模式识别技术

大作业

成牧原 12022215226

云南大学信息学院

2022 年 12 月 15 日

2022 年 12 月 15 日

Problem 1

(1) 图 1 显示了 4 种装配零件的彩色图像：螺母 (nuts)、螺钉 (screws)、支架 (brackets) 和垫圈 (washers)。请你设计一组特征（不超过 4 种），能很好地把 4 种零件分开。分析、讨论你的特征和背后的原理。
(2) 编写代码提取每种模式的上述 4 种特征，绘制每对特征的二维散点图（如, f1 vs. f2, f3 vs. f4, 等），讨论你的结果。
(3) 计算原始空间中每一对样本之间的欧式距离。首先，你得把所有图像都转换为同样大小，比如 32×32 彩色像素（提示：在 MATLAB 中，可使用函数 imresize）。把距离组成 30×30 的矩阵进行显示（提示：在 MATLAB 中，可使用函数 imagesc）。在二维特征空间中，重复上述过程。比较两种距离矩阵，在原始图像空间中和特征空间中的类别可分离性有何不同？讨论你的结果。

图 1: 数据集

Solution.

(1)(2) 基于以下三个特征使用多种算法进行分类，分别为

1) 图像中圆形的面积占比，

$$F_1 = \frac{\sum Area_{circle}}{W \times H}$$

可以帮助分别出垫圈和螺母，可以看到一个垫圈通常是一个圆环，即两个圆。而螺母只有一个圆形，面积占比相对较小。在计算完成后将 F1 缩放到 [0,1] 之间。

$$F_1(i) = \frac{F_(i) - min(F_1)}{max(F_1) - min(F_1)}$$

2) 图像中 4 连域的个数，因为支架有很多孔，可以辅助分离螺钉和支架。四连域，如果二值图像的一个像素点的上下左右四个像素点值都为 1，则认为这是一个四连域，计算完成后也进行归一化。

3) 二值化图像的和除以图像像素点总数，可以帮助分离出螺钉，因为螺钉较细，在图像中占的空间不大。

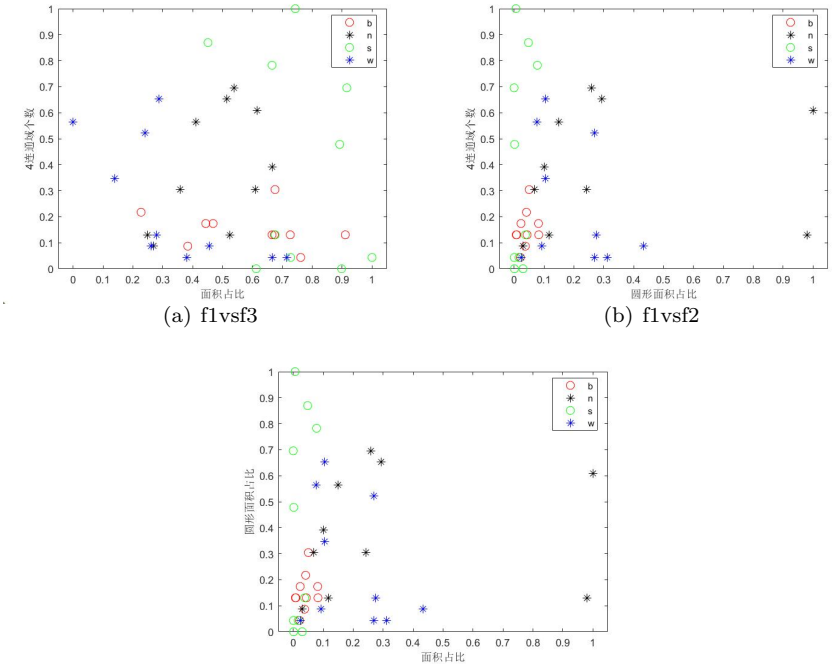
$$F_3 = \frac{\sum_i^W \sum_j^H pixel(i,j)}{W \times H}$$

，计算得到面积后同样进行归一化

在提取特征和绘制散点图之前，需要对图像进行一些预处理，主要包含了以下两个过程

1) 图像灰度化，二值化

2) 图像缩放为 64×64



(a) f1vsf3　　(b) f1vsf2

(c) f2vsf3

图 2: 特征散点图

通过观察发现，图 2(a) 中的散点比较分散，而 (b)(c) 相对比较密集。使用 SVM 进行分类 (不设置测试集)，得到如下结果。

分类算法	准确率 (f1 vs f3)	准确率 (f2 vs f3)	准确率 (f1 vs f2)
精细高斯 SVM	90%	85%	90%
三次 SVM	77.5%	85%	87.5%

