一、 题型

分析：3×5′

简答：5×5′

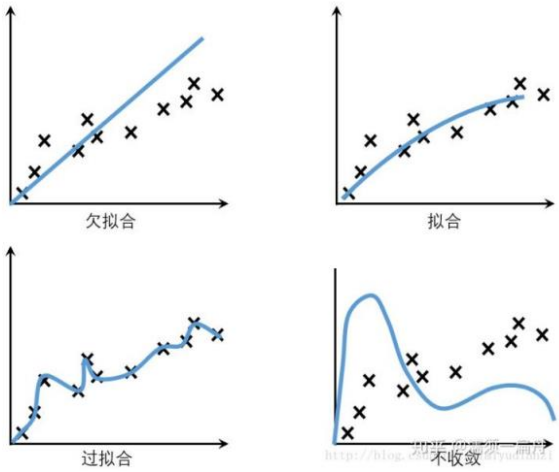
计算：3×10′

代码：2×15′

二、 分析题

2.1 分析欠拟合、过拟合

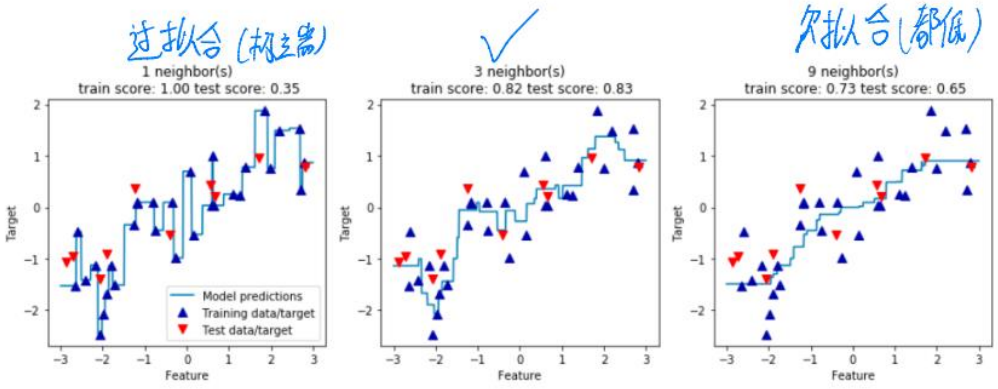
【例】



欠拟合：预测结果与真实结果相差太大，对数据的拟合效果太差。

过拟合：虽然对数据的拟合效果很好，但是这是基于训练集上训练得到的模型复杂，对于新样本的预测结果并不一定很好。

【例】



不同n_neighbors值的k近邻回归的预测结果对比

2.2 分析K值选择

原理

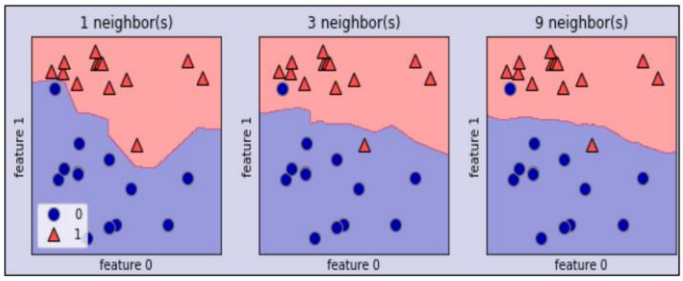
选择较小的K值，就相当于用较小范围中的训练实例进行预测，近似误差减小，估计误差增大，意味着整体模型变得复杂，容易过拟合，容易受到异常点的影响。

