《机器学习》复习题

一、单项选择题

1、一监狱人脸识别准入系统用来识别待进入人员的身份，此系统一共包括识别 4 种不同的人员：狱警，小偷，送餐员，其他。下面哪种学习方法最适合此种应用需求：

A. K-means 聚类问题　　B. 回归问题　　C.二分类问题　　√D.多分类问题

2、数据科学家可能会同时使用多个算法（模型）进行预测， 并且最后把这些算法的结果集成起来进行最后的预测（集成学习），以下对集成学习说法正确的是

A.　单个模型之间有高相关性

√B. 单个模型之间有低相关性

C. 在集成学习中使用“平均权重”而不是“投票”会比较好

D. 单个模型都是用的一个算法

3、bootstrap 数据的含义是

A.　有放回的从整体 M 中抽样 m 个特征

B.　无放回的从整体 M 中抽样 m 个特征

√C. 有放回的从整体 N 中抽样 n 个样本

D.无放回的从整体 N 中抽样 n 个样本

4、对于在原空间中线性不可分问题，支持向量机（）。

A. 在原空间中寻找非线性函数的划分数据

B. 无法处理

C. 在原空间中寻找线性函数划分数据

√D. 将数据映射到核空间中

5、回归问题和分类问题的区别是？

A. 回归问题有标签，分类问题没有

B. 回归问题输出值是离散的，分类问题输出值是连续的

√C. 回归问题输出值是连续的，分类问题输出值是离散的

C.　回归问题与分类问题在输入属性值上要求不同

6、以下关于降维的说法不正确的是？

A.降维是将训练样本从高维空间转换到低维空间

√B.降维不会对数据产生损伤

C.通过降维可以更有效地发掘有意义的数据结构

D.降维将有助于实现数据可视化

7、以下关于训练集、验证集和测试集说法不正确的是( )。

A.　测试集是纯粹是用于测试模型泛化能力

B.　√ 训练集是用来训练以及评估模型性能

C.　验证集用于调整模型参数

D.　以上说法都不对

8、下列哪种方法可以用来缓解过拟合的产生：( )。

A.增加更多的特征

√B.正则化

C.增加模型的复杂度

D.以上都是

9、下面哪个情形不适合作为 K-Means 迭代终止的条件？

A.　前后两次迭代中，每个聚类中的成员不变

√B. 前后两次迭代中，每个聚类中样本的个数不变

C. 前后两次迭代中，每个聚类的中心点不变

10、下列说法正确的是[A]

A 每一轮提高错误样本的权重。

B 每一轮降低错误样本的权重。

C Bagging 每一轮使用的训练集相同。

D Boosting 的每一轮之间可以并行。

11、K-means 无法聚以下哪种形状样本？（B ）

A. 圆形分布　　B. 螺旋分布　　C. 带状分布　　D. 凸多边形分布

12、属于监督学习的机器学习算法是(A)

A. 贝叶斯分类器　　B. 主成分分析

C. K-Means　　D. 高斯混合聚类

13、属于无监督学习的机器学习算法是(C)

A. 支持向量机　　B. Logistic 回归

C. 层次聚类　　D. 决策树

14、朴素贝叶斯分类器的特点是(C)

A. 假设样本服从正态分布　　B. 假设样本服从多项式分布

C. 假设样本各维属性独立　　D. 假设样本各维属性存在依赖

15、下列属于线性分类方法的是(B)

A. 决策树　　B. 感知机　　C. 最近邻　　D. 集成学习

16、SVM 的原理的简单描述，可概括为(C)

A. 最小均方误差分类　　B. 最小距离分类

C. 最大间隔分类　　D. 最近邻分类

17、以下对支持向量机中的支撑向量描述正确的是(C)

A. 最大特征向量　　B. 最优投影向量

C. 最大间隔支撑面上的向量　　D. 最速下降方向

18、关于决策树节点划分指标描述正确的是(B)

A. 类别非纯度越大越好　　B. 信息增益越大越好

C. 信息增益率越小越好　　D. 基尼指数越大越好

19、以下描述中，属于决策树策略的是(D)