根据分类的结果来讨论，我认为使用 f1 和 f2 进行分类效果最好。

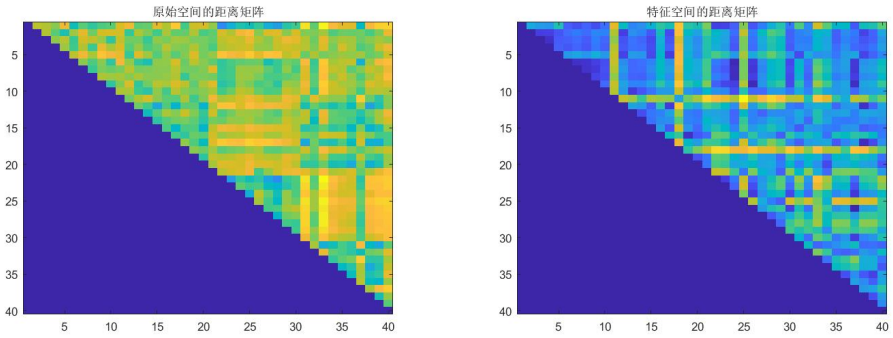
(3) 本题中共有 40 个样本，如果两两求距离矩阵，那该矩阵的大小应该为 40×40, 如果只是将一个样本与另外 30 个不同类别的样本求距离，那矩阵的大小应该为 40×30。将计算得到的矩阵 (40×40) 画出来 (矩阵是对称阵，所以只画一半)。本题的原始空间欧氏距离定义为两幅二值图像 (矩阵) 对应位置相减的平方和开根，即

$$distance = \sqrt{\sum_{i=1}^{W}\sum_{j=1}^{H}(A(i,j) - B(i,j))^2}$$

其中 A，B 为两个样本 (图像),W,H 为图像的宽和高。而特征空间的欧氏距离定义为

$$featureDistance = \sqrt{\sum_{i}^{3}((Fi_A - Fi_B)^2)}$$

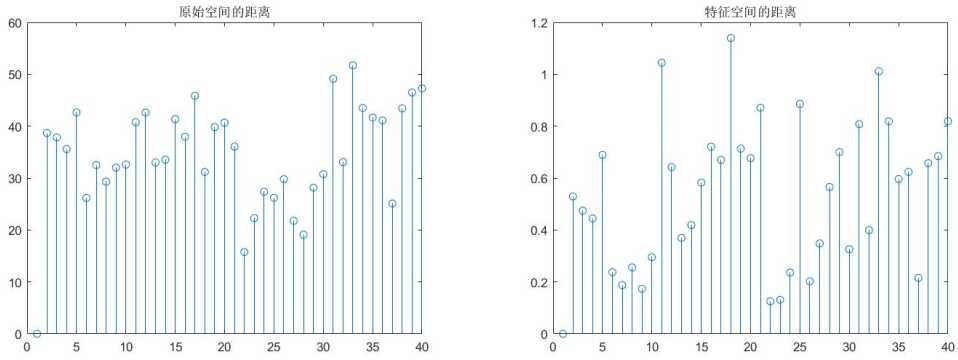
即每个特征的差的平方和开根号。



(a) 原始空间的欧氏距离矩阵　　(b) 特征空间的欧式距离矩阵

图 3: 距离矩阵

因为数据较多，不方便观察，所以用第 1 幅图片与另外 39 幅图片的的欧式距离排列成一个向量，然后绘制针状图。



(a) 非特征空间欧氏距离　　(b) 特征空间欧式距离

图 4: 欧式距离针状图

从距离上看，不同类别之间并没有非常明显的差异，因此我认为这个分离性比较差。使用前文提到的三个特征中的 f1 和 f3 输入 SVM 或者 KNN(不设置测试集) 均能得到不错的分类效果。

分类算法	准确率
精细高斯 SVM	90%
精细 KNN	100%
线性判别	52.5%

以上分类结果均可在 MATLAB 的 Classification Learner APP 中计算得到。

2022 年 12 月 15 日

Problem 2

考虑一个医学诊断问题，用快速生化检验筛查病人。根据下列似然函数，健康者的检验返回结果接近 0，受感染者的返回结果接近 1：

$$\begin{aligned} p(v|w_1) &= N(\mu = 0, \sigma = 0.3) \\ p(v|w_2) &= N(\mu = 1, \sigma = 0.1) \end{aligned} \tag{1}$$

假设平均 1 万个患者中有 1 人受感染，且误诊的代价如下：(1) 将健康者诊断为“感染者”：预计病人综合医疗费用为 2 万人民币。(2) 将感染者诊断为“健康者”：预计由于误诊导致的医疗费用为 100 万人民币。根据下列准则，分析并确定决策规则：

1) 最大似然

2) 最大后验概率

3) 最小贝叶斯风险

Solution.