选择较大的K值，就相当于用较大范围中的训练实例进行预测，近似误差增大，估计误差减小，意味着整体的模型变得简单，容易欠拟合，容易受到样本均衡的问题。

PS：K与近似误差变化趋势相同，与估计误差变化趋势相反。

【例】

观察分类界面的变化：



随着k的变化，分类界面边更加平滑

k

大：学习的近似误差会增大，模型变得简单。

小：敏感性增强，估计误差会增大，模型变得复杂。

2. 3 模型的选择

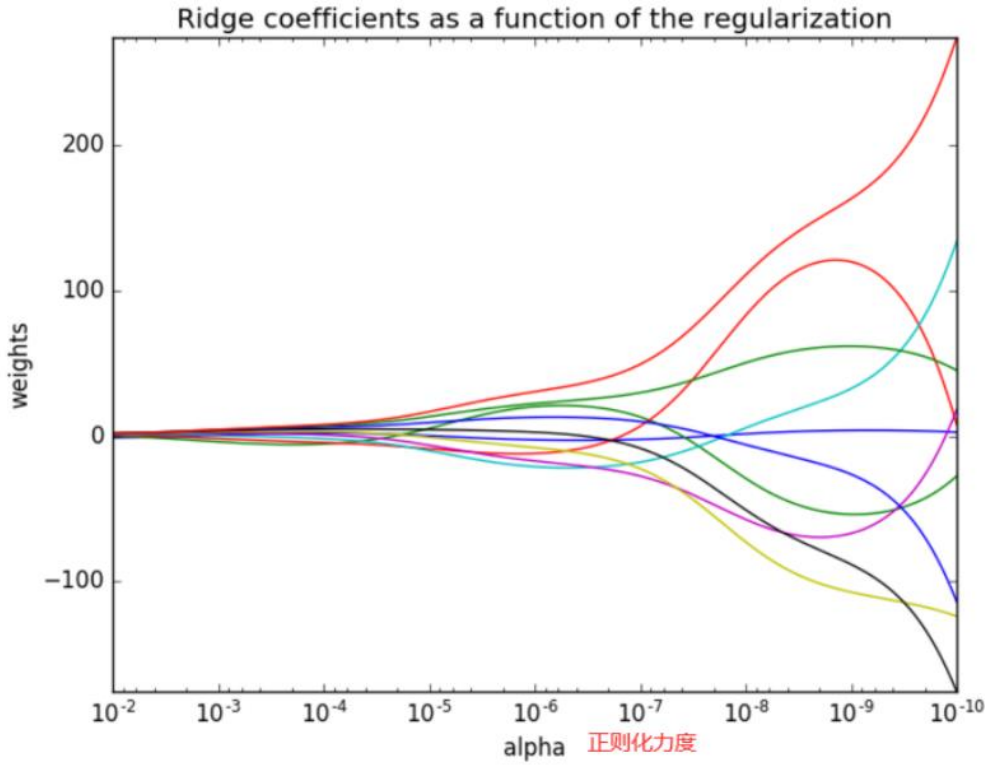
【例】

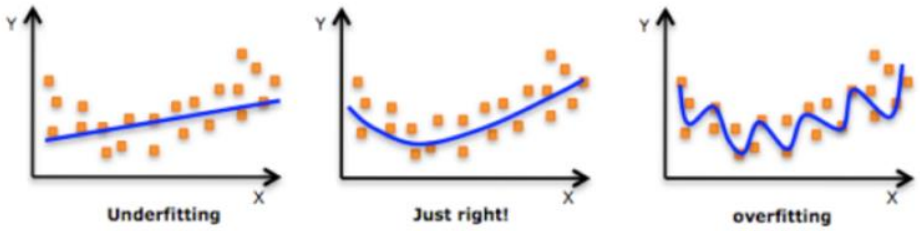
思考：使用机器学习，考虑的一个重点就是泛化能力

- **那么模型越复杂好，还是越简单好呢？**
- **而此时的泛化能力对应的又是什么样的呢？**

复杂，泛化↑，训练精度↓，测试精度↑

2. 4 正则化分析



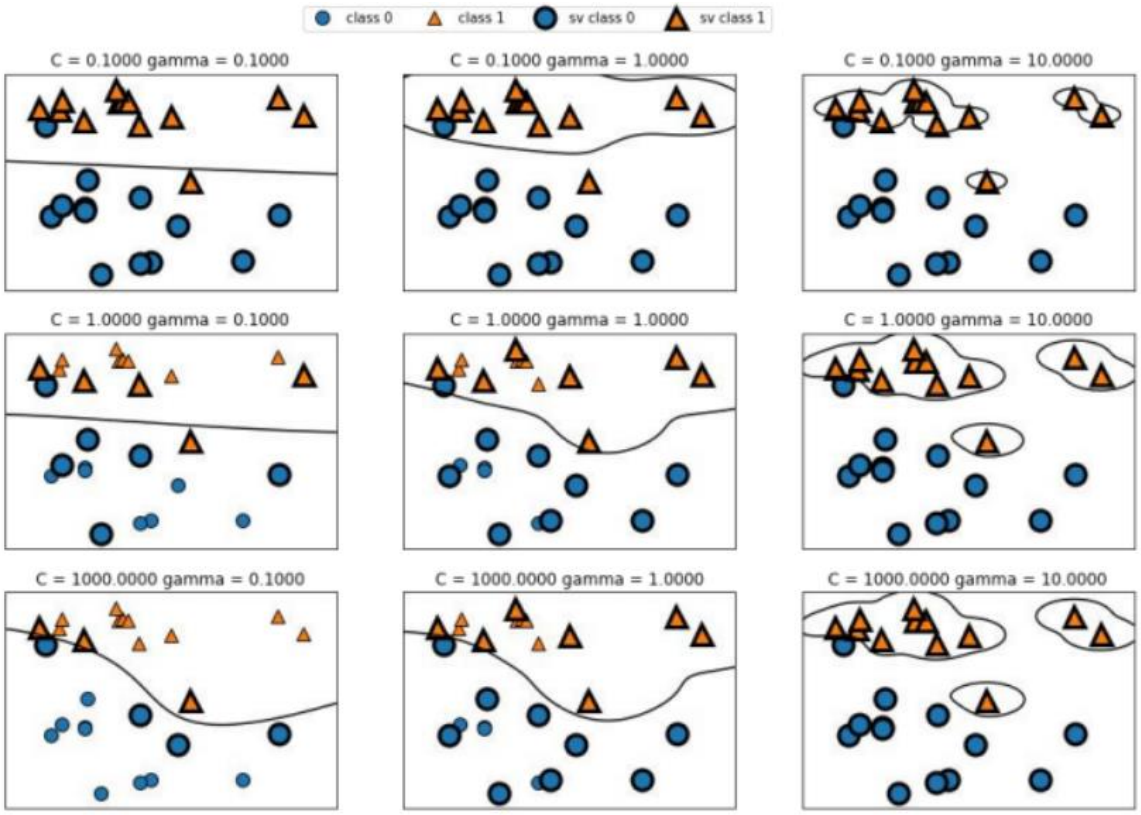


$\theta_0 + \theta_1 x$ $\theta_0 + \theta_1 x + \theta_2 x^2$ $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