A. 最优投影方向　　B. 梯度下降方法

C. 最大特征值　　D. 最大信息增益

20、集成学习中基分类器的选择如何，学习效率通常越好(D)

A. 分类器相似　　B. 都为线性分类器

C. 都为非线性分类器　　D. 分类器多样，差异大

21. 集成学习中，每个基分类器的正确率的最低要求(A)

A. 50%以上　　B. 60%以上　　C. 70%以上　　D. 80%以上

22. 下面属于 Bagging 方法的特点是(A)

A. 构造训练集时采用 Bootstraping 的方式

B. 每一轮训练时样本权重不同

C. 分类器必须按顺序训练

D. 预测结果时，分类器的比重不同

23. 下面属于 Boosting 方法的特点是(D)

A. 构造训练集时采用 Bootstraping 的方式

B. 每一轮训练时样本权重相同

C. 分类器可以并行训练

D. 预测结果时，分类器的比重不同

24. 随机森林方法属于(B)

A. 梯度下降优化　　B. Bagging 方法

C. Boosting 方法　　D. 线性分类

25. 回归问题和分类问题的区别(A)

A. 前者预测函数值为连续值，后者为离散值

B. 前者预测函数值为离散值，后者为连续值

C. 前者是无监督学习

D. 后者是无监督学习

26. 正则化的回归分析，可以避免(B)

A. 线性化　　B. 过拟合　　C. 欠拟合　　D. 连续值逼近

27. 密度聚类方法充分考虑了样本间的什么关系(C)

A. 范数距离　　B. 集合运算

C. 密度可达　　D. 样本与集合运算

28. 主成分分析方法是一种什么方法(C)

A. 分类方法　　B. 回归方法　　C. 降维方法　　D. 参数估计方法

29. PCA 在做降维处理时，优先选取哪些特征(A)

A. 中心化样本的协方差矩阵的最大特征值对应特征向量

B. 最大间隔投影方向

C. 最小类内聚类

D. 最速梯度方向

30. 过拟合现象中(A)

A. 训练样本的测试误差最小，测试样本的正确识别率却很低

B. 训练样本的测试误差最小，测试样本的正确识别率也很高

C. 模型的泛化能力很高

D. 通常为线性模型

31. 下列关于过拟合现象的描述中，哪个是正确的（A）

A. 训练误差小，测试误差大　　B. 训练误差小，测试误差小

C. 模型的泛化能力高　　D. 其余选项都不对

32. 有两个样本点，第一个点为正样本,它的特征向量是(0,-1);第二个点为负样本,它的特征向量是(2,3),从这两个样本点组成的训练集构建一个线性 SVM 分类器的分类面方程是（C）

A. 2x+y=4　　B. x+2y=5　　C. x+2y=3　　D. 以上都不对

33. 下方法中属于无监督学习算法的是（D）

A. 线性回归　　B. 支持向量机

C. 决策树　　D. K-Means 聚类

34. Bootstrap 数据是什么意思（C）

A. 有放回地从总共 M 个特征中抽样 m 个特征

B. 无放回地从总共 M 个特征中抽样 m 个特征

C. 有放回地从总共 N 个样本中抽样 n 个样本

D. 无放回地从总共 N 个样本中抽样 n 个样本

35. 在 Logistic Regression 中,如果同时加入 L1 和 L2 范数,会产生什么效果（A）

A. 可以做特征选择，并在一定程度上防止过拟合

B. 能解决维度灾难问题

C. 能加快计算速度

D. 可以获得更准确的结果

36. 关于特征选择，下列对 Ridge 回归和 Lasso 回归说法正确的是？（B）

A. Ridge 回归适用于特征选择

B. Lasso 回归适用于特征选择

C. 两个都适用于特征选择

D. 以上说法都不对

37. 假如使用一个较复杂的回归模型来拟合样本数据，使用 Ridge 回归，调试正则化参数 λ，来降低模型复杂度。若 λ 较大时，关于偏差（bias）和方差（variance），下列说法正确的是？（C）