约定 v 为病人的检测值，x 为检测结果 (x=0 为阴性，x=1 为阳性),H 为健康，I 为感染。

(1) 我们可以根据某个病人的检验结果带入两个正态分布中计算似然函数值，如果 $p(v|w_1)$ 的概率大，那就决策为健康，如果 $p(v|w_2)$ 的概率大，那就决策为健康.

(2) 设定一个阈值 T，当病人指标 v 大于 T 时，检验结果为 x=1(阳性)，否则 x=0(阴性)，则这个人感染病毒的后验概率是

$$P(I|x) = \frac{P(x|I)P(I)}{P(x)}$$

其中 P(x|I) 表示当染病时，检验结果为 x 的概率，P(I) 表示人感染病毒的先验概率，P(x) 表示检验结果为 x 的概率。又有

$$P(x) = P(x|I) \times P(I) + P(x|H) \times P(H)$$

$$P(H) = 1 - P(I)$$

根据上述信息，我们可以计算得到检验结果为阳性时，人真的感染病毒的后验概率为 $P(I|x)$，随后通过枚举 T 的值或者通过求导等办法来求出使得后验概率最大的 T。假设染病概率分布函数为

$$F_1(x) = \frac{1}{\sqrt{2\pi} \cdot 0.1} \int_{\infty}^{x} e^{-\frac{(v-1)^2}{2 \cdot 0.1^2}} dx \tag{2}$$

健康的概率分布函数为

$$F_1(x) = \frac{1}{\sqrt{2\pi} \cdot 0.3} \int_{\infty}^{x} e^{-\frac{(v-0)^2}{2 \cdot 0.3^2}} dx \tag{3}$$

求 $P(I|x=1)$

$$\begin{aligned} P(I|x=1) &= \frac{P(x=1|I)P(I)}{P(x=1)} \\ &= \frac{(1-F_1(T))\times 0.0001}{P(x=1)} \\ &= \frac{(1-F_1(T))\times 0.0001}{P(x=1)} \\ &= \frac{(1-F_1(T))\times 0.0001}{(1-F_1(T))\times 0.0001+(1-F_0(T))\times(1-0.0001)} \end{aligned} \tag{4}$$

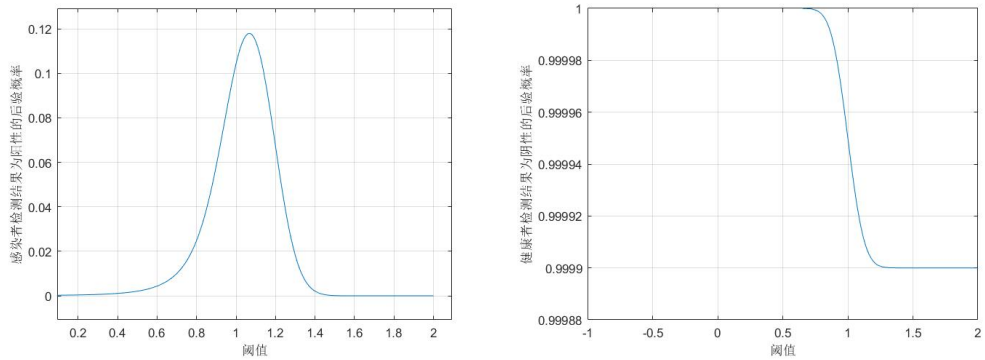
我们的目标是找出一个 T 使得公式 (3) 的值最大，同理我们求一下 $P(H|x=0)$

$$\begin{aligned} P(H|x=0) &= \frac{P(x=0|H)P(H)}{P(x=0)} \\ &= \frac{F_0(T)\times(1-0.0001)}{P(x=0)} \\ &= \frac{F_0(T)\times(1-0.0001)}{P(x=0)} \\ &= \frac{F_0(T)\times(1-0.0001)}{F_0(T)\times(1-0.0001))+F_1(T)\times 0.0001} \end{aligned} \tag{5}$$