正则化力度越强，高次项系数值越小，模型变得简单，从而不容易过拟合。

正则化力度越弱，则与之相反。

2.5 SVM的C和gamma值变化分析



参数C

影响支持向量与决策平面之间的距离。

C值大，正则化弱，决策边界越弯曲，分类越严格，易导致过拟合。

C值小，正则化强，决策边界趋于线性，有更大的错误容忍度，泛化能力增强。

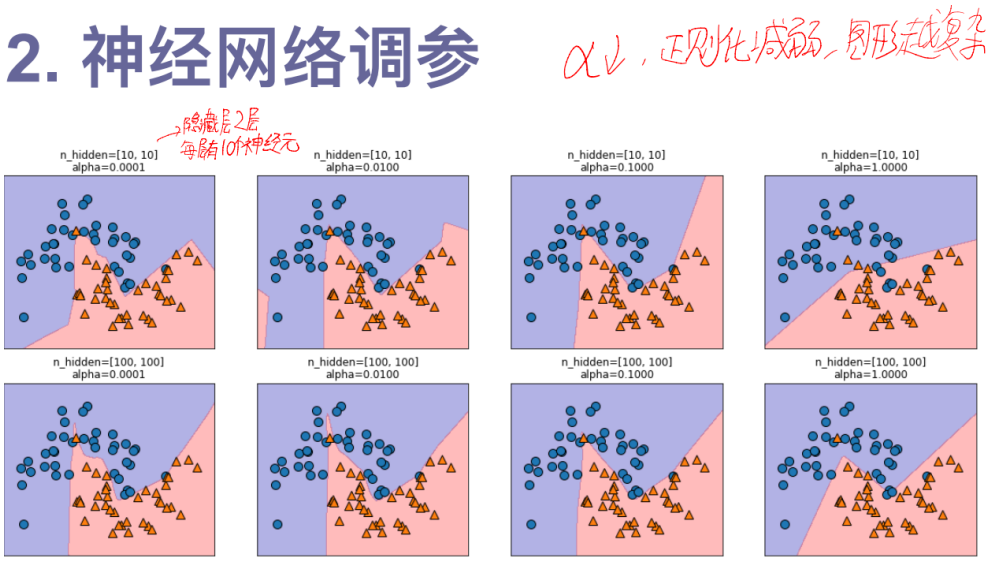
参数gamma

用于将低维数据其映射到高维空间，使得数据在高维空间中线性可分。

gamma值大，模型变复杂，分类越精准。

gamma值小，决策边界变化慢，模型复杂度较低。

2. 6 神经网络 α 值及隐藏层个数变化分析



不同隐单元个数与alpha参数的不同设定下的决策函数

隐藏层数量适中，模型在训练集和测试集上的表现都比较好。

隐藏层数量过少，无法很好地拟合训练数据，导致欠拟合。

隐藏层数量过多，训练过程变得困难，训练集表现很好，但测试集表现较差，导致过拟合。

三、 简答题

3.1 监督学习/无监督学习区别

3.1.1 监督学习

利用一组已知类别的样本，训练模型的参数，使其达到所要求性能的过程。

从标记的训练数据来推断一个功能的机器学习任务。

目标是对对未见过的新数据做出准确预测。

3.1.2 无监督学习

利用无标签的数据学习数据的分布或数据与数据之间的关系。

缺乏足够的先验知识，难以人工标注类别。

进行人工类别标注的成本太高。

3.1.3 应用场景

对图像中的行人和汽车进行分类检测

监督学习会提前标记一些已知的行人和汽车样本，告诉机器什么是行人，什么是汽车。而无监督学习则不会提前标记，机器会在未知的情况下开始。

列举常见的监督式学习任务

线性回归、逻辑回归、K近邻算法、感知机、支持向量机(SVM)、决策树、随机森林、朴素贝叶斯。

垃圾邮箱检测、广告点击率、是否患病，金融诈骗、虚假账号

使用监督学习。属于二分类算法，使用逻辑回归解决。

3.2 分类问题与回归问题的区别

分类问题

目标是预测类别标签。

回归问题

目标是预测一个连续值。

输出不同

分类：输出物体所属的类别，是离散的

回归：输出具体的值，是连续的

目的不同

分类：寻找决策面

回归：寻找最优拟合曲线

3.3 泛化、过拟合、欠拟合

3.3.1 泛化

泛化能力是指模型对新样本的适应能力，亦即预测新样本的准确性

3.3.2 过拟合

模型过于复杂，拟合时过分关注训练集的细节。在训练集拟合效果好，在测试集拟合效果差。模型越复杂，近似误差越小，估计误差越大。

过拟合常见原因

建模样本选取有误，比如样本数量太少，样本标签错误。

样本噪音干扰过大。

参数太多，模型过于复杂。

过拟合解决方法

重新清洗数据集。

增大数据的训练量。

正则化(通过限制高次项的系数防止过拟合)。

减少特征维度。

3.3.3 欠拟合

模型过于简单，拟合时对数据没有做到充分考虑，拟合程度不高。在训练集拟合效果差，在测试集拟合效果差。

解决方法

继续学习

添加多项式特征

添加其他特征项

3.4 数据特征工程三个步骤

特征提取

通过转换函数将特征数据转换为更加适合算法模型的特征数据。

特征预处理

将任意数据转换为机器学习的数字特征。数值型数据可借助标准缩放、归一化、标准化。

特征降维

降低随机变量的个数。

3.5 为什么要进行归一化

特征的单位或大小相差较大，或者某特征的方差相比其他的特征要大出几个数量级，对目标结果影响大，使得一些算法无法学习到其他特征。

3.6 最小-最大规范化及标准化的特点

3.6.1 最大-最小规范化

最大值和最小值是变化的

最大值与最小值容易受异常点影响

鲁棒性差

适合传统精确小数据场景

3.6.2 标准化

受异常值影响较小

适合现代嘈杂大数据场景

3.7 K近邻算法

3.7.1 KNN算法思想

判断未知样本的类别。以全部训练样本作为代表点，计算未知样本与所有训练样本的距离，并以最近邻者的类别作为决策未知样本类别的唯一依据。

3.7.2 KNN基本思路

选择未知样本一定范围内确定个数的K个样本，若K个样本大多数属于某一类型，则未知样本判定为该类型。

3.7.3 KNN核心步骤

STEP1：载入数据，初始化k值。

STEP2：预测类别，遍历每一个训练样本，计算测试数据到每一个训练样本数据的距离。

STEP3：按距离排序，取前k个最相近的训练数据。

STEP4：取频率最高的类别作为测试数据的预测类别。

3.7.4 KNN算法优缺点

优点

容易理解，无需过多调参，仅有的两个重要参数：邻居数量K、选择何种距离度量方法，通常用欧式距离。

缺点

训练集很大，预测速度慢，计算量大。

受限于K值的选择。K值大发生过拟合，异常点对结果影响大；K值小发生欠拟合，受到样本均衡的问题。

3.8 近似误差与估计误差

近似误差

是对现有训练集的训练误差，关注训练集。近似误差过小，可能出现过拟合，没有接近最佳模型。

估计误差

是对测试集的测试误差，关注测试集。估计误差小，模型对未知数据的预测能力好，接近最佳模型。

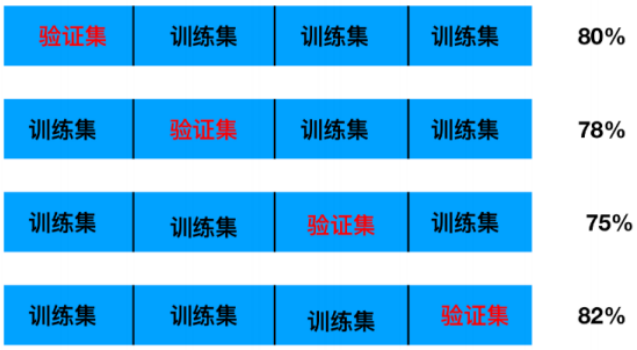
3.9 交叉验证与网格搜索

3.9.1 为什么用（K折）交叉验证

将拿到的训练数据分为训练集和验证集。将数据分成K份，其中一份作为验证集。然后经过K次的测试，每次都更换不同的验证集。可得到K组模型的结果，取平均值作为最终结果。

不能提高模型的准确率，但可以提高模型可信度，即找到更优的K值。

助记



3.9.2 为什么用网格搜索

网格搜索把超参数的值，通过字典的形式传递，然后进行选择最优值。

能够选择和调优参数。

3.10 什么是线性回归

线性回归是利用回归方程对一个或多个自变量和因变量之间关系进行建模的一种分析方式。

3.11 L1 正则化和L2 正则化的区别

L1 正则化

使得其中某些w的值直接等于 0，删除这些特征的影响。

适用于Lasso回归。

L2 正则化

使得某一些w的值都很小，都接近于 0，削弱某个特征的影响，防止过拟合。

适用于Ridge回归。

3.12 决策树

3.12.1 什么是决策树

决策树是一种树形结构，由决策结点、分支和叶子结点组成。

决策节点表示在样本的一个属性上进行的划分。

分支表示对于决策结点进行划分的输出。

叶结点代表经过分支到达的类。

通过把数据样本分配到某叶子结点，确定数据集中样本所属的分类。

3.12.2 构造决策树的步骤

