样本包含 **30 个特征**，并拥有 **1 个标签**，指明该样本代表的病例为**恶性（M）**还是**良性（B）**。这些特征是根据乳腺肿块的细针抽吸物（FNA）的数字化图像（如图 1 所示）计算得出的，它们描述了图像中存在的细胞核的特征。

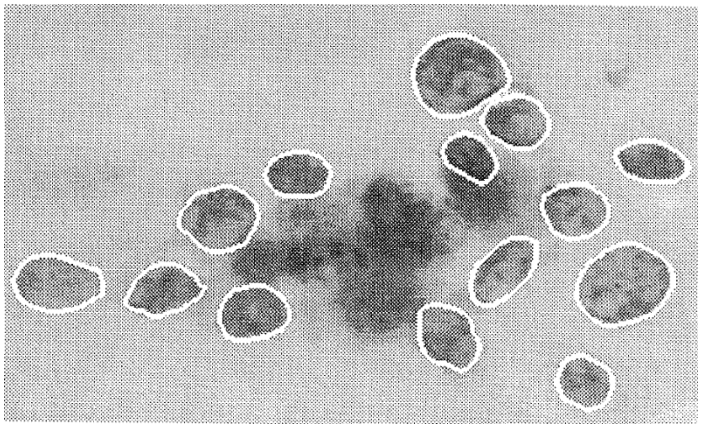
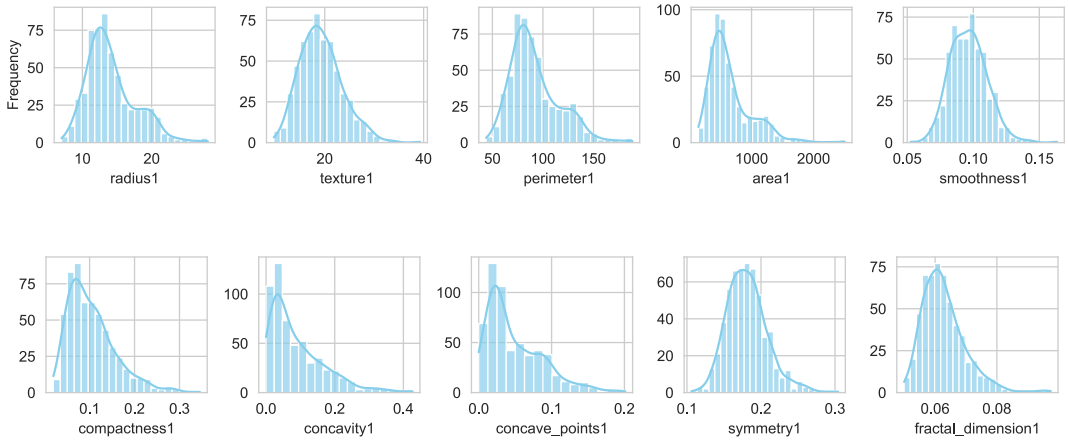


图 1 数字化图像样例（来自[1]）

2 数据基本特征分析

2.1 属性直方图

30 个特征可分为 **3 组**，每一组分别表示一个细胞的十个数字特征，初步研究每一个特征的基本分布，绘制属性直方图及核密度曲线如图 2 所示，可以看出所有属性基本都近似呈现出中间高、两边低的正态分布。



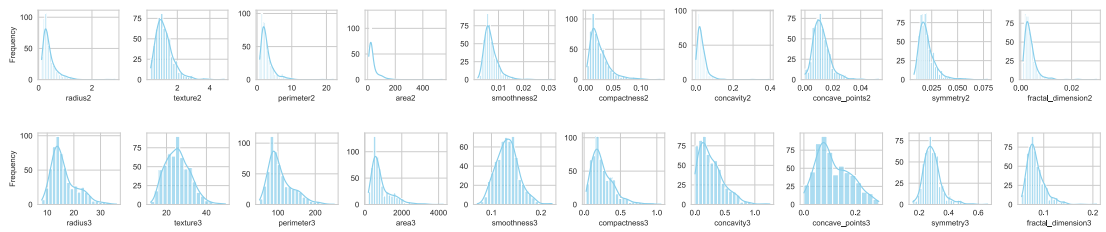


图 2 30 个特征的直方图与核密度曲线

2.2 正态性检验

为进一步确定任意属性是否近似服从正态分布，考虑对每一个属性进行 Shapiro-Wilk 正态性检验。该检验的统计量为：

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

其中，$x_{(i)}$表示样本中第i个小的数，$\bar{x}$为样本均值，a_i为常量，由有序独立同正态分布的统计量计算得出。得到的检验结果如图 3 所示。在显著性水平$\alpha = 0.05$的条件下，所有属性服从正态分布的概率值$p - value$均远低于显著性水平，因此否定某一属性服从正态分布的假设。

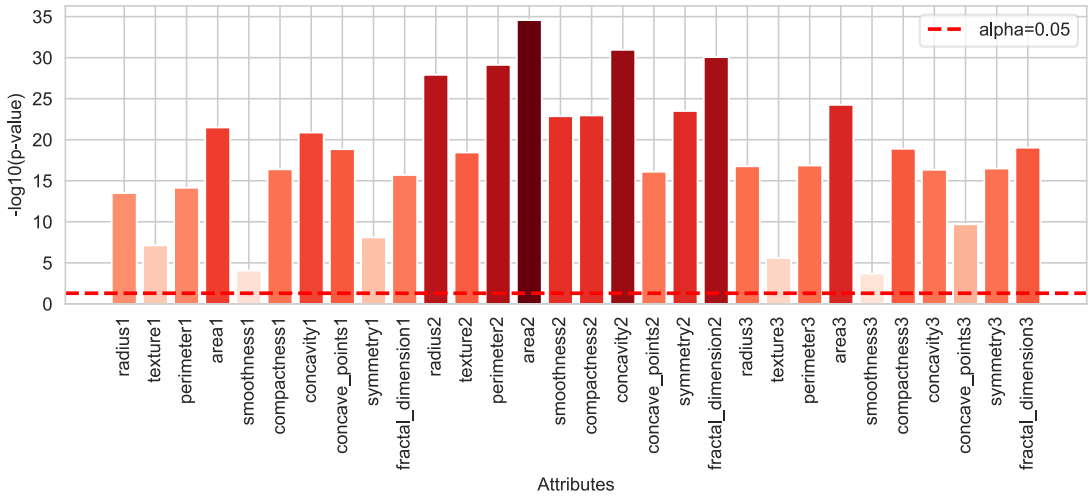


图 3 30 个特征正态性检验 p-value 可视化直方图

但是考虑到样本数目较小，在细胞维度测量数据难以避免噪声，另外，2.1 节中的直方图确实展现出大部分的属性具有正态分布的特征，因此，仍认为各项属性基本满足正态分布而不影响后续数据分析及假设检验。

2.3 最大似然估计与无偏估计

为进一步展现数据的分布特性，在 2.2 节默认各属性服从正态分布的前提下，对各正态分布的均值及方差进行估计。

对均值的最大似然估计及无偏估计为：

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_{(i)}$$

对方差的最大似然估计为：

$$\hat{\Sigma}_{ml} = \frac{1}{n}A$$

$$A = \sum_{i=1}^{n}(X_{(i)} - \bar{X})(X_{(i)} - \bar{X})^T = X^T[I_n - \frac{1}{n}1_n 1_n^T]X$$