设定阈值查找范围为 0-2，步长为 0.01 进行搜索，当 T=1.07 时，$P(I|x=1)=0.1180$ 最大，而又有

$$P(H|x=0)\geq 0.9999 \quad when \quad 0\leq T\leq 2$$

，因此可以认为 T 的值对 $P(H|x=0)$ 影响不大。所以主要根据 $P(I|x=1)$ 判断，当指标小于 T=1.07 时，判定为健康人，当大于 1.07 时，判定为感染者。



(a) P(I|X=1)　　(b) P(H|X=0)

图 5: T-后验概率曲线

(3) 本小题仍然是确定阈值，同样我们将其设为 T。我们根据后验概率来计算期望风险。误诊的概率是

$$\begin{aligned} P(H|x=1) &= \frac{P(x=1|H)P(H)}{P(x=1)} \\ P(I|x=0) &= \frac{P(x=0|I)P(I)}{P(x=0)} \end{aligned} \tag{6}$$

期望风险为

$$R(T) = 2 \times P(H|v > T) + 100 \times P(I|v < T) \tag{7}$$

与 (2) 采用同样的方式求解

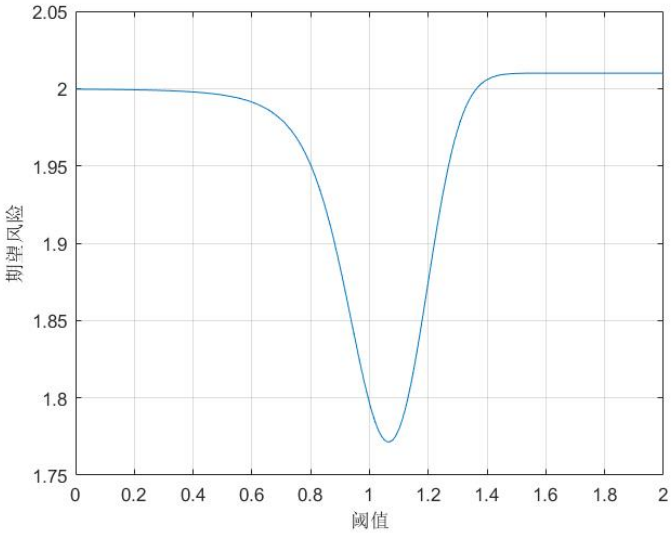


图 6: 不同阈值下的期望风险

当 T=1.06 时，风险期望最小，等于 1.7716 万。

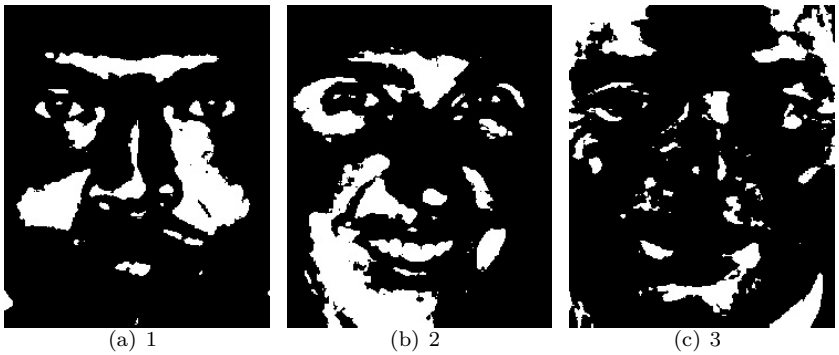
Problem 3

文件夹“csfac”包含 53 幅人脸正面图像，对这些人脸进行 PCA 分解。
(1) 生成平均脸的图像，
(2) 生成前六个特征向量的图像（即“特征脸”），
(3) 绘制相应主成分的二维 PCA 散点图。讨论您的结果。