构造一系列的if/else问题，这些问题称为测试。算法搜遍所有可能的测试，找出对目标变量来说信息量最大的那一个，作为一个结点。

根据是否满足该结点的条件分为左右子结点，对当前数据进行划分。

反复递归，直到所有叶结点都只包含单一类别，称为纯的。

3.12.3 节点选择算法区别

名称	分支方式	备注
ID3	信息增益	只能对离散值的数据集构成决策树
C4.5	信息增益率	优化后解决ID3分支过程中总喜欢偏向选择值较多的属性

PS：C4.5 可处理连续属性，但只适合于能够驻留内存的数据集。

CART	Gini系数	可以进行分类和回归，可以处理离散属性，也可以处理连续属性

3.12.4 决策树如何防止过拟合

随着树的增长，测试集的精度会先升后降，导致过拟合的出现，可采用预剪枝和后剪枝。

预剪枝

控制树的最大深度。

控制叶节点的最大个数。

限制每个节点所包含的最小样本数目为x，如果该节点总样本小于x，则不再分

后剪枝

在已生成过拟合决策树上进行剪枝。

3.12.5 决策树优缺点

优点

易理解，易可视化。

不用对特征进行预处理。

缺点

泛化能力差，即使预剪枝，也容易过拟合。

3.13 决策集成

3.13.1 集成算法优势

集成多个算法，用不同的算法对同一组数据进行分析，投票决定各个算法的最好的结果。

优化训练数据，防止欠拟合。

提升泛化性能，防止过拟合。

3.13.2 随机森林算法思想

用随机的方式建立一个森林，由很多的决策树组成，每棵树相互独立。

随机化方法是：随机选取样本、随机选取特征。

分类问题采用软投票策略，少数服从多数；回归问题则是对结果取平均值。

可提高模型泛化能力，降低过拟合。

3.13.3 梯度提升回归树算法思想

利用梯度下降思想，使用损失函数的负梯度在当前模型的值，作为提升树中残差的近似值。每棵树相互关联。

3.14 支持向量机（SVM）算法思想

寻找最优决策边界，使数据集的边缘点到分界面的距离最大。

若线性不可分，则用核函数，增加维度，减少计算代价。

3.15 降维与特征提取

3.15.1 降维应用场景

回归分析。

聚类分析。

3.15.2 两种降维方式

特征选择：特征是否发散，特征与目标的相关性。

主成分分析。

3.15.3 为什么要降维

降低数据维度。

降低算法计算开销。

去除噪声。

使数据集更易使用。

使结果易理解。

3.16 主成分分析（PCA）

3.16.1 PCA算法思想

将n维特征映射到k维上(k<n)，这k维是全新的正交特征也被称为主成分，是在原有n维特征的基础上重新构造出来的k维特征。

根据新特征对解释数据的重要性来选择它的一个子集。

3.16.2 PCA的目标

找到一个轴，使得样本空间的所有点映射到这个轴后，方差最大。

使预测结果的均方误差（MSE）尽量小。

3.16.3 为什么要舍弃部分信息

降低数据维度。

降低算法计算开销。

去除噪声。

避免过拟合。

3.16.4 t-SNE算法思想

t-SNE适用于高维空间到低维空间的函数映射不是线性关系的情况。通过保持相似样本之间的距离关系，将高维数据映射到低维空间。使用概率模型来表达相似性，通过t分布计算低维空间的点之间的相似度。

3.17 K-means聚类

3.17.1 K聚类目标

聚类将一组数据分为几个簇，使得同一个簇内的数据非常相似，不同簇内的数据点非常不同。

3.17.2 聚类与分类的区别

聚类算法属于无监督学习算法，分类算法属于监督学习算法。

聚类：将一组数据分为几个簇，同一个簇内的数据非常相似，不同簇内的数据点非常不同。

分类：将数据分为预定义的类别。

3.17.3 K-means算法的局限性

对离群点和孤立点敏感。

给定数据集中簇的正确个数，k均值可能找不到正确聚类。

每个簇都是凸形，只能找到相对简单的形状。

四、 计算题

4. 1 分类模型评估指标计算

混淆矩阵

- 对于二分类模型，预测情况与实际情况会得出2*2=4种组合，形成**混淆矩阵**

	预测正例	预测反例
实际正例	TP: True Positive	FN: False Negative
实际反例	FP: False Positive	TN: True Negative

- P(positive)　　N(Negitive)
- TP：本身是正类，结果划分是正类
- FP：本身是负类，结果划分为正类
- TN：本身是负类，结果划分为负类
- FN：本身是正类，结果划分为负类

TN：预测正确，预测结果为负类.

- 1.准确率：所有正确的分类的比重

  Accuracy=（TP+TN）/（TP+TN+FN+FP）
- 2. 查准率： 正类且被预测也是正类的比率占所有预测为正类的占比

  Precision = TP/（TP+FP）
- 3. 召回率Recall: 预测的正确的正类占实际正类的占比

  Recall = TP /（TP+FN）
- TP和TN是相互影响的，TP是正类预测也是正类，如果这个TP的值越大，TN的值肯定是下降的。根据查准率和召回率的公式，TP越大，查准率越高，但是召回率就不一定了。我们用F-score 来描述两者的关系：

$$F_{Score} = (1+\beta^2)\frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}$$

- β如果取1,表示Precision与Recall一样重要
- β如果取小于1,表示Precision比Recall重要
- β如果取大于1,表示Recall比Precision重要

4.2 数据规范化（归一化）

4.2.1 最小-最大规范化

$$X' = \frac{x - min}{max - min} \qquad X'' = X' * (mx - mi) + mi$$

max、 min：分别表示一列中最大值和最小值

mx 、mi：分别表示指定区间的最大值和最小值

4.2.2 标准化

通过对原始数据进行变换把数据变换为均值为0，标准差为1的范围内

$$X' = \frac{x - \text{mean}}{\sigma}$$

标准差：$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$

4.2.3 例题

【例】归一化原因

消除特征间的尺度差异：不同特征往往具有不同的数值范围和尺度。这种尺度差异可能导致某些特征在模型中的影响力远远大于其他特点，从而影响模型的性能。通过归一化，可以将所有特征的尺度统一，确保模型能公平对待每一个特征。