无偏估计为：

$$\hat{\Sigma}_{ub} = \frac{1}{n-1}A$$

部分估计结果如表 1 所示。

Attribute	mean	max likelihood variance	unbiased variance
radius1	14.12729	12.39709	12.41892
texture1	19.28965	18.4664	18.49891
perimeter1	91.96903	589.4028	590.4405
area1	654.8891	123625.9	123843.6
smoothness1	0.09636	0.000197	0.000198
compactness1	0.104341	0.002784	0.002789

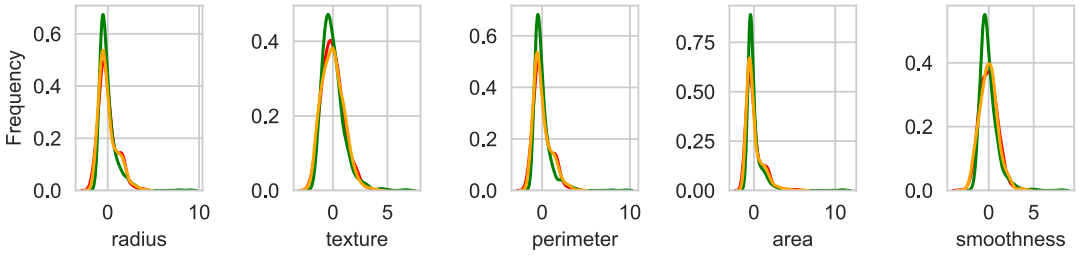
表 1 部分特征参数的最大似然估计和无偏估计

2.4 组间及类间差异

2.4.1 组间：三总体均值向量的检验

正如前面提到，数据集中包含三组指标，每一组指标包含十个指标，一个样本的每组指标为对一个细胞图像进行测量得到的数据结果，那么对于同一个样本这三组指标间是否有较大差异呢？通过观察数据发现，每组之间的数值差异很大，猜想可能是量纲或者是三个不同的细胞采集自不同区域，但是对于同一组指标不同样本的采集区域相同。因此先对每一个属性进行 z-score 标准化，去除量纲及其他无关因素的影响。

接下来，对三组指标的对应十个指标分布绘制核密度曲线，如图 4 所示。可以观察到不同组别的每个对应指标有相似的分布，即均展现出正态分布的特性，均值及协方差都类似。



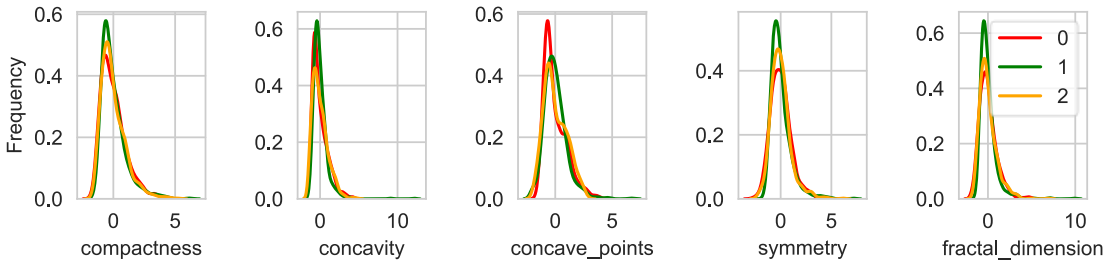


图 4 不同组别属性的核密度曲线对比图

既然都表现出肉眼可见的相似性，就有必要对此猜想进行假设检验。为贴切课程所学，我假设已知三总体的协方差相等，则可做出的假设为三总体均值向量相等：

$$H_0: \mu_1 = \mu_2 = \mu_3, \quad H_1: \mu_1 \neq \mu_2 \ or \ \mu_1 \neq \mu_3 \ or \ \mu_2 \neq \mu_3$$

此时统计量为：

$$\Lambda = \frac{|A|}{|T|} \sim \Lambda(p, n-k, k-1), \quad A = A_1 + A_2 + A_3$$

再转换为 F 统计量：

$$F = \frac{(n-k)-p+1}{p} \frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}$$

取显著性水平$\alpha = 0.01$，对上述假设进行检验。最终$p - value = 1.0 > \alpha$，接受原假设，三总体均值向量相等。可以注意到，$p - value$已经无线趋近于 1，说明了在三总体协方差相等的假设下，可以认为三组指标的均值向量相等。

2.4.2 类间：两总体均值向量的检验

该数据集为针对乳腺癌的数据集，包含了良性和恶性两种标签，自然地将数据分为两类。后续的判别分析需要建立在两个总体的均值有显著差异的基础上，因此需要首先判别这两个总体之间的均值向量是否有差异。

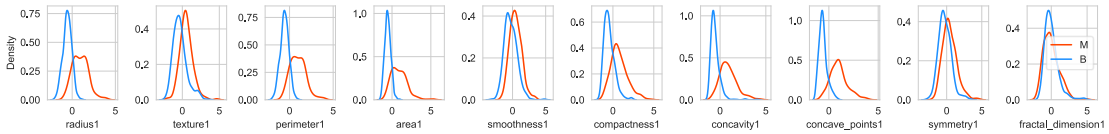
首先通过绘制不同指标的核密度曲线，初步分析各指标在不同总体间是否有显著差异。从图 5 中可以看出，大部分指标在不同总体间有较为显著的差异。则做出两总体各指标均值相等的假设：

$$H_0^j: \mu_1^j = \mu_2^j, \quad H_1^j: \mu_1^j \neq \mu_2^j$$

为简化实验，默认各指标的协方差相等但未知。则检验统计量为：

$$(n+m-2)\frac{nm}{n+m}(\bar{X}-\bar{Y})^T(A_1+A_2)^{-1}(\bar{X}-\bar{Y}) \sim T^2(p, n+m-2)$$

在显著性水平$\alpha = 0.01$的条件下，除了$symmetry2$这个指标以$p - value = 0.013$接受该假设，其余指标均拒绝该假设，说明均值向量有显著差异。此外，对两个总体整体也进行检验，最终$p - value$远远小于显著性水平，认为两个总体间的均值向量有显著差异。



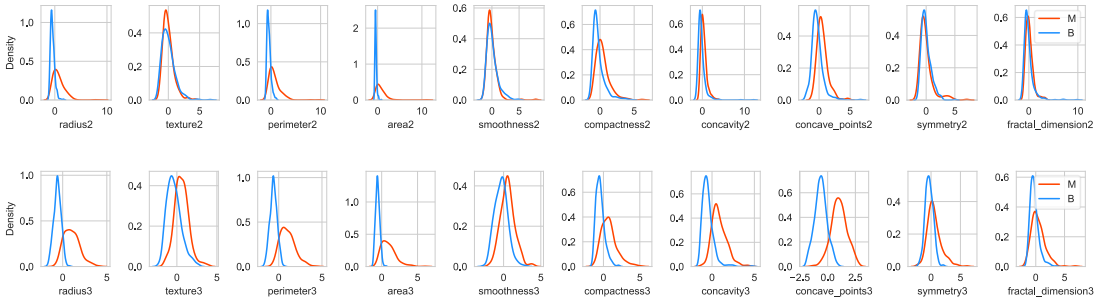


图 5 30 个属性不同类别之间的核密度曲线对比图

2.5 属性相关性分析

三组指标之间存在明显相关性的结论已经可以从均值向量的检验中得出。为进一步探明三组指标内部是否存在较为明显的相关性，对组内指标进行相关性分析。绘制的相关系数热力图如图 6 所示。从图中可以发现，各组内 radius、perimeter、area 三个指标之间相关性都非常高，甚至接近 1，这表明这三项指标之间可以相互表示，信息相互重叠。在后续的分析中需要注意这三项指标，避免重复信息的叠加。