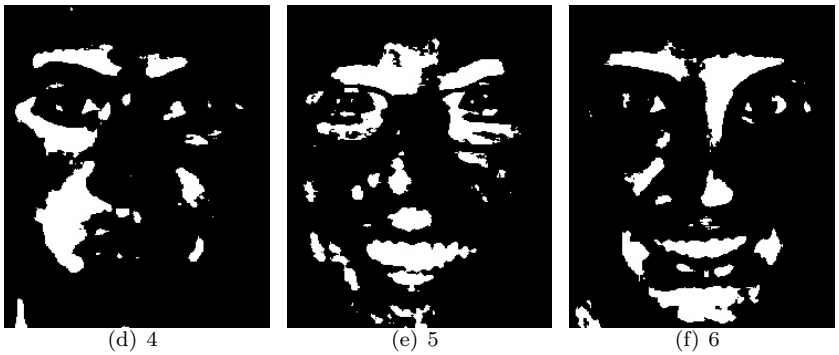
Solution.
(1) 将所有图片求均值即可得到平均脸。

图 7: 平均脸

(2) 为了防止计算量过大，先将图片转为灰度图片，然后通过 PCA 降维到 50 个特征，然后绘制前六个特征脸



(a) 1 (b) 2 (c) 3



(d) 4 (e) 5 (f) 6

图 8: 前六个特征脸

(3) 绘制散点图并讨论结果

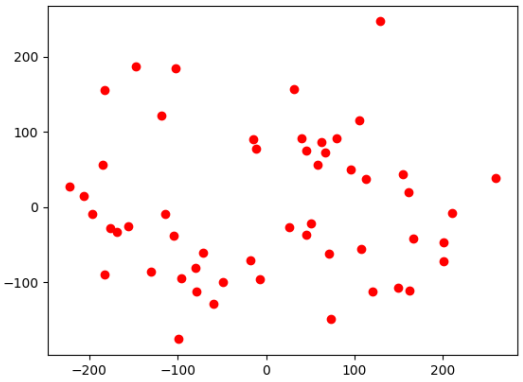


图 9: PCA 散点图

根据图 9，可以看到，这 53 个样本比较分散，并没有出现特别明显的聚集现象。说明这 53 个样本之间的相似程度比较低，这将有益于对这 53 幅图像进行分类。

2022 年 12 月 15 日

Problem 4

数据集 hw1p4dat 中包含下述三维问题的数据。
(1) 绘制数据集中每一对特征的二维散点图，并对数据的结构进行分析。
(2) 估计数据的平均向量和协方差矩阵。协方差矩阵中非对角线项与 (1) 中的散点图一致吗？为什么一致或不一致？
(3) 使用你在 (2) 中估计的均值向量和协方差矩阵生成高斯分布的数据集。(提示：可使用命令 mvnrnd)
(4) 使用你在 (3) 中生成的数据集，重复做 (1)。这里的散点图与 (1) 中的一致吗？为什么一致或不一致？讨论你的结果。

Solution.
(1) 绘制得到的散点图如下

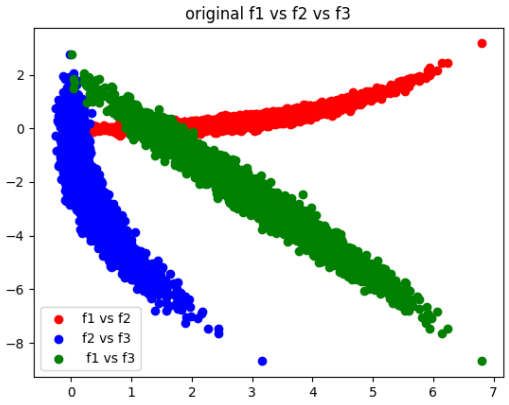


图 10: 原始数据的散点图

不难发现，各个特征之间都不是互相独立的，而且 f2 和 f3 负相关，f1 和 f3 负相关，f1 和 f2 正相关。且 f1 vs f2 的散点图，f2 vs f3 的散点图的线性的程度明显低于 f1 vs f3。
(2) 平均向量就是三个特征的平均值组成的向量

$$\hat{v} = \begin{pmatrix} 3.0323392488991194 & 0.36865053884307447 & -2.551861607445957 \end{pmatrix}$$

使用 np.random.cov 求出的协方差矩阵是

$$cov = \begin{pmatrix} 0.97136798 & 0.29223774 & -1.45366002 \\ 0.29223774 & 0.1133424 & -0.43705089 \\ -1.45366002 & -0.43705089 & 2.26436308 \end{pmatrix}$$

不难发现 cov(1,2) >0，则 f1 和 f2 正相关，cov(1,3) <0，则 f1 和 f3 负相关。同时 cov(2,3) <0 则 f2 和 f3 负相关。这与 (1) 中目测的结论一致。
(3)(4) 使用 np.random.multivariateNormal 生成 4995 个数据。并绘制散点图