【例】

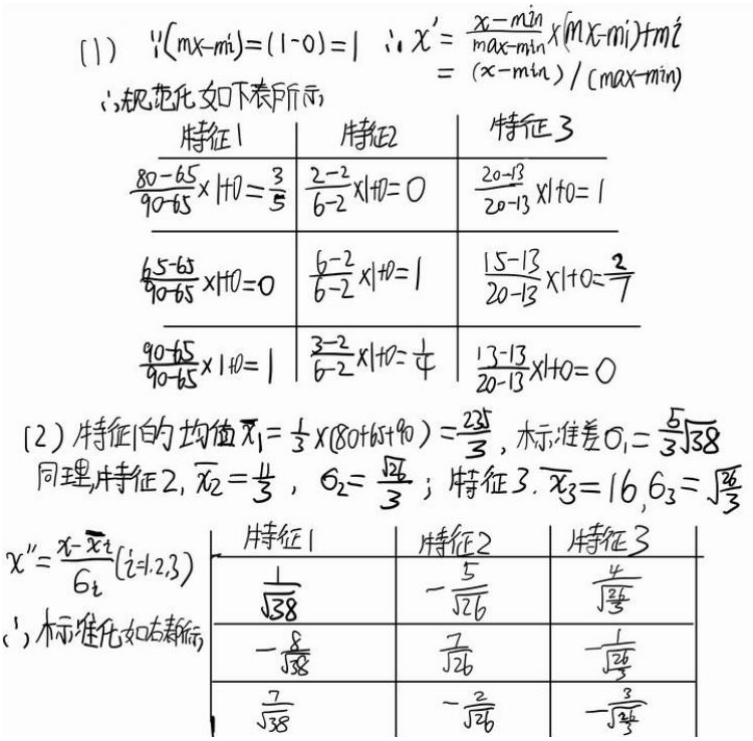
(知识点：数据归一化)已知3个样本，每个样本有3个特征，如下表1所示：

表1

特征1	特征2	特征3
80	2	20
65	6	15
90	3	13

(1) 请采用最小-最大规范化法将上述数据规范到[0,1]之间。

((2) 请采用标准化处理将上述数据变换到均值为0，标准差为1的范围内。



4.3 熵（Entropy）

1.1 概念

- **信息理论：随机变量不确定度的度量。**
- **熵越大，数据的不确定性越高；熵越小，数据的不确定性越低。**反之，p_i↑，H则↓

$$H = -\sum_{i=1}^{k} p_i log(p_i)$$

Σ (sigma)

注意：log是以2为底，lg是以10为底

1.2 案例

手算

{1/3,1/3,1/3}

$$H = -\frac{1}{3}log(\frac{1}{3})-\frac{1}{3}log(\frac{1}{3})-\frac{1}{3}log(\frac{1}{3}) = 1.0986$$

{1/10,2/10,7/10}

$$H = -\frac{1}{10}log(\frac{1}{10})-\frac{2}{10}log(\frac{2}{10})-\frac{7}{10}log(\frac{7}{10}) = 0.8018$$

{1,0,0}

$$H = -1log(1)=0$$

PS：信息熵越小越好

4.4 信息增益(Gain)

4.4.1 公式

$$Gain(D,a) = H(D) - H(D|a)$$

PS：信息增益越大，该属性对标签的影响越大，越选择该属性。

4.4.2 例题

【例】

2.2 案例

已知：第一列：论坛号码，第二列：性别，第三列：活跃度，最后一列：用户是否流失。

问题：性别和活跃度两个特征,哪个对用户流失影响更大?

uin	gender	act_info	is_lost
1	男	高	0
2	女	中	0
3	男	低	1
4	女	高	0
5	男	高	0
6	男	中	0
7	男	中	1
8	女	中	0
9	女	低	1
10	女	中	0
11	女	高	0
12	男	低	1
13	女	低	1
14	男	高	0
15	男	高	0

常考 手算题(15-20')

	positive	negative	汇总
整体	5	10	15
男性	3	5	8
女性	2	5	7
高	0	6	6
中	1	4	5
低	4	0	4

Positive：已流失
Negative：未流失

- 计算类别信息熵—整体熵

$$H(D)=-\frac{5}{15}log(\frac{5}{15})-\frac{10}{15}log(\frac{10}{15})=0.9182$$ →总体信息熵

- 计算性别属性信息增益(a="性别")

$$H(D,\text{性别})=\sum_{v=1}^{V}\frac{D^v}{D}H(D^v)=\frac{D^1}{D}H(D^1)+\frac{D^2}{D}H(D^2)$$

$$H(D^1)=-\frac{3}{8}log(\frac{3}{8})-\frac{5}{8}log(\frac{5}{8})=0.9543$$

$$H(D^2)=-\frac{2}{7}log(\frac{2}{7})-\frac{5}{7}log(\frac{5}{7})=0.8631$$

$$Gain(D,\text{性别})=H(D)-H(D\mid a)=H(D)-\frac{8}{15}H(D^1)-\frac{7}{15}H(D^2)=0.0064$$

- 计算活跃度属性信息增益(a="活跃度")

$$H(D^1)=0$$ = $-\frac{0}{6}log\frac{0}{6}-\frac{6}{6}log\frac{6}{6}$

$$H(D^2)=0.7219$$

$$H(D^3)=0$$ = $-\frac{4}{4}log\frac{4}{4}-\frac{0}{4}log\frac{0}{4}$

$$Gain(D,\text{活跃度})=H(D)-H(D\mid a)=H(D)-\frac{6}{15}H(D^1)-\frac{5}{15}H(D^2)-\frac{4}{15}H(D^3)=0.6776$$

- 活跃度的信息增益比性别的增益大，也就是说，活跃度对用户流失的影响更大**(即采用信息熵小的特征（不确定性小）作为划分节点）**。

4.5 信息增益率

4.5.1 公式

- **增益率：用信息增益Gain(D,a)和属性a对应的"固有值"(intrinsic value)的比值来共同定义。**

$$Gain\_ration(D,a)=\frac{Gain(D,a)}{IV(a)}\qquad IV(a)=-\sum_{v=1}^{V}\frac{D^v}{D}log(\frac{D^v}{D})$$

PS：信息增益率越大，说明该属性纯度提升越高，越选择该属性。

4.5.2 例题

【例】数据和章节 3.4 的例题一致

a. 计算类别信息熵
b. 计算性别属性的信息熵
c. 计算活跃度属性的信息熵

- 计算属性IV

→求属性本身的信息熵

$$IV(\text{性别})=-\frac{7}{15}log(\frac{7}{15})-\frac{8}{15}log(\frac{8}{15})=0.9968$$

$$IV(\text{活跃度})=-\frac{6}{15}log(\frac{6}{15})-\frac{5}{15}log(\frac{5}{15})-\frac{4}{15}log(\frac{4}{15})=1.5656$$