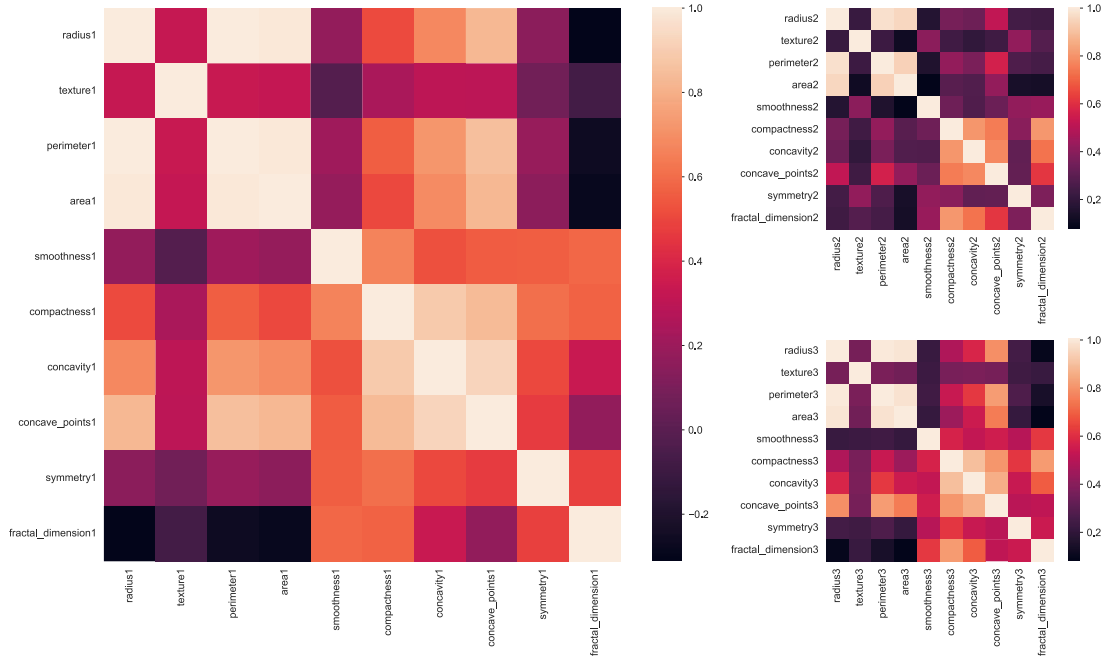


图 6 不同组别间 10 个特征的相关系数热力图

3 回归分析

在 2.5 节中，发现了 radius、perimeter、area 三个指标之间具有较强的相关性，因此每个指标都能够以较小的损失表示为其余两个指标的线性组合。在本节中，期望用第一组中的其余 9 个指标对第一项指标 radius1 进行拟合，即线性回归。表示为：

$$radius1 = C\beta + \varepsilon$$

其中$C=[1_n, texture1, perimeter, \cdots, fractal_dimension]$，$\varepsilon \sim N_n(0, \sigma^2 I_n)$。

为了便于可视化，在求解出回归方程后，施加 one-hot 形式的 mask，遮盖住其余指标而只关注 radius1 和指定的一个指标，即只展现这二者在回归方程中的相关性。可视化结果如图 7 所示。可以发现 radius1 与 perimeter1 的数据点呈现良好的线性相关性，而回归方程也完美地刻画了、拟合了这种线性相关性。对于 radius1 与其他指标，呈现的大多是随机性，因此回归方程难以拟合这种离散的缺乏线性相关性的数据。而 area1 属性似乎与 radius1 呈现出类似 log 函数的分布特性。

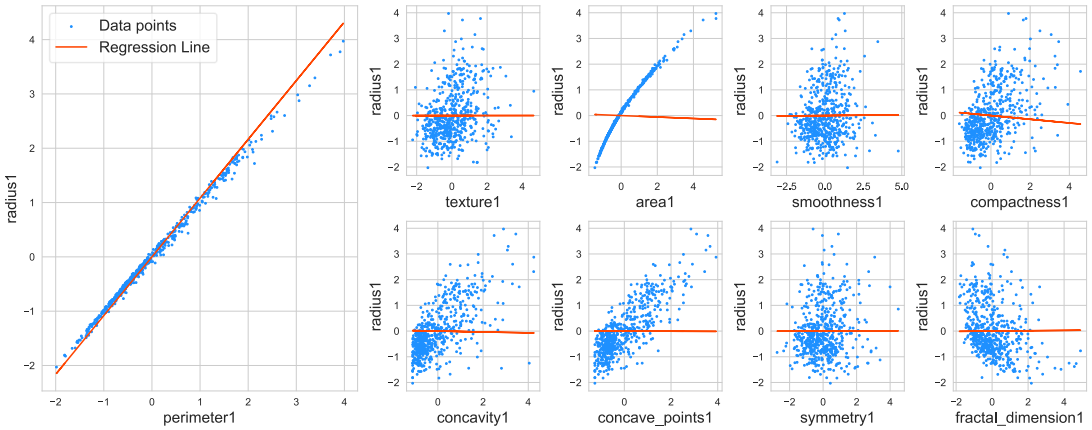


图 7 组别 1 中 radius1 与其他 9 个属性线性回归结果的可视化图像

实际上，如果考究这些特征的具体含义（做深度学习做久了自己都养成不看特征的具体含义的坏习惯了……），在考虑细胞为类圆形的情况下，半径 radius1 当然与周长 perimeter1 为线性相关性，在未归一化指标下二者的回归系数应当为2π，而面积 area1 当然与半径 radius1 为平方关系，其分布呈现出 log 函数的特性也就非常自然了。而至于其他指标，在不挖掘其高维语义的情况下，仅仅探究数值之间的联系，自然呈现随机性。

由于 radius 属性与 perimeter 属性的强相关性，拟合的均方误差为 0.0006，可以认为是测量误差以及细胞不完全类圆造成的。

4 聚类分析

4.1 系统聚类

系统聚类首先将每一个样品作为一个类别，然后按照一定的距离度量方式，计算每个类别之间的距离，选择距离最短的两个类别进行合并，经过不断地迭代更新，最终将样本聚为指定的类别数。这种无监督的方法，在不知道类别数时可以通过离差平方和指导的统计量确定类的个数，从而挖掘出总体的一些浅层特征。在明确类别个数的情况下，也是一种很好的分类手段。

对数据进行聚类分析，采用离差平方和的距离度量方式的聚类谱系图如图所示。从谱系图中可以清晰地看到，在类别数为 2 的时候，两个类别之间的合并距离有非常大的跨越，表现出数据呈现 2 类被的空间分布，当然可以明确这两个类别分别是恶性肿瘤与良性肿瘤。现在只需统计每个簇之中占比较大的种类，便可确定簇的类别，进而达到分类的目的。分类的结果将在后续特定章节进一步论述对比。