A. 若 λ 较大时，偏差减小，方差减小

B. 若 λ 较大时，偏差减小，方差增大

C. 若 λ 较大时，偏差增大，方差减小

D. 若 λ 较大时，偏差增大，方差增大

38. 如果在大型数据集上训练决策树。为了花费更少的时间来训练这个模型，下列哪种做法是正确的？（C）

A. 增加树的深度

B. 增加学习率

C. 减小树的深度

D. 减少树的数量

39、以下哪些方法不可以直接来对文本分类？ （A）

A、Kmeans

B、决策树

C、支持向量机

D、KNN

40、评估完模型之后，发现模型存在高偏差（high bias），应该如何解决？（B）

A. 减少模型的特征数量

B. 增加模型的特征数量

C. 增加样本数量

D. 以上说法都正确

解析：如果模型存在高偏差(high bias)，意味着模型过于简单。为了使模型更加健壮，我们可以在特征空间中添加更多的特征。而添加样本数量将减少方差。

41、下面关于 ID3 算法中说法错误的是（ D ）

A ID3 算法要求特征必须离散化

B 信息增益可以用熵，而不是 GINI 系数来计算

C 选取信息增益最大的特征，作为树的根节点

D ID3 算法是一个二叉树模型

42、在决策树中，用作分裂节点的 information gain 说法不正确的是（A）

A 较小不纯度的节点需要更多的信息来区分总体

B 信息增益可以使用熵得到

C 信息增益更加倾向于选择有较多取值的属性

43、一个 SVM 存在欠拟合问题，下面怎么做能提高模型的性能（A）

A 增大惩罚参数 C

B 减小惩罚参数 C

C 减小核函数系数(gamma 值)

解析：C >0 称为惩罚参数，是调和二者的系数，C 值大时对误差分类的惩罚增大，当 C 越大，趋近无穷的时候，表示不允许分类误差的存在，margin 越小，容易过拟合。

C 值小时对误差分类的惩罚减小,当 C 趋于 0 时，表示我们不再关注分类是否正确，只要求 margin 越大，容易欠拟合。

44、评估模型之后，得出模型存在偏差，下列哪种方法可能解决这一问题（B）

A 减少模型中特征的数量

B 向模型中增加更多的特征

C 增加更多的数据

D B 和 C

E 以上全是

解析：高偏差意味这模型不够复杂(欠拟合)，为了模型更加的强大，我们需要向特征空间中增加特征。增加样本能够降低方差

45、bootstrap 数据的含义是（C）

A 有放回的从整体 M 中抽样 m 个特征

B 无放回的从整体 M 中抽样 m 个特征

C 有放回的从整体 N 中抽样 n 个样本

D 无放回的从整体 N 中抽样 n 个样本

46、一个计算机程序从经验 E 中学习任务 T，并用 P 来衡量表现。并且，T 的表现 P 随着经验 E 的增加而提高。假设我们给一个学习算法输入了很多历史天气的数据，让它学会预测天气。什么是 P 的合理选择？（C）

A. 计算大量历史气象数据的过程

B. 以上都不

C. 正确预测未来日期天气的概率

D. 天气预报任务

二、多项选择题

1. Adaboost 方法中，需要迭代调整的两个重要参数是(AB)

A. 样本权重　　B. 分类器权重

C. 梯度变化率　　D. 梯度

2. 以下对层次聚类描述正确的(BD)

A. 监督学习　　B. 自顶向下寻找最优划分

C. 集成学习　　D. 自底向上寻找最优合并

3. 支持向量机可能解决的问题(ABC)

A. 线性分类　　B. 非线性分类　　C. 回归分析　　D. BP 算法

4. 影响 K-Means 聚类算法结果的主要因素有(BC)

A. 样本顺序　　B. 相似性度量　　C. 初始聚类中心　　D. 样本类别

5. 下面关于集成学习的描述，正确的是(AD)

A. Bagging 方法可以并行训练

B. Bagging 方法基学习器的比重不同

C. Boosting 方法可以并行训练

D. Boosting 方法基学习器的比重不同

6. 对于 PCA 说法正确的是(ABD)