- 计算信息增益率

$$Gain\_ratio(D,\text{性别})=\frac{Gain(D,\text{性别})}{IV(\text{性别})}=\frac{0.0064}{0.9968}=0.0064$$

$$Gain\_ratio(D,\text{活跃度})=\frac{Gain(D,\text{活跃度})}{IV(\text{活跃度})}=\frac{0.6776}{1.5656}=0.4328$$

比之前更小，纯度提升

活跃度的信息增益率更高，构建决策树的时候，优先选择

4.6 基尼值和基尼指数

4.6.1 公式

- **基尼值Gini(D)**：从数据集D中随机抽取两个样本，其类别标记不一致的概率。Gini(D)值越小，数据集D的纯度越高。

$$Gini(D)=\sum_{k=1}^{N}\sum_{k'\neq k}p_k p_{k'}=1-\sum_{k=1}^{N}p_k^2$$

- **基尼指数Gini_index(D)**：一般选择使划分后基尼指数最小的属性为最优划分属性。

$$Gini\_index(D,a)=\sum_{v=1}^{V}\frac{D^v}{D}Gini(D^v)$$

4.6.2 例题

【例】

序号	是否有房	婚姻状况	年收入	是否拖欠贷款
1	yes	single	125k	no
2	no	married	100k	no
3	no	single	70k	no
4	yes	married	120k	no
5	no	divorced	95k	yes
6	no	married	60k	no
7	yes	divorced	220k	no
8	no	single	85k	yes
9	no	married	75k	no
10	no	single	90k	yes

- 根据是否有房来进行划分

$$Gini(\text{左节点})=1-(\frac{0}{3})^2-(\frac{3}{3})^2=0$$

$$Gini(\text{右节点})=1-(\frac{3}{7})^2-(\frac{4}{7})^2=0.4898$$

$$Gini\_index(D,\text{是否有房})=\frac{7}{10}*0.4898+\frac{3}{10}*0=0.343$$

		是否有房	
		yes	no
是否拖欠贷款	yes	0	3
	no	3	4

- 根据婚姻状况来划分
  - {married}|{single,divorced}

$$Gini\_index(D,\text{婚姻状况})=\frac{4}{10}*0+\frac{6}{10}*[1-\left(\frac{3}{6}\right)^2-\left(\frac{3}{6}\right)^2]=0.3$$

  - {single}|{married,divorced}

$$Gini\_index(D,\text{婚姻状况})=\frac{4}{10}*0.5+\frac{6}{10}*[1-\left(\frac{1}{6}\right)^2-\left(\frac{5}{6}\right)^2]=0.367$$

  - {divorced}|{single, married}

$$Gini\_index(D,\text{婚姻状况})=\frac{2}{10}*0.5+\frac{8}{10}*[1-\left(\frac{2}{8}\right)^2-\left(\frac{6}{8}\right)^2]=0.4$$

- 根据年收入来划分

对于年收入属性为数值型属性，首先需要对数据按升序排序，然后从小到大一次用相邻的中间值作为分隔将样本划分为两组。

节点为87.5时：$Gini\_index(D,\text{收入}) = \frac{6}{10}*[1-(\frac{3}{6})^2-(\frac{3}{6})^2] + \frac{4}{10}*(\quad) = \sim$

是否拖欠贷款	no	no	no	yes	yes	yes	no	no	no	no
年收入	60	70	75	85	90	95	100	120	125	220
相邻值中点	/	65	72.5	80	87.5	92.5	97.5	110	122.5	172.5	/
Gini_index	/	0.4	0.375	0.343	0.417	0.4	0.3	0.343	0.375	0.4	/

节点为65时：$Gini\_index(D,\text{收入}) = \frac{1}{10}*0+\frac{9}{10}*[1-\left(\frac{6}{9}\right)^2-\left(\frac{3}{9}\right)^2]=0.4$

$(1-(\frac{1}{1})^2)$

【例】

1. 计算题

假设有一个数据集包含8个样本，每个样本有三个属性：年龄、性别和收入。数据集如下：

样本	年龄	性别	收入	是否买车（是/否）
1	青年	男	6	是
2	青年	女	5	是
3	中年	男	12	是
4	中年	男	10	是
5	青年	女	4	否
6	青年	男	7	否
7	中年	男	11	是
8	中年	女	13	是

请使用Gini指数作为评价指标来使用决策树算法时的划分属性是哪个？

①根据年龄划分

		年龄	
		青年	中年
是否买车	是	2	4
	否	2	0

$Gini\_index(D,\text{年龄}) = \frac{6}{8}\times[1-(\frac{4}{6})^2-(\frac{2}{6})^2] + \frac{2}{8}\times[1-(\frac{2}{2})^2-(\frac{0}{2})^2]$

$= \frac{4}{9}$

②根据性别划分

		性别	
		男	女
是否买车	是	4	2
	否	1	1

$Gini\_index(D,\text{性别}) = \frac{6}{8}\times[1-(\frac{4}{6})^2-(\frac{2}{6})^2] + \frac{2}{8}\times[1-(\frac{1}{2})^2-(\frac{1}{2})^2]$

$= \frac{11}{30}$

③根据收入划分

是否买车　否　是　是　否　是　是　是　是

收入　4　5　6　7　10　11　12　13

相邻中点　4.5　5.5　6.5　8.5　10.5　11.5　12.5

Gini_index　$\frac{3}{14}$　$\frac{9}{28}$　$\frac{11}{30}$　$\frac{4}{9}$　$\frac{3}{10}$　$\frac{1}{3}$　$\frac{5}{14}$

节点为4.5时，$Gini\_index(D,\text{收入}) = \frac{1}{8}\times[1-(1)^2] + \frac{7}{8}\times[1-(\frac{6}{7})^2-(\frac{1}{7})^2] = \frac{3}{14}$