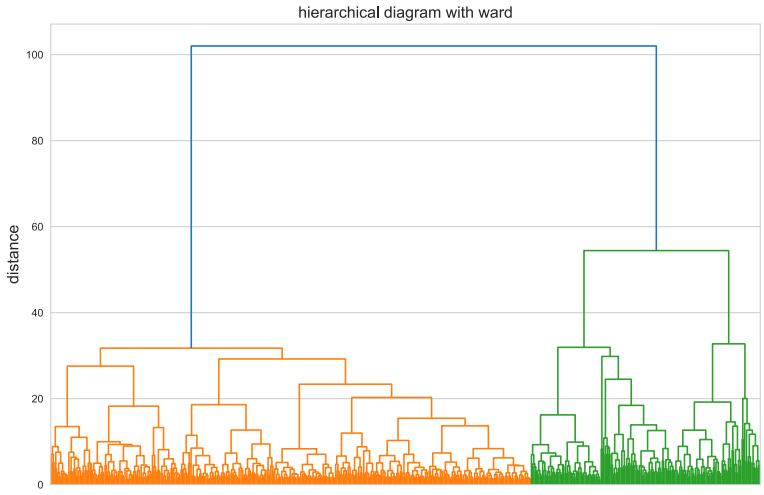


图 8 系统聚类谱系图

4.2 动态聚类

动态聚类与系统聚类不同，其步骤特性要求需要指定类别个数。获得先验知识类别个数后，该方法随机初始化相应个数的簇中心（传统），并利用指定的距离度量方法，将样本归为距离簇中心最近的簇，然后更新簇中心（按批和逐个修改法有差异），进行迭代更新。在利用先验知识明确类别个数时，这种聚类方法也是一种很好的分类方法。

对数据进行 K-means 聚类，聚类结果如图 9 所示。为便于可视化以及消除之前指出的强相关属性对距离度量的影响，这里首先采用主成分分析将数据降低至 2 维。聚类后将簇中数目较多的样本类别作为簇的标签。

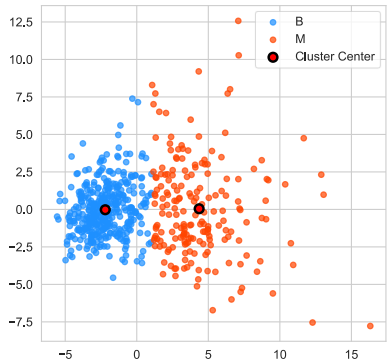


图 9 二维动态聚类结果

两个簇的分界面非常清晰，因为采用欧式距离作为距离度量的方式，分界线基本为两个簇中心连线确定的中垂线。分类结果可视化如图 10 所示。可以发现在两个类别交界处，存在许多误分类样本，表明线性判别面难以完美刻画两个类别的分界面。事实上，这些难以被线性分界面划分的样本，就是支持向量机中寻找的支持向量，是机器学习中重视的难样本（hard sample）。

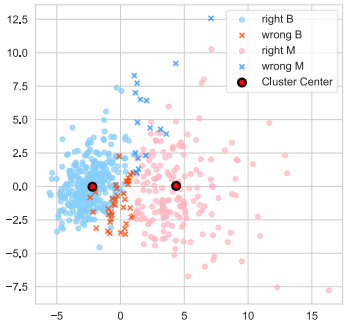


图 10 二维动态聚类准确率结果

5 判别分析

5.1 距离判别（马氏距离）

个人认为距离判别以结果为导向来看就是一步到位的动态聚类。动态聚类的动态之处在于不断更新簇中心，而距离判别则直接使用先验知识确定簇中心，然后快进到动态聚类最后一步，直接根据一定的距离度量准则得出聚类结果。

为了获取先验的簇中心，将样本按 7：3 分为训练集和测试集，使用训练集计算估计均值向量得到两个类的类中心以及马氏距离中使用的两个类别的协方差矩阵，然后根据马氏距离对测试集进行距离判别，最终准确率为 26.08%。值得注意的是因为使用马氏距离，这里我并没有进行任何降维手段或者归一化、标准化处理，而是直接使用原始数据，虽然没有量纲影响，但是有强相关属性，这可能导致信息的重叠，而使得这些强相关属性代表的信息在距离度量中占据主导地位。

5.2 Fisher 判别

相比于距离判别，贝叶斯判别考虑了先验概率和类间散度，但是个人认为这样的先验概率于类间散度在小样本中对分类带来的影响微乎其微，反而会带来难以解释的结果、难以避免的偏置，因此在这个数据集上不进行贝叶斯判别。而 Fisher 判别则考虑了类间散度与类内散度，采用最小化类内散度、最大化类间散度的优化目标，从而利用投影将数据投影到易于分类的投影空间。这样的优化目标对我的启发很深，因为只需要非常完美地抵近这两个优化目标的帕累托前沿，再取一个适合当前任务的点，就能够对数据进行很好地分类。而我也在我的研究中按照这种思想，设计出先聚类再分类架构的深度学习架构，并取得了较好的结果。

同样，按照 7：3 分割训练集与测试集，假设投影函数为

$$f(X) = a^T X$$

利用训练集最大化广义瑞利商即可求得投影方程系数：

$$max\ \frac{a^T B a}{a^T A a}$$

将数据投影到一维后，取两类样本均值的中心为分界点，带入测试集，即可进行分类。最后的分类精度为 97.07%，显著高于判别分析。

为了进一步可视化 Fisher 判别的原理和过程，以及与主成分分析（还没有开始正式分析，但迫不及待）的区别，首先利用主成分分析将数据降维至两维，虽然这样先利用主成分分析的方式对比较 Fisher 判别与主成分分析有一点影响，但并不会过多地影响二者的本质区别以及可视化结果。分别利用 Fisher 判别与主成分分析将这个二维数据投影至一维，并绘制出 Fisher 判别直线、Fisher 投影直线、PCA 投影直线。为展现不同类别数据的分布，将其真实标签标注在图中。可视化结果如图 11 所示。

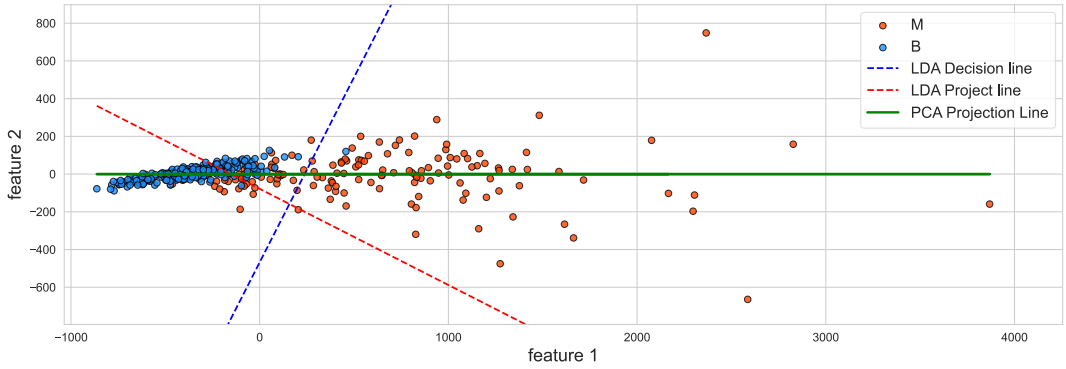


图 11 Fisher 判别分析与主成分分析对比图

可以看见，在图中的两类数据在降维至二维后，分布为明显的两簇，具有基本的线性可分性，Fisher(LDA)判别直线能够较好地将两类样本分开。此外，应该注意到在判别直线的右下方只有一个错分的蓝色负样本点，而其左上方有许多错分的橙色正样本点，个人认为导致这样的不平衡因素主要有两点：

- 蓝色负样本点（良性肿瘤）聚集密集，若判别直线向上偏移，会导致错分样本急剧增加，这样似乎并不具有说服力，因为此时橙色错分点在减少。但是如果考虑 Fisher 的分类原理，是通过最大化类间散度，最小化类内散度的优化目标实现的。若判别直线向上偏移（其实应该考虑投影直线更正确，因为判别直线的偏置项是由投影直线上的两类样本中心确定的），由于蓝色样本点聚集密集，其与橙色样本点在投影空间中的类间散度会增加，而这是与优化目标相悖的；
- 另一方面，橙色样本点分散稀疏，当投影直线向上偏移时，橙色样本点的类内散度会急剧增加，这同样与优化目标不符。