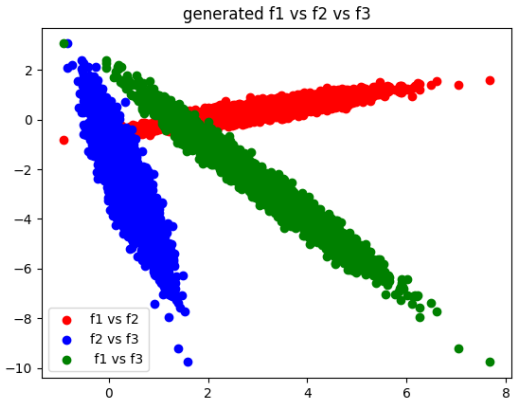


图 11: 根据平均向量和协方差矩阵生成的数据的散点图

可以明显观察到生成的数据线性程度明显高于原始的数据，所以散点图并不一致。我认为，任意两组一维数据都可以求出协方差和平均向量，但根据协方差和平均向量只能估计出数据之间是否独立，是否正相关或者负相关。数据之间是否遵循线性关系则无法估计。最后给出根据生成的数据得到的协方差和平均向量

$$\hat{v_g} = \begin{pmatrix} 3.0517519941316706 & 0.37372607740464564 & -2.589954667422647 \end{pmatrix}$$

$$cov_g = \begin{pmatrix} 0.97137977 & 0.29448247 & -1.45318156 \\ 0.29448247 & 0.1141593 & -0.44060412 \\ -1.45318156 & -0.44060412 & 2.26073453 \end{pmatrix}$$

Problem 5

数据集 hw1p5data 中包含由非线性函数 $y = f(x) + n$ 合成的数据，其中 n 为加性噪声。请你研究多项式函数能在多大程度上可以用来表示这个关系。

(1) 随机选择 $n = 10$ 个数据点作为训练数据，其余数据点用作测试样本。建立一阶多项式模型（例如 $y = ax + b$,，提示：在 MATLAB 中，使用命令 polyfit）。以测试样本对测试样本的形式，绘制模型输出。计算模型的均方误差（MSE，模型预测值与正确输出值之间的平方误差平均值）。

(2) 对于 2-10 阶多项式，重复做 (1)。

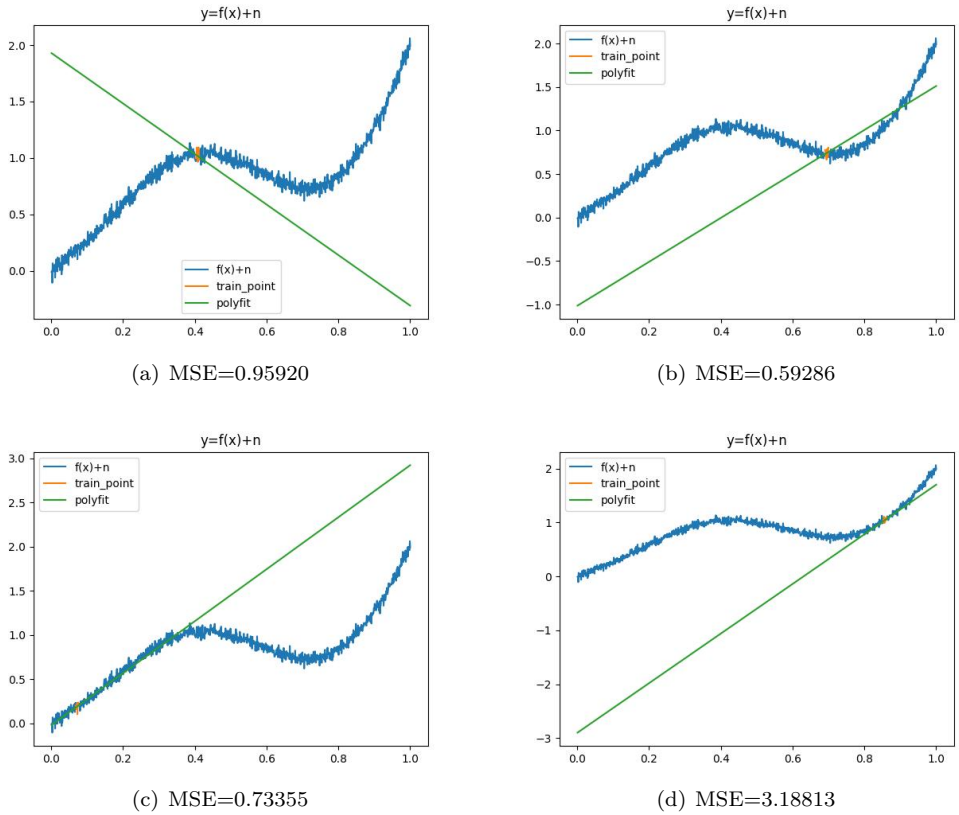
(3) 重复做 (1) 和 (2) 各 100 次，估计每一阶多项式重复 100 次的平均 MSE。绘制 log(MSE)（即对数尺度的 MSE）与多项式阶数的关系图。