A. 必须在使用 PCA 前规范化数据

B. 应该选择使得模型有最大 variance 的主成分

C. 应该选择使得模型有最小 variance 的主成分

D. 可以使用 PCA 在低维度上做数据可视化

7. 给定两个特征向量，以下哪些方法可以计算这两个向量相似度（ABD）

A. 欧式距离　　B. 夹角余弦　　C. 信息熵　　D. 曼哈顿距离

8.下列关于 sigmoid 函数描述正确的是（ABC）

A. 取值范围为（0,1），他可以将一个实数映射到（0,1）的区间，看做概率值；

B. 只能做二分类

C. 阈值一般设置在 0.5，大于该值的概率表示正例

D. 只有该函数能将实数映射到（0,1）区间

9、机器学习中 L1 正则化和 L2 正则化的区别是？ （AD）

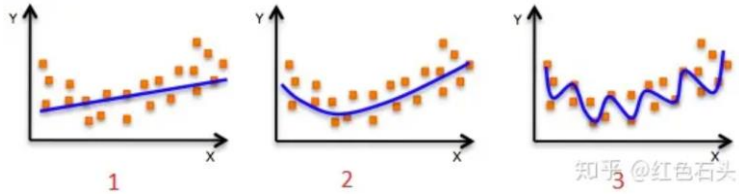
A、使用 L1 可以得到稀疏的权值

B、使用 L1 可以得到平滑的权值

C、使用 L2 可以得到稀疏的权值

D、使用 L2 可以得到平滑的权值

10、下面三张图展示了对同一训练样本，使用不同的模型拟合的效果（蓝色曲线）。那么，我们可以得出哪些结论（ACD）？



A. 第 1 个模型的训练误差大于第 2 个、第 3 个模型

B. 最好的模型是第 3 个，因为它的训练误差最小

C. 第 2 个模型最为“健壮”，因为它对未知样本的拟合效果最好

D. 第 3 个模型发生了过拟合

11、**下列哪种方法可以用来减小过拟合？（ABCD）**

A. 更多的训练数据

B. L1 正则化

C. L2 正则化

D. 减小模型的复杂度

三、判断题

1、梯度下降，就是沿着函数的负梯度方向更新自变量，使得函数的取值越来越小，直至达到全局最小或者局部最小。√

2、最小二乘法是基于预测值和真实值的均方差最小化的方法来估计线性回归学习器的参数 w 和 b。√

3、监督学习的学习数据既有特征（feature），也有标签（label）。√

4、线性回归主要用于解决回归问题，其因变量是连续的值。√

5、聚类生成的组称为簇，簇内任意对象之间具有较高的相似度，而簇间任意对象之间具有较高的相异度。√

6、k 近邻学习是一种常用的监督学习方法，其工作机制为：给定测试样本，基于某种距

离度量找出训练集中与其最靠近的 k 个训练样本，然后基于这 k 个邻居信息进行预测。因此 k 近邻算法的核心是 k 值和距离度量的选取。√

7、数据集一般划分为训练集、验证集和测试集三部分，训练集用于建模，验证集（开发集）用于模型验证与矫正，测试集用于模型的最终评估。√

8、正则化是为了防止模型过拟合而引入额外信息，对模型原有逻辑进行外部干预和修正，从而提高模型的泛化能力。√

9、Lasso 回归是对线性回归的优化，在线性回归的基础上，对损失函数增加了一个 L1 正则项，目的是降低方差，提高模型泛化能力。√

10、岭回归是对线性回归的优化，在线性回归的基础上，对损失函数增加了一个 L2 正则项，目的是降低方差，提高模型泛化能力。√

11.逻辑回归是有监督学习。 （√）

四、简答题

1、什么是机器学习？简述机器学习的一般过程。

答：机器学习是通过算法使得机器从大量历史数据中学习规律，从而对新样本做分类或预测。一般分为训练阶段、测试阶段和工作阶段。训练阶段的主要工作是根据训练数据建立模型，测试阶段的主要工作是利用验证集对模型评估与选择，工作阶段的主要工作是利用建立好的模型对新的数据进行预测与分类。