∴收入的Gini指数最低，是$\frac{3}{14}$

∴选"收入"属性进行划分。

4.7 贝叶斯（朴素贝叶斯）分类器

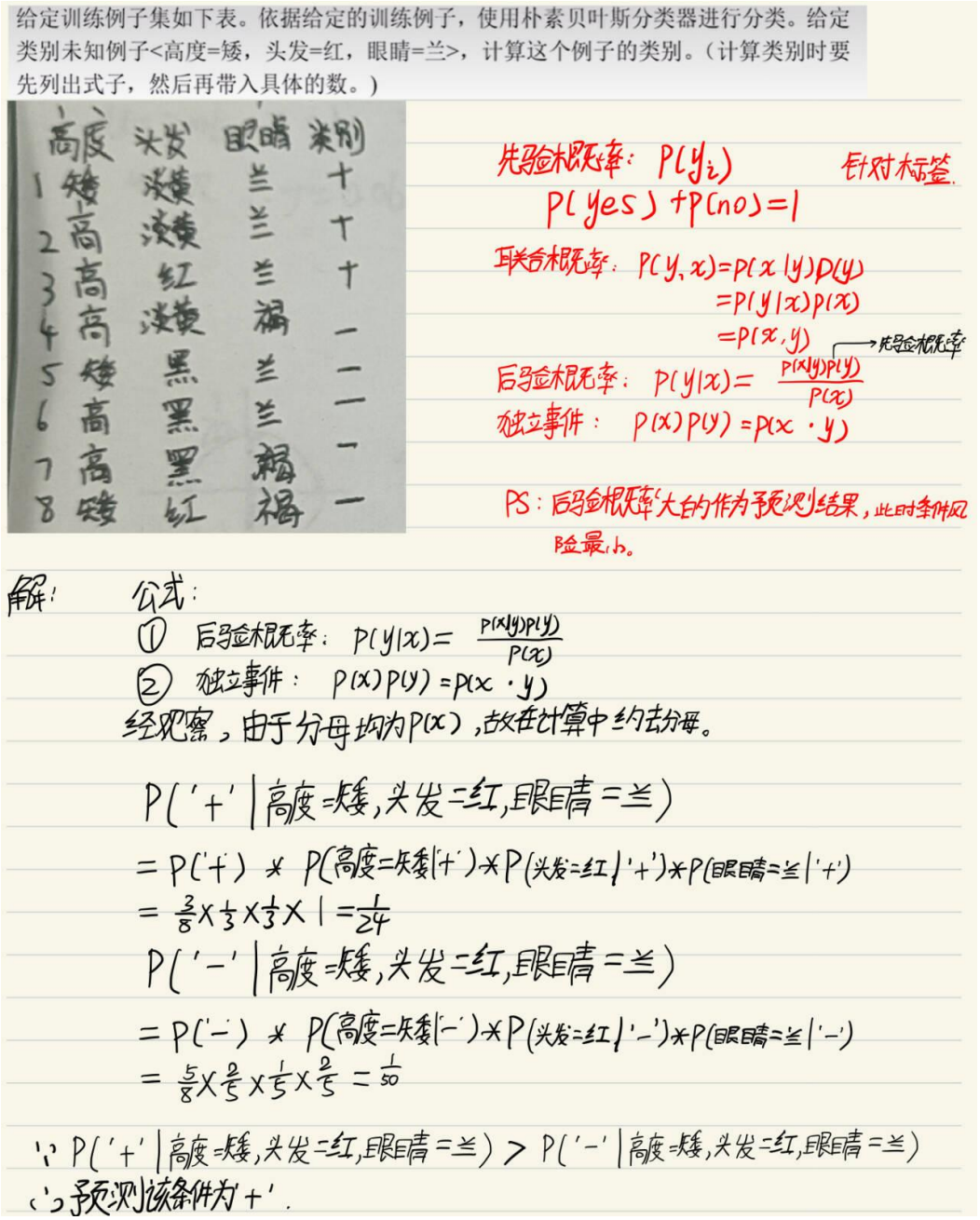
【例】

西瓜数据集如表所示，包括四个属性：色泽、根蒂、纹理、脐部以及目标变量是否是好瓜。试采用贝叶斯分类器来判断四个属性中哪个属性为最优划分属性。

编号	色泽	根蒂	纹理	脐部	好瓜
1	青绿	蜷缩	清晰	凹陷	是
2	浅白	硬挺	清晰	平坦	是
3	乌黑	稍蜷	稍糊	稍凹	否

解：计算每个属性的信息增益，然后选择信息增益最大的属性。

【例】



【例】

编号	色泽	根蒂	敲声	纹理	好瓜
1	青绿	蜷缩	浊响	清晰	是
2	乌黑	蜷缩	沉闷	清晰	是
3	乌黑	蜷缩	浊响	清晰	是
4	青绿	蜷缩	沉闷	清晰	是
5	浅白	蜷缩	浊响	清晰	是
6	青绿	稍蜷	浊响	清晰	是
7	乌黑	稍蜷	浊响	稍糊	是
8	乌黑	稍蜷	浊响	清晰	是
9	乌黑	稍蜷	沉闷	稍糊	否
10	青绿	硬挺	清脆	清晰	否
11	浅白	硬挺	清脆	模糊	否
12	浅白	蜷缩	浊响	模糊	否
13	青绿	稍蜷	浊响	稍糊	否
14	浅白	稍蜷	沉闷	稍糊	否
15	乌黑	稍蜷	浊响	清晰	否
16	浅白	蜷缩	浊响	模糊	否
17	青绿	蜷缩	沉闷	稍糊	否

（青绿，稍蜷，浊响，清晰）？？

$$y = f(x) = \underset{c_k}{argmax}\, P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} \mid Y = c_k)$$

p(是|青绿，稍蜷，浊响，清晰)=p(是)*p(青绿|是)*p(稍蜷|是)*p(浊响|是)*p(清晰|是)=8/17*3/8*3/8*6/8*7/8

先验概率 条件概率

p(否|青绿，稍蜷，浊响，清晰)=p(否)*p(青绿|否)*p(稍蜷|否)*p(浊响|否)*p(清晰|否)=9/17*3/9*4/9*4/9*2/9

【例】

1. 计算题

假设有一个数据集包含10个样本，每个样本有三个属性：体重（正常/偏重）、身高（矮/中等/高）、年龄（青年/中年/老年），类别为是否患病。数据集如下：

序号	体重	身高	年龄	类别
1	正常	矮	青年	是
2	偏重	中等	中年	是
3	正常	高	老年	否
4	偏重	中等	老年	否
5	偏重	矮	青年	是
6	正常	中等	中年	是
7	偏重	高	老年	否
8	正常	高	中年	否
9	正常	中等	老年	否
10	偏重	高	青年	是

现在要使用朴素贝叶斯分类器对一个新样本进行分类，新样本的属性是：体重为偏重、身高为高、年龄为青年。请根据以上数据集计算朴素贝叶斯分类器的分类结果。

解：

$P(是 \mid 偏重, 高, 青年) = P(是) * P(偏重|是) * P(高|是) * P(青年|是)$
$= 5/10 * 3/5 * 1/5 * 3/5$
$= 0.036$

$P(否 \mid 偏重, 高, 青年) = P(否) * P(偏重|否) * P(高|否) * P(青年|否)$
$= 5/10 * 2/5 * 3/5 * 0$
$= 0$