因此，出现了这样的不平衡情况。主成分分析的投影曲线为横轴，实际上，这是由于将数据降低至二维时使用了主成分分析，而主成分分析就是利用将数据按最大化方差的方向进行投影达到降维目的。降低至二维时，横轴为数据最大方差方向，纵轴为第二大方差方向，当再次使用主成分分析时，只需要将数据投影到最大方差方向即横轴。

观察投影直线可知，主成分分析只考虑了将所有数据的方差最大化，而不会考虑图中三条直线形成的难分类的三角区域。Fisher 判别降维方法则考虑了类间散度与类内散度，使得被投影的数据有更好的线性可分性。通常来说，有监督的 Fisher 降维方法在分类任务中是优于无监督的 PCA 降维方法的。

6 主成分分析

6.1 主成分个数选择

主成分分析的本质是在降维过程中减少数据方差损失。自然的，这种数据方差代表了数据中蕴含的信息，毕竟一堆有一定规律的散点总是要比一个点的蕴含的信息量大。那么在不考虑下游任务对数据维度要求的情况下，将数据降维到多少维是一个很重要的问题。这很显然是一个权衡利弊（trade-off）的问题。数据维度太高，不利于处理大规模数据，数据维度太低，数据信息会损失很多。对于这类问题，很好的一个解决办法就是绘制帕累托前沿面，然后根据问题实际情况选择合理的方案。而在没有实际问题所带来的偏好的情况下，有一个办法就是选择前沿面剧烈变化的地方，因为一般来说，剧烈变化点就是不同的偏好的折中点。

主成分个数的选择一般也遵循这个方法，通过绘制数据方差与降维维度之间的变化关系图，选择随着主成分个数增加，方差增益急剧减少的点。通常来说，在维持数据方差 80%以上的情况下，选择维度最低的降维方案。

绘制主成分个数与数据方差的关系图如图 12 所示。

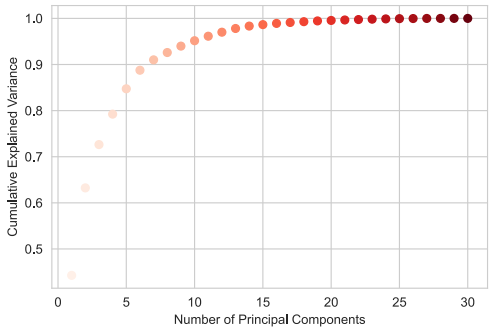


图 12 主成分分析累计方差图

可以发现，在主成分个数为 5 时，蕴含的数据方差已经能够接近 85%，当主成分个数为 15 时，方差的解释占比已经接近 99%，而后随着主成分个数的增加，方差的增益已经微乎其微。这表明各属性之间包含的大部分信息有重叠，只需要不到 20 个增强的属性就能刻画数据分布。一般来说，会选择主成分个数为 5 至 10。

6.2 主成分降维

主成分分析的明显优点在于，通过保证数据方差不显著减少，用更少的属性刻画原有特征空间，这样能够在降低数据维度的同时，减少原始信息中大量冗余与重复，特别是在直接基于距离进行分类的任务中，它能够克服相关性较强的变量对距离贡献累积性。

为探明主成分降维对数据信息的保留情况，以及对下游分类任务的影响，通过改变选择的主成分个数，将数据降低至不同维度，进一步通过距离判别测试分类的准确度。绘制变化散点图如图 13 所示。可以明显观察到，当主成分个数为 1 时，距离判别准确率低于不使用主成分降维；当主成分个数到达 5 后，随着主成分个数的增加，距离判别准确率逐渐提升；当主成分个数为 15 时，距离判别准确率急剧上升，达到最大值；而如果再继续增加主成分个数，准确率却略微降低。

这样的变化过程完美展示了主成分分析降维对数据信息的影响，即较少的主成分难以刻画原有的特征空间；随着主成分个数的增加，对原始信息的保留程度逐渐提升，包含的原有数据信息也逐渐增加，通过判别分析对数据进行分类的准确率也因此得到提升；但是当主成分个数过多时，原始信息的刻画已经饱和，继续引入的主成分可能会带来信息的冗余重叠与噪声，反而影响距离判别的准确率。因此，利用主成分分析进行数据降维时，应该合理谨慎地选择主成分个数，避免主成分个数太少而缺失信息，或者主成分个数过多而带来信息的冗余。

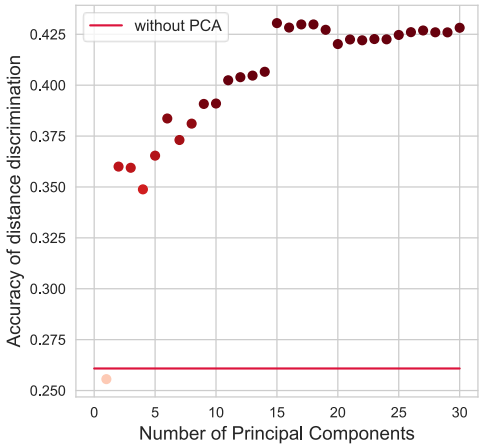


图 13 不同主成分个数下马氏距离判别的准确率散点图

7 因子分析

因子分析的主要目标是将多个变量综合为少数几个因子，即构建因子模型如下：

$$X - \mu = AF + \varepsilon$$

F为公共因子，ε为特殊因子，A为因子载荷阵。这样通过几个增强属性刻画原有特征空间的方式与主成分分析十分相似，溯源来说，因子分析是主成分分析的发展。但是二者有着较大的区别。主成分分析通过分析数据本身的方差信息，保证方差即信息损失尽量小，达到刻画原有特征空间的目的。因子分析通过压缩数据信息，“归纳总结”出一组变量，同时刻画这组变量构成的特征空间与原始空间之间的差异性，达到解释原有空间的目的。因子分析构建的是一个数学模型，更偏向于数学建模类的方法，可解释性也因此更强。

7.1 因子数目选择

同样的，因子的数目也需要通过先验知识确定，但是这个因子数目选择相较于主成分个数的选择似乎更难以解释理解。每一个主成分都携带着原始信息的一部分方差或者说信息，第一主成分最多，其他次之，这很好理解。而每一个因子都负责解释一部分原始信息，这显得因子是一个隐变量，在确定因子载荷阵之前，无法说明解释的是什么信息，这一定程度上对因子数目的选择造成了一定的困难。

现有选择因子数目的方法是通过分解样本协方差阵，然后根据特征值的大小进行选择（因子分析参数估计时的主成分法）：

$$S = \sum_{i=1}^{p} \lambda_i l_i l_i^T$$

这样的处理方法使得它与主成分分析的差别实际上基本只剩下口头解释了，因为从求解方法来说很难区别。绘制 Scree 图（每个因子的特征值的降序排列）如图 14 所示。和之前主成分个数与数据方差之间的关系图类似，都呈现长尾分布。在因子序号 5 处前后有斜率变化，在因子序号 15 之后，特征值几乎为 0。

图 14 主成分法求解因子模型特征值的分布

进一步分析，选择因子数目为 30，与原始特征数目相等，将求得的因子载荷阵可视化如图 15 所示。可以观察到$Factor1-5$解释了大部分变量，特别是$Factor4$对变量 3 个$texture$指标的解释性很强且主要解释这类指标。此外，在$Factor15$之后的因子对于的因子载荷阵中的元素值几乎为 0，表明这些因子几乎不参与对任何变量的解释，可以丢弃。基于以上分析，可以确定因子数目在 5 至 15。对于进行方差最大化旋转的因子分析相关的图可以参见附录，结果基本一致。