2、监督学习和非监督学习是什么

监督学习，是其训练集的数据是提前分好类，带有标签的数据，进行学习到模型以及参数，当用测试集进行测试时，给出 D 测={Xi}=>{yi}

非监督学习，需要将一系列没有标签的训练数据，输入到算法中，需要根据样本之间的相似性对样本集进行分类或者分析。

3、简述 K 折交叉验证与留一法的基本思想及其特点。

答：K 折交叉验证：将数据划分为 K 个大小相等的互斥子集；然后用其中的 K-1 个子集作为训练集，余下的那个子集作为测试集；这样就可以进行 K 次训练和测试，最终返回的是这 K 个测试结果的平均值。其稳定性和保真性在很大程度上取决与 K 的取值。

留一法：每次取一个样本作为测试集，其余样本组成的集合作为训练集，训练和测试的次数等于样本的个数。留一法的评估结果往往被认为是比较准确的，其最大的缺陷是当数据集较大时，模型的开销非常大。

4、简述什么是欠拟合和过拟合、产生的原因以及如何解决。

答：欠拟合：模型在训练集上的误差较高。原因：模型过于简单，没有很好的捕捉到数据特征，不能很好的拟合数据。解决方法：模型复杂化、增加更多的特征，使输入数据具有更强的表达能力等。

过拟合：在训练集上误差低，测试集上误差高。原因：模型把数据学习的太彻底，以至于把噪声数据的特征也学习到了，这样就会导致在后期测试的时候不能够很好地识别数据，模型泛化能力太差。解决方法：降维、增加训练数据、正则约束等。

5、简述线性回归与逻辑回归的区别。

答：（1）任务不同：回归模型是对连续的量进行预测；分类模型是对离散值/类别进行；（2）输出不同:回归模型的输出是一个连续的量，范围在[-∞,+∞]，分类模型的输出是数据属于某种类别的概率，范围在[0,1]之间；（3）参数估计方法不同：线性回归中

使用的是最小化平方误差损失函数，对偏离真实值越远的数据惩罚越严重；逻辑回归使用对数似然函数进行参数估计，使用交叉熵作为损失函数，对预测错误的惩罚是随着输出的增大，逐渐逼近一个常数。

6、简述决策树算法中剪枝的目的以及常用的两种剪枝方式的基本过程。

答：目的：剪枝是决策树学习算法对付“过拟合”的主要手段，通过主动去掉一些分支来降低过拟合的风险。基本策略有“预剪枝”和“后剪枝”。

“预剪枝”对每个结点划分前先进行估计，若当前结点的划分不能带来决策树泛化性能的提升，则停止划分，并标记为叶结点。

“后剪枝”先从训练集生成一棵完整的决策树，然后自底向上对非叶子结点进行考察，若该结点对应的子树用叶结点能带来决策树泛化性能的提升，则将该子树替换为叶结点。

7、简述 K 均值聚类算法的流程。

答：假设有 m 条数据，n 个特性，则 K 均值聚类算法的流程如下：

（1）随机选取 k 个点作为起始中心(k 行 n 列的矩阵，每个特征都有自己的中心)；（2）遍历数据集中的每一条数据，计算它与每个中心的距离；（3）将数据分配到距离最近的中心所在的簇；（4）使用每个簇中的数据的均值作为新的簇中心；（5）如果簇的组成点发生变化，则跳转执行第 2 步；否则，结束聚类。

8、简述什么是降维以及 PCA 算法的流程。

答：降维是通过某种数学变换将原始高维属性空间转变为一个低维子空间，保留重要性比较高的特征维度，去除冗余的特征。

主元成分分析 PCA 使用最广泛的数据降维算法，其一般流程如下：（1）样本零均值化；（2）计算数据的协方差矩阵；（3）计算协方差矩阵的特征值与特征向量；（4）按照特征值，将特征向量从大到小进行排序；（5）选取前 k 个特征向量作为转换矩阵；（6）零均值化后的数据与转换矩阵做矩阵乘法获得降维后的数据。