4.8 K-means聚类

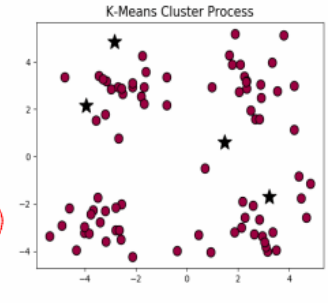
4.8.1 聚类步骤与例题

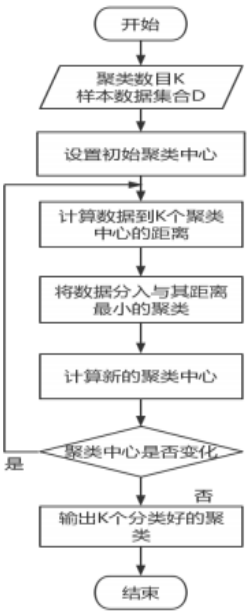
1. 随机设置K个特征空间内的点作为初始的聚类中心

2. 对于其他每个点计算到K个中心的距离，未知的点选择最近的一个聚类中心点作为标记类别。

3. 接着对着标记的聚类，重新计算出每个聚类的新中心点(**平均值**)。新中心点坐标 $\left(\frac{x_1+x_2+\cdots+x_n}{n}, \frac{y_1+y_2+\cdots+y_n}{n}\right)$

4. 如果计算得出的新中心点与原中心点一样，那么结束，否则重新进行第二步步骤。⇒指 簇包含的内容保持不变，不代表中心点坐标不变。





PS：若初始聚类中心未给出，则先画图观察，再假设中心点，可简化聚类迭代次数。

【例】

2.假设有以下7个样本点的坐标数据：样本点1: (2, 3)，样本点2: (3, 3)，样本点3: (1, 2)，样本点4: (5, 4)，样本点5: (6, 5)，样本点6: (6, 4)，样本点7: (7, 6)，请使用K-means聚类算法将这7个样本点分成两个簇，给出聚类迭代次数和每个簇的中心点坐标。

依次计样本点1~7为P1~P7

第1次聚类：假设选 P1、P2 为初始聚类中心、

计算中心：

	P1	P2
P3	1.4142	2.2361
P4	3.1623	2.2361
P5	4.4721	3.6056
P6	4.3589	3.1623
P7	5.8310	5

第1次聚类结果是：

簇A：P1，P3

簇B：P2，P4，P5，P6，P7

∴ A、B两组新的聚类中心分别为：P8(1.5, 2.5)，P9(5.4, 4.4)

第2次聚类：

计算中心：

	P8	P9
P1	0.7071	3.6770
P2	1.5811	2.7785
P3	0.7071	5.0120
P4	3.8079	0.5657
P5	5.1478	0.8485
P6	4.7434	0.7211
P7	6.5192	2.2627

第2次聚类结果是：

簇A：P1，P2，P3

簇B：P4，P5，P6，P7

∴ A、B两组新的聚类中心分别为：P10(2, $\frac{8}{3}$)，P11(6, 4.75)

第3次聚类：

计算中心：	P10	P11
P1	0.6667	4.3661
P2	1.0541	3.4731
P3	1.2019	5.7064
P4	3.2830	1.25
P5	4.6308	0.25
P6	4.2164	0.75
P7	6.0093	1.6008

P11(6,4.75)

第3次聚类结果是：

簇A：P1, P2, P3

簇B：P4, P5, P6, P7

由于第3次的聚类结果与第2次相同，所以聚类结束，最终结果为：

簇A：P1, P2, P3

簇B：P4, P5, P6, P7

簇A的中心坐标为 $(2, \frac{8}{3})$

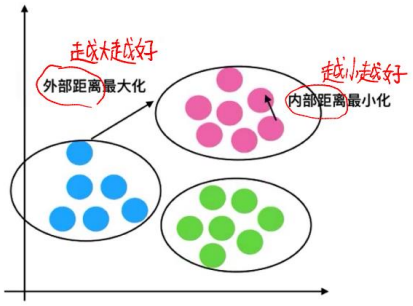
簇B的中心坐标为 (6, 4.75)

聚类迭代次数为3.

4.8.2 轮廓系数与例题

1. 轮廓系数

$$sc_i = \frac{b_i - a_i}{\max(b_i, a_i)}$$



- bi为样本i到其他族群的所有样本的平均值；ai为样本i到本身簇的距离平均值。
- 取值范围(-1,1)

如果bi>>ai:趋近于1效果越好，若bi<< ai:趋近于-1效果越不好

越大越好

【例】

例：使用聚类轮廓系数的方法，对k-means算法所举的例子进行度量。

	X	Y
P_1	0	0
P_2	1	2
P_3	3	1
P_4	8	8
P_5	9	10
P_6	10	7

簇A：P1、P2、P3

簇B：P4、P5、P6

(1)分别计算p1与p2和p3的距离，并计算平均值：

a（p1）=(2.24+3.16)/2=2.7

(2)分别计算p1与p4、p5、p6之间的距离，并计算平均值：

b(p1)=(11.31+13.45+12.20)/3=12.32

(3)计算p1的轮廓系数：

s(p1)=(12.32-2.7)/12.32=0.78

(4)同理，计算p2,p3的轮廓系数分别为：

s(p2)=(10.28-2.24)/10.28=0.78

s(p3)=(9.55-2.7)/9.55=0.71

(5)计算簇A中的轮廓系数的平均值：

s=(0.78+0.78+0.71)/3=0.76

结果：簇内紧凑，不同簇距离较远

五、 代码题

5.1.1 算法思想

实验三：线性回归应用（两次实验：回归、分类）

1．实验目的

A. 理解并掌握线性回归的基本算法

B. 掌握数据加载、模型训练以及绘图的基本方法。

2．实验内容和要求

A. 下载或者导入 wave 数据集，训练线性回归模型（lasso,ridge），比较不同正则化并输出模型的预测结果，**并保存模型**。

B. 下载或者导入波士顿房价数据集，在不同 apha 值情况下，对比训练 lasso 模型，**并保存模型后加载模型，**输出预测结果，并绘制出不同模型的系数对比图（图可以打印）。

实验六：SVM 算法应用（实验报告）

1．实验目的

A. 理解并掌握 SVM 的基本算法

B. 掌握利用 SVM 进行分类和回归。

C. 掌握创建虚拟数据集的方法。

D. 掌握数据加载、模型训练以及绘图的基本方法。

2．实验内容和要求

A. 下载鸢尾花数据集，分别使用 SVM 来进行分类，要求总结出参数 C，gama 值在 SVM 中的作用并能够绘制出决策边界。

B. 下载或者导入波士顿房价预测数据集，利用 SVM 的回归模型(SVR)实现房价预测，要求对采用不同的核函数效果进行对比。