(4) 对训练集大小为 $n = 15, 20, 25, 50, 100, 200$，重复做 (1)-(3)。

(5) 讨论模型的 $log(MSE)$ 随多项式阶数和用于训练模型的样本数如何变化。讨论你的结果。

Solution.

(1) 随机抽取连续的 10 个训练点，拟合一个线性函数，进行四次实验得到如下拟合结果

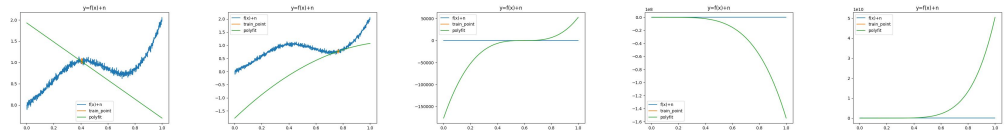


(a) MSE=0.95920　(b) MSE=0.59286

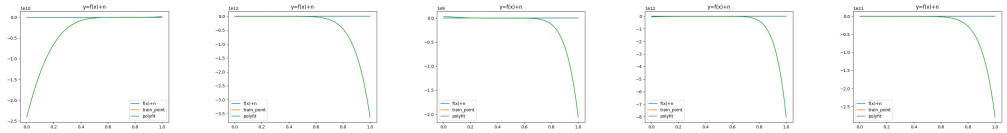
(c) MSE=0.73355　(d) MSE=3.18813

图 12: 四次实验结果

(2)1 阶到 10 阶的模型输出如下图



(a) MSE=0.95920 (b) MSE=1.07296 (c) MSE=2830996 (d) MSE=243047 (e) MSE=$1.9{\cdot}10^{20}$



(f) MSE=$4.6\cdot10^{19}$ (g) MSE=$6.9\cdot10^{23}$ (h) MSE=$1.7\cdot10^{17}$ (i) MSE=$2.4\cdot10^{24}$ (j) MSE=$3.6\cdot10^{41}$

图 13: 1-10 阶的 MSE

图 13 中 (a)-(j) 分别表示 1-10 阶的模型输出。

(3) 每一阶重复 100 次求 log(MSE) 得到

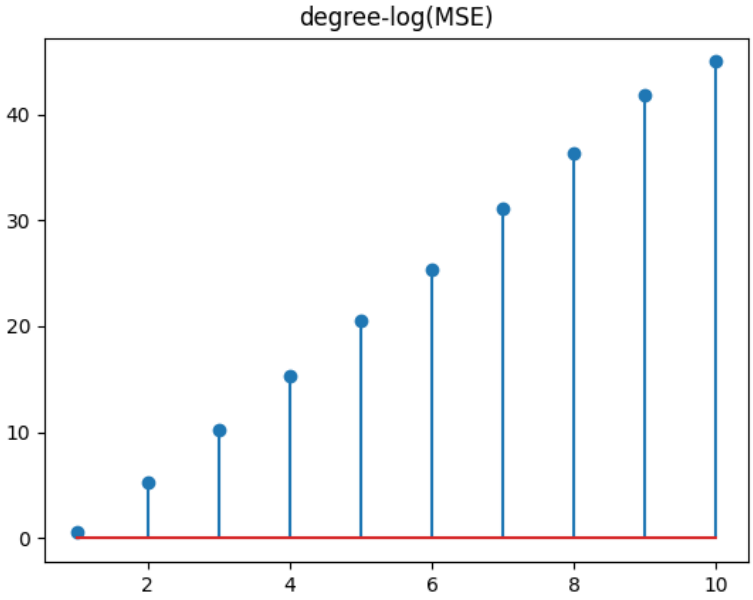
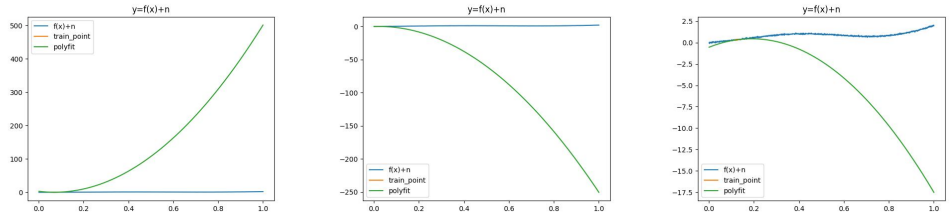