9、什么是最大似然学习？写出最大似然估计值的一般步骤。

（1）最大似然估计是已经知道了结果，然后寻找使该结果出现可能性最大的条件，以此作为估计值。

（2）1.写出似然函数，2.对似然函数取对数，并整理 3.求导，令导数为零，得到似然方程 4，解似然方程，得到的参数即为所求。

10、分类器分为哪几种模型，分别简要介绍一下。

分成判别式模型和生成式模型。

生成式模型：由数据学习联合概率分布 P(X,Y),然后由 P(Y|X)=P(X,Y)/P(X)，求出概率分布 P(Y|X)作为预测的模型，该方法表示了给定输入 X 与输出 Y 之间的生成关系

判别式模型：由数据直接学习决策函数 y=f(x)或者条件概率分布 P(Y|X)作为预测模型，判别方法关心的是对于给定输入 X 应预测出什么样的输出 Y

判别式：逻辑回归，线性回归，SVM，决策树，神经网络

生成式：朴素贝叶斯，贝叶斯网、高斯混合模型

11、简要叙述正则化项中 L1 和 L2 方法。

分析：1 正则化和L2 正则化可以看作是损失函数的惩罚项。L1 正则化是指权值向量w 中各个元素的绝对值之和，通常表示为||w||1。L2 正则化是指权值向量w 中各个元素的平方和，然后再求平方根（可以看到Ridge 回归的L2 正则化项有平方符号），通常表示为||w||2。

L1 正则化可以产生稀疏权值矩阵，即产生一个稀疏模型，可以用于特征选择

L2 正则化可以防止模型过拟合；一定程度上，L1 也可以防止过拟合

12、什么是训练数据集和测试数据集？

在类似于机器学习的各个信息科学相关领域中，一组数据被用来发现潜在的预测关系，称为“训练数据集”。训练数据集是提供给学习者的案例，而试验数据集是用于测试由学习者提出的假设关系的准确度。

13. 在数据处理时，为什么通常要进行标准化处理。

答案：在实际问题中，我们使用的样本通常是多维数据，每一维对应一个特征，这些特征的量纲和数量级都是不一样的，这时需要对数据进行标准化处理，是所有的特征具有同样的尺度。

14. 请写出通过条件概率公式和全概率公式推出贝叶斯公式的过程

分析：条件概率：$P(A|B)=\frac{P(AB)}{P(B)}$，$P(B|A)=\frac{P(AB)}{P(A)}$

全概率：$P(A)=\sum_i P(A|B_i)P(B_i)$

贝叶斯公式：$P(B_i|A)=\frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$

15. 当学习率过高或过低时会怎样？

当模型的学习率过低时，模型的训练速度会变得非常慢，因为其每次对权重的更新会变得非常小。模型将需要大量更新才能到达局部最优点。

如果学习率过高，模型很可能无法收敛，因为权重的更新过大。在加权的步骤中，模型有可能无法实现局部优化，然后使模型难以更新到最优点（因为每步更新都跳得过远，导致模型在局部最优点附近摇摆）。

16. 阐述偏置和方差的概念以及它们之间的权衡关系

偏置是当前模型的平均预测结果与我们需要预测的实际结果之间的差异。当模型的偏置较高时，说明其不够关注训练数据。这会使得模型过于简单，无法在训练和测试上同时实现优良的准确度。这个现象也被称为「欠拟合」。

方差（variance）可以简单理解为是模型输出在一个数据点上的分布（或聚类）。方差越大，模型越有可能更密切关注训练数据，而无法提供在从未见过的数据上的泛化能力。由此造成的结果是，模型可在训练数据集上取得非常好的结果，但在测试数据集上的表现却非常差。这个现象被称为过拟合。

偏置和方差之间需要保持平衡。如果我们的模型过于简单，有非常少的参数，那么它就可能有较高的偏置和较低的方差。另一方面，如果我们的模型有大量参数，则其将有较高的方差和较低的偏置

17、经验误差(empirical error)与泛化误差(generalization error)分别指？

经验误差：也叫训练误差(training error)，模型在训练集上的误差。

泛化误差：模型在新样本集(测试集)上的误差

18、简述 K 折交叉验证(k-fold crossValidation)。

数据集大小为 N，分成 K 份，则每份含有样本 N/K 个。每次选择其中 1 份作为测试集，另外 K-1 份作为训练集，共 K 种情况。

在每种情况中，训练集训练模型，用测试集测试模型，计算模型的泛化误差。

将 K 种情况下，模型的泛化误差取均值，得到模型最终的泛化误差

19、随机森林的随机性体现在哪里？

随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定。随机森林的随机性体现在每颗树的训练样本是随机的，树中每个节点的分裂属性集合也是随机选择确定的。有了这 2 个随机的保证，随机森林就不会产生过拟合的现象了。