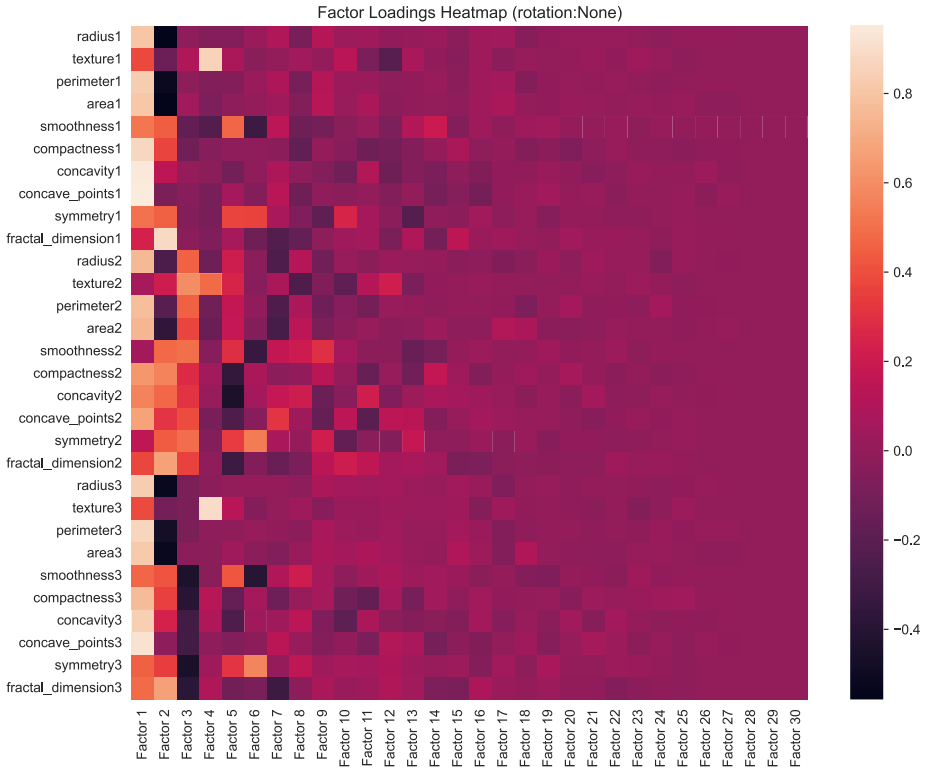


图 15 因子载荷阵可视化热力图

7.2 变量共同度及因子贡献度分析

由上一节内容，这里取 15 个因子进行分析处理，因子载荷矩阵（使用了方差最大化旋转）可视化如图 16 所示。可以注意到每个因子对变量都有一定的解释，前两个因子对大部分变量都具有解释性，同时这也说明了这些变量之间具有一定程度的相关性，$Factor3$对变量 3 个$smoothness$指标的解释性很强$Factor4$对变量 3 个$texture$指标的解释性依然很强，$Factor5$对变量 3 个$symmetry$指标的解释性很强，之后的因子大多是针对某一个变量进行偏置的补充。

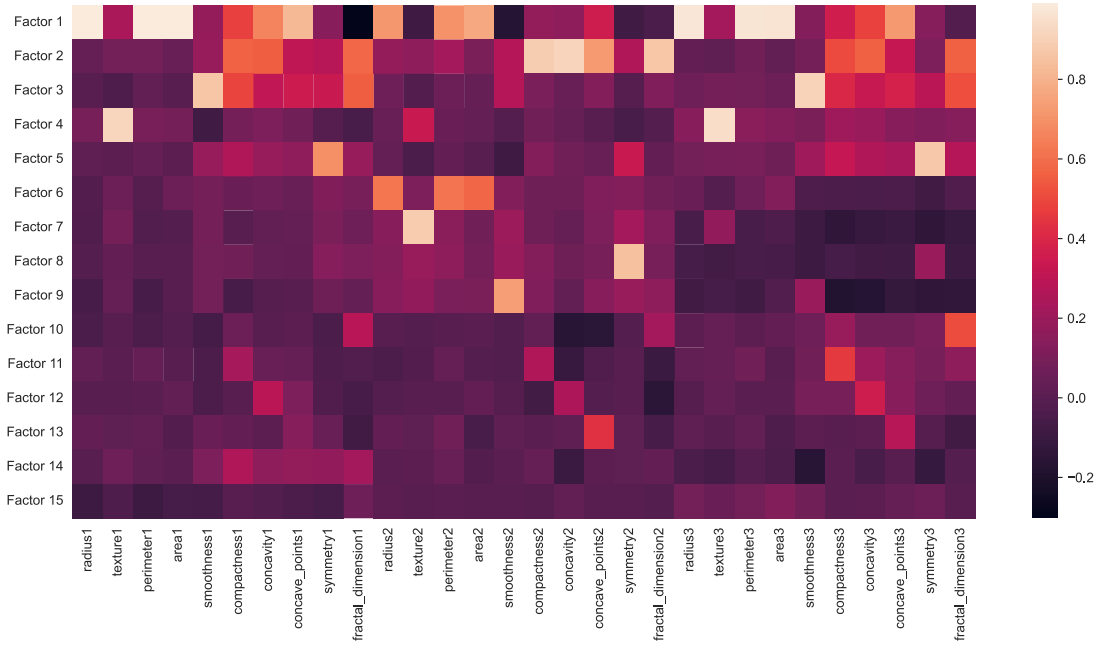


图 16 15 个因子载荷矩阵可视化热力图

7.2.1 共同度分析

因子载荷矩阵A中的各行元素的平方和称为变量的共同度，变量X_i的共同度记为：

$$h_i^2 = \sum_{j=1}^{m} a_{ij}^2, \qquad i = 1,2,\dots,p$$

这个值同样可以通过方差来解释，变量X_i的方差可分解为：

$$Var(X_i) = h_i^2 + \sigma_i^2$$

h_i^2为公共因子F的方差，σ_i^2为特殊因子的方差，因此，前者越大，说明公共因子包含的该变量方差越大，对该变量的解释程度越大，类似于能够从公共因子还原出变量X_i，反过来也反映了变量X_i对公共因子的依赖程度。将每个因子的贡献度可视化如图 17 所示。从图中可以看到，几乎每个变量对公共因子都有较大的依赖性，部分变量甚至接近 1，说明每个变量能够很好地被公共因子刻画，而这些公共因子构成的因子模型相应地能够刻画原始数据空间。