图 14: degree-log(MSE)

(4) 因为生成的图像过多，故仅选择 2 阶的 6 张图片进行分析，其他的图片放在工程目录中

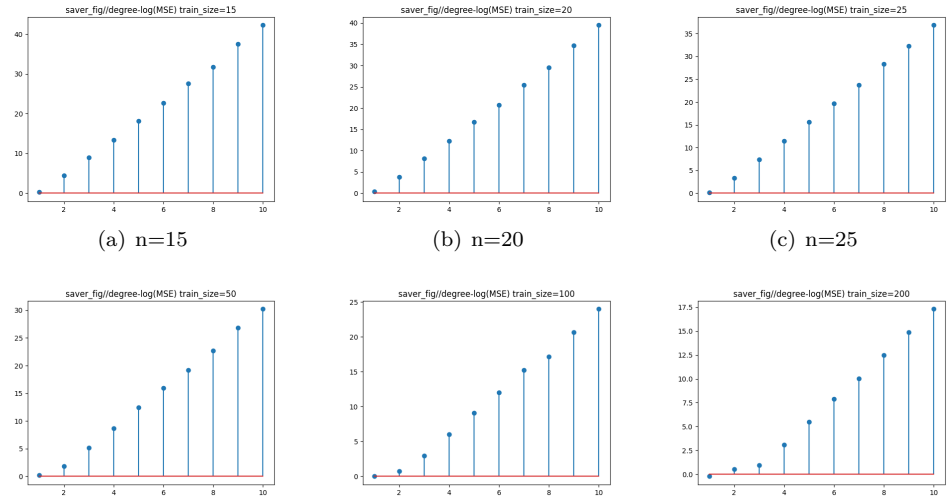


(a) n=15,MSE=46546 (b) n=20,MSE=12506 (c) n=25,MSE=59.44

(d) n=50,MSE=45.210 (e) n=100,MSE=0.73420 (f) n=200,MSE=6.3359

图 15: 训练集大小不同的情况下 2 阶拟合的结果

可以明显观察到随着训练集的变大，拟合的效果变好，误差也随着训练集变大而变小。



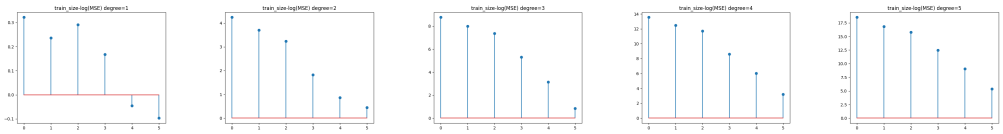
(a) n=15 (b) n=20 (c) n=25

(d) n=50 (e) n=100 (f) n=200

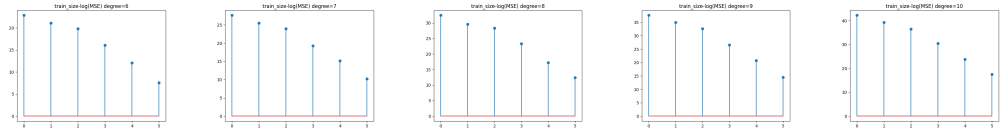
图 16: 训练集大小不同的情况下阶数和 log(MSE) 的关系

可以明显观察到，无论测试集的大小，log(MSE) 与阶数呈现出明显的线性关系。而当训练集变大时，log(MSE) 明显变小，阶数比较大时尤为明显。

(5) 固定阶数，计算每个训练尺寸下的平均 MSE(计算 100 次的均值)，绘制训练尺寸和 log(MSE) 的关系图。



(a) 1 阶 (b) 2 阶 (c) 3 阶 (d) 4 阶 (e) 5 阶



(f) 6 阶 (g) 7 阶 (h) 8 阶 (i) 9 阶 (j) 10 阶

图 17: 1-10 阶的 trainSize-log(MSE)

综合 (1)-(4) 和图 17, 首先可以得到两个明显的结论，对于 f(x)，当训练数据不变时，拟合的多项式阶数越高，误差越大；当拟合的多项式阶数固定时，如图 17 所示，训练数据越多，误差越小。