随机森林是用一种随机的方式建立的一个森林，森林是由很多棵决策树组成的，每棵树所分配的训练样本是随机的，树中每个节点的分裂属性集合也是随机选择确定的。

20、线性回归中正则化的目的是什么吗？L1 正则化和 L2 正则化有什么不同？

在线性回归中，正则化是用来防止模型过拟合的一种技术，通过引入额外的信息（惩罚项）来约束或减少系数的大小。

正则化的主要目的是改善模型的泛化能力，即在未见数据上的表现。通过对系数大小施加惩罚，正则化帮助确保模型不会过于依赖训练数据中的特定观测，从而减少模型在新数据上预测时的误差。

L1 正则化（也称为 Lasso 回归）通过向损失函数添加系数的绝对值之和的惩罚项，促使一部分系数精确地缩减到 0。这种性质使 L1 正则化成为一种有效的特征选择方法，因为它可以减少模型中变量的数量，留下对输出变量有显著影响的那些变量。

L2 正则化（也称为岭回归）通过向损失函数添加系数的平方和的惩罚项，倾向于使系数值均匀地缩小，但不会将它们完全缩减到 0。L2 正则化有助于处理特征间的多重共线性问题，通过保留所有特征但减小系数值的影响，从而提高模型的稳定性和泛化能力。

简而言之，L1 正则化倾向于产生稀疏系数矩阵，有助于特征选择，而 L2 正则化则通过惩罚大的系数值来防止过拟合，保证模型的复杂度得到适当的控制。在实际应用中，根据数据的特点和需求选择适当的正则化方法是非常重要的。

21、逻辑回归能用于多分类问题吗？如果可以，应如何应用？

逻辑回归虽然最初是为二分类问题设计的，但它也可以被扩展到多分类问题。处理多分类问题的常用方法有两种：一对多（One-vs-Rest， OvR）和多项逻辑回归（Multinomial Logistic Regression）。

一对多（OvR）方法涉及将多分类问题分解为多个二分类问题。对于每个类别，模型将这个类别与所有其他类别对比，构建一个二分类模型。因此，如果有 N 个类别，就会有 N 个二分类模型。在预测时，所有模型都会被用来评估一个观测值，然后选择概率最高的类别作为最终预测。

多项逻辑回归（又称为 Softmax 回归）是另一种方法，它直接在一个模型中处理多个类别。与逻辑函数用于二分类一样，多项逻辑回归使用 Softmax 函数来将线性函数的输出转换为概率分布。这种方法不需要像 OvR 那样为每个类别构建多个模型，因此在某些情况下可能更有效率。

22、讨论数据归一化和标准化对 KNN 算法的影响。

在使用 KNN 算法时，数据预处理是一个关键步骤，尤其是数据归一化和标准化。这两种技术对 KNN 算法的性能有着显著影响，原因在于 KNN 是基于距离的算法，而距离计算对数据的尺度非常敏感。

数据归一化（Normalization）：归一化是将所有特征缩放到[0, 1]区间的过程。这种方

法对于确保没有一个特征会因为其数值范围大而对距离计算产生不成比例的影响非常有效。例如，如果一个特征的范围是 0 到 1，而另一个特征的范围是 0 到 1000，未归一化的数据会导致后者对距离计算的贡献远大于前者。

数据标准化（Standardization）：标准化是将数据特征缩放到具有零均值和单位方差的过程。这对于 KNN 来说也是非常重要的，因为它确保了所有特征在距离计算中具有相同的重要性，无论它们的原始尺度或分布如何。标准化有助于减少特征之间尺度差异造成的偏差，从而提高 KNN 算法的准确性和效率。