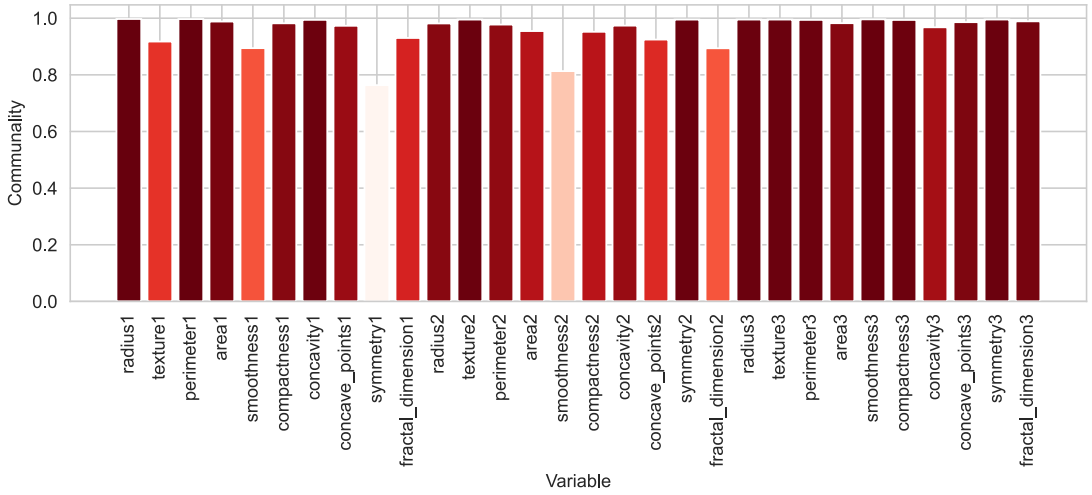


图 17 30 个属性变量的共同度可视化直方图

7.2.2 贡献度分析

因子载荷矩阵A的各列的平方和称为公共因子的贡献度，公共因子F_j对X的贡献度记为：

$$q_j^2 = \sum_{t=1}^{p} a_{ij}^2, \quad j = 1,2,\ldots,m$$

它的统计意义正好与共同度倒置，q_j^2表示第j个公共因子对X所有分量的总影响，也就是第j个公共因子在因子模型刻画原始特征空间时的贡献程度，它是衡量第j个公共因子与其他公共因子之间相对重要性的指标之一。将公共因子贡献度可视化如下图所示。从图中可以看出，$Factor1$的贡献度占比非常大，且随着因子序数的增大贡献度减小，这主要是采用主成分法求解导致的。

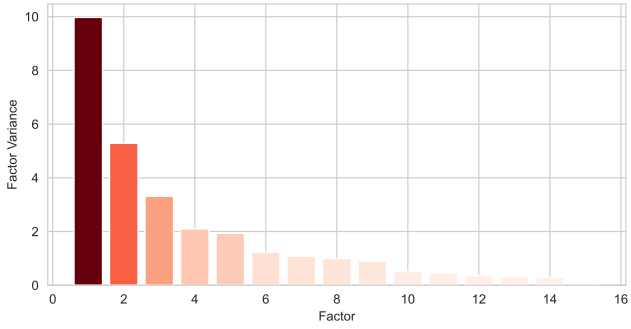


图 18 15 个公共因子的贡献度可视化直方图

7.3 因子得分

通过数据计算出因子载荷阵后，便能够计算出每个样本在该因子模型中的因子得分向量，即在因子模型刻画的空间中的表示。因子得分与主成分降维后得到的主成分得分一样，可以用于下游任务，进行分类、聚类、回归等任务。使用因子分析将数据降维至二维，可视化如图 19 所示。可以发现，两个簇已经具有良好的线性可分性。

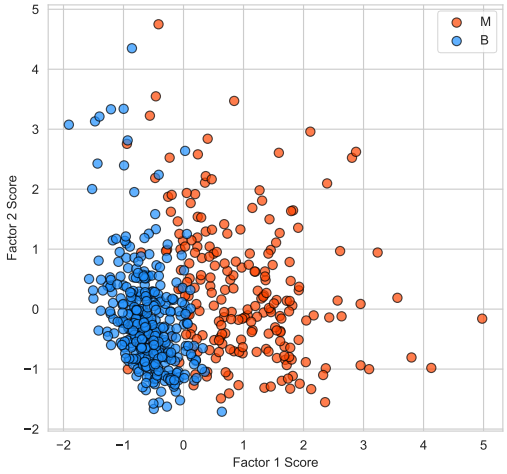


图 19 两个因子的因子模型得出的数据因子得分可视化散点图

8 降维与分类汇总

在《多元统计分析》这门课程中学习了 3 种可以用于**降维**的方法——主成分分析、Fisher 判别分析、因子分析，3 种可以用于**分类**的方法——距离判别、Fisher 判别、聚类判别（在知道类别个数时聚类也不失为一种很好的分类方法，可以看作增强版的距离判别方法）。为了进一步掌握这些方法，并对它们的优缺点进行比较分析，在本次大作业的最后，希望通过对数据进行先降维再分类的方式，进行对比分析。与之前一致，将数据 z-score 标准化后，按照 7：3 划分为训练集和测试集（更为严谨的做法是先划分再标准化）。

8.1 多种降维方法

分别利用线性判别分析（LDA）、主成分分析（PCA）、因子分析（FA）对训练数据和测试数据**分开**进行降维处理，根据之前章节的分析，降维维度分别设置为 **1、2、15**，其中 LDA 最多只能将数据降维至$C-1$维，即 1 维，故不参与 2、15 维的降维比较。

将训练数据降维至 1 维，绘制各类方法降维后的数据分布如图 20 所示。可以看到，每一类降维方法均可以将两类样本分开。其中，经过 LDA 降维后，两类样本的直方图交叉最少，两类样本几乎能够被完全分开。而 PCA 与 FA 降维后，两类样本的直方图有较大的重叠，较难分开。

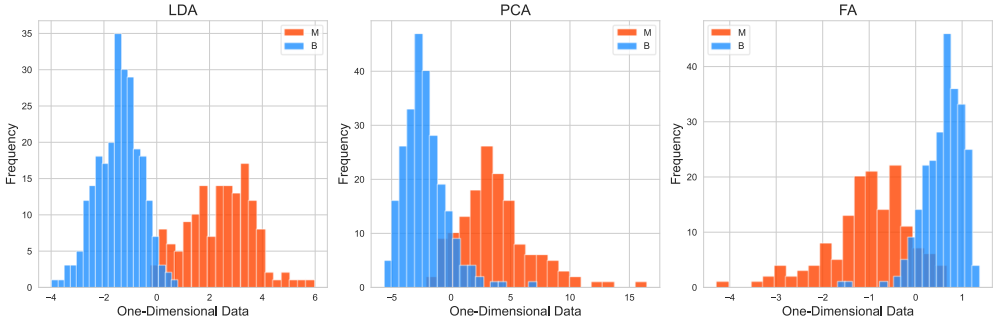


图 20 一维情形下不同降维方法的直方图可视化

将训练数据通过 PCA 与 FA 降维至 2 维，绘制降维后的数据散点分布如图 21 所示。对于两个方法，降维后的数据都呈现出良好的线性可分性，而至于具体效果将通过后续的分类来衡量。

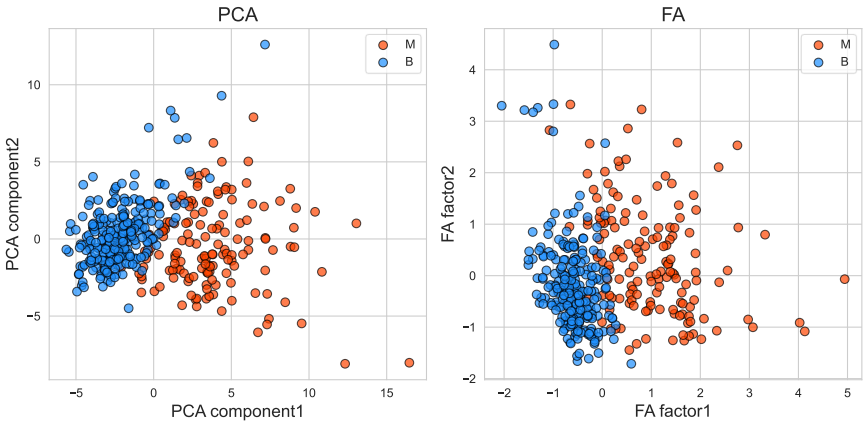


图 21 二维情形下 PCA 与 FA 的降维可视化散点图

8.2 多种分类方法

由于降维时已经利用过线性判别分析，故这里就不再将线性判别纳入分类方法的对比。分别考虑四种分类方法，广义平方距离判别、K-means 聚类判别、支持向量机、随机森林。其中，为了平衡无监督分类算法与有监督分类算法，广义平方距离判别将利用训练数据得出计算两类样本的协方差作为两类总体的协方差的估计，计算两类样本的均值作为类别中心的估计，以及先验概率；K-means 聚类判别将利用训练数据得出两类样本的中心作为初始的聚类中心，而这对于无监督的聚类算法来说非常重要。以上安排构建在，数据集划分后，训练集与测试集同分布的前提下。

统计结果如表 2 所示，横向为不同的降维方法，后缀数字代表降维的维数，纵向为不同分类方法。

分类方法	LDA-1	PCA-1	FA-1	PCA-2	FA-2	PCA-15	FA-15
广义平方距离判别	0.9883	0.9123	0.1053	0.9474	0.9357	0.9123	0.8772
K-means 聚类判别	0.9825	0.9123	0.1053	0.9064	0.9415	0.9240	0.9474
支持向量机	0.9883	0.9064	0.0819	0.9474	0.9240	0.9532	0.9591
随机森林	0.9766	0.8889	0.1345	0.9357	0.9240	0.9474	0.9123

表 2 不同降维方法、不同降维维度、不同分类方法最终结果

8.3 结果分析

8.3.1 不同降维方法

从表中可以看出，LDA 降维后的数据在不同分类方法中的均得到非常好的划分，准确率最高，展现了 LDA 最小化类内散度、最大化类间散度优化目标的优异性。而 FA-1 则效果很差，而与之有类似求解方法的 PCA-1 却表现地较好，虽然二者的理论基础不同，但是能够造成这么大的差异是我不能想象、不能理解、更不能解释的。特别是在二者一维情形下直方

图如此相似的情况下，这里我暂且留下疑问（或许程序写错了没发现……）。

PCA 与 FA 随着降维后维度的增加，不同分类器的结果相比一维情形下基本都在提升，部分结果稍有下降。对于广义平方距离判别，可以非常直观的发现，当维度从 2 升至 15 时，其分类准确率比较明显地下降，推测这是由于维度的上升导致噪声引入，从而干扰了距离准则的判别。

8.3.2 不同分类方法

为使得变化趋势更加清晰，删除 FA-1，绘制折线图如图 22 所示。从折线图中可以看出，不同分类方法在不同的降维方法下有不同的表现，支持向量机在 LDA、PCA-2、PCA-15、FA-15 中均得分最高，展现出 SVM 强悍的分类能力；而根据之前的分析，广义平方距离判别在数据中含有噪声时，容易被干扰，导致性能下降；而 K-means 虽然也是基于距离判别，但应该是因为其类别中心是动态更新的，因此能够在一定程度上抵抗这些噪声的干扰。

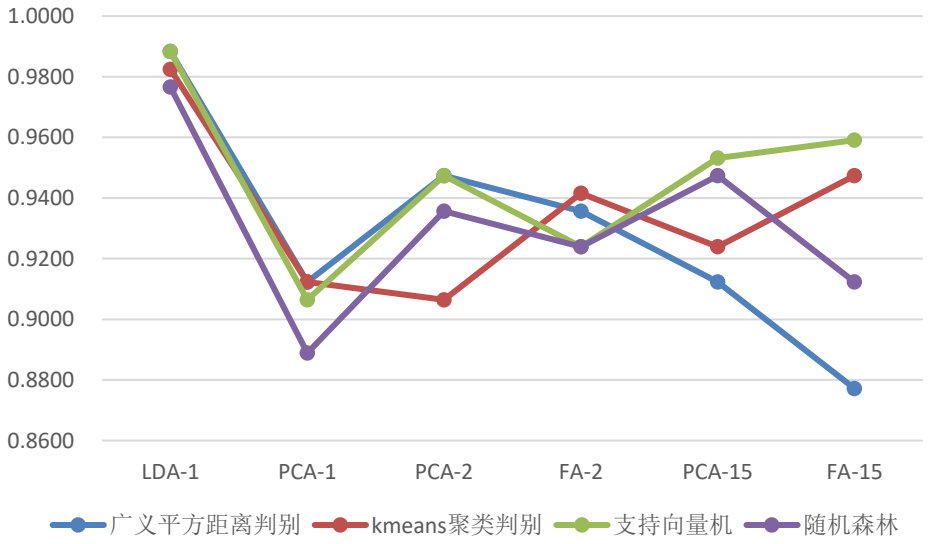


图 22 不同降维方法、不同降维维度、不同分类方法最终结果折线图

9 总结

在本次大作业中，首先对数据集进行了基础的数字特征分析，检验了数据的正态性、两类总体的可分性，在此基础上，进一步通过对本课程中几乎所有相关知识的应用，对该数据集进行详实充分的分析。同时，利用该数据集，对多种降维、分类方法进行充实的可视化，展现了方法的基本原理，洞察了方法的本质，理清了不同方法之间的差异。最后一部分对多种降维、分类方法联合应用、对比分析，也进一步展现了统计分析方法的魅力。

非常高兴在期末前能有这样一段空闲时间，进行这样一种层层递进的分析过程，既包含了抽丝剥茧的充实，也充满了拨云见日的喜悦。

10 参考文献

[1] Street W N, Wolberg W H, Mangasarian O L. Nuclear feature extraction for breast tumor diagnosis[C]//Biomedical image processing and biomedical visualization. SPIE, 1993, 1905: 861-870.

附录

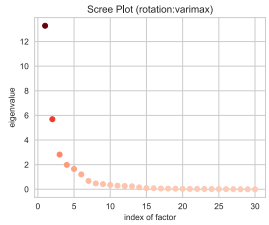


图 23 主成分法求解因子模型特征值的分布（经过方差最大化旋转）

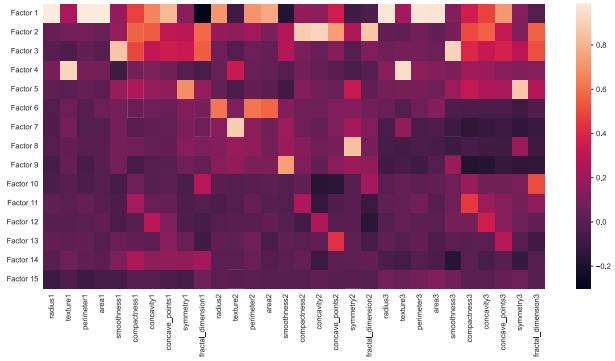


图 24 15 个因子载荷矩阵可视化热力图（经过方差最大化旋转）

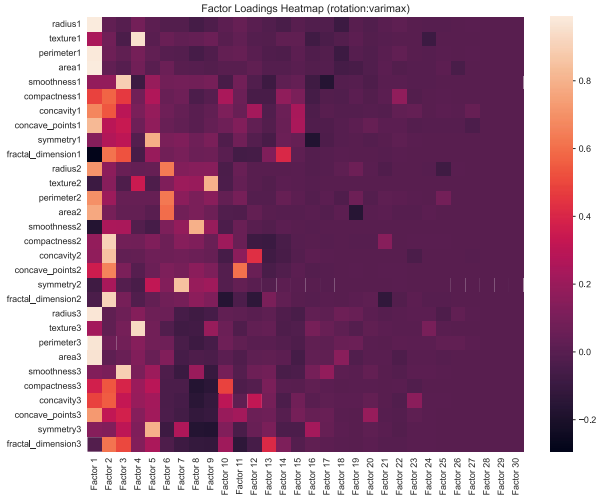


图 25 因子载荷阵可视化热力图（经过方差最大化旋转）