总体而言，数据归一化和标准化可以显著提高 KNN 算法的性能，尤其是在处理具有不同尺度和分布的特征时。通过确保所有特征在相同的尺度上，KNN 算法能够更公平地比较不同特征之间的距离，从而做出更准确的预测。因此，对于 KNN 算法，适当的数据预处理不仅是提高模型性能的重要步骤，也是确保模型公正性的关键。

23、为什么它被称为“朴素”贝叶斯？

“朴素”贝叶斯的称呼来源于它在建模时采用的一个基本假设，即所有的特征在给定类别的条件下都是相互独立的。这个假设被认为是“朴素”的（这个假设好“天真”），因为在现实世界中，特征之间往往是有关联的。

24、想象你有一个决策树模型，在你的训练数据上准确率达到 100%，但在测试数据上只有 60%。这可能是什么原因，你会如何解决？

这种现象通常表明模型出现了过拟合。过拟合是指模型在训练数据上表现得非常好，几乎或完全达到了 100%的准确率，但是在未见过的测试数据上表现不佳。这意味着模型学习到了训练数据中的噪声和特定的细节，而没有捕捉到更广泛的、能够泛化到新数据的模式。

解决过拟合的方法通常包括：

简化模型：减少模型复杂度，例如，通过限制决策树的深度，减少分支的数量，或者移除不重要的特征。

增加数据：如果可能的话，增加更多的训练数据，可以帮助模型学习到更广泛的数据分布特性。

使用正则化：在模型训练过程中引入正则化项（如决策树的剪枝）可以防止模型变得过于复杂。

交叉验证：使用交叉验证来更准确地评估模型的泛化能力。

集成学习：使用集成学习方法如随机森林，可以减少过拟合，因为它们通过结合多个模型的预测来提高泛化能力。

通过这些方法，可以减少模型对训练数据的过度拟合，提高其在未见数据上的表现。

25、什么是支持向量，以及它们在 SVM 中为什么重要？

支持向量是在 SVM 模型中，距离决策超平面最近的那些数据点。它们被称为“支持”向量，因为它们直接影响到超平面的位置和方向。在 SVM 中，超平面的确定是为了最大化两个类别之间的间隔，而这个间隔就是由最靠近超平面的数据点——也就是支持向量——决定的。

26、请解释 SVM 中软间隔的概念

在支持向量机（SVM）的上下文中，软间隔（Soft Margin）的概念是指允许某些数据点违反最初的间隔规则的一种机制。这意味着在寻找分割超平面时，一些数据点可以被允许位于它们应该分隔开的边界之内，或者甚至在错误的一侧。这种方法主要用于处理非线性可分的数据集，以及在数据中存在轻微的噪声或异常值时。

27、在什么情况下你会选择使用 DBSCAN 聚类算法而不是 k-means？

DBSCAN（Density-Based Spatial Clustering of Applications with Noise）是一种基于密度的聚类算法，与 k-means 算法相比，有其独特的优势和适用场景：

数据集含有异常值：DBSCAN 对异常值具有很好的鲁棒性。它能够识别并处理噪声点，这些噪声点不会被划分到任何聚类中，而 k-means 由于是基于距离的方法，异常值会影响聚类中心的计算。

聚类形状的多样性：当数据集中的聚类形状不仅仅是球形时，DBSCAN 能够识别任意形状的聚类，因为它是基于密度的聚类，而 k-means 则假设聚类是球形，对于非球形聚类的识别能力有限。

聚类数量未知：DBSCAN 不需要事先指定聚类的数量，它通过密度连接的概念自动确定聚类的数量。相反，k-means 需要预先设定聚类的数量 k，这在不了解数据集特性的情况下可能难以确定。

数据集的规模和密度差异大：DBSCAN 能够很好地处理规模和密度差异大的数据集，因为它通过局部密度估计来形成聚类，而 k-means 在处理规模和密度差异大的数据集时可能不会得到满意的聚类结果。

综上所述，当数据集含有异常值、聚类形状多样、聚类数量未知或数据集的规模和密度差异大时，选择 DBSCAN 聚类算法而不是 k-means 将更为合适。

28、解释主成分分析（PCA）的基本原理。它是如何确定主成分的？

主成分分析（PCA）是一种统计方法，用于通过正交变换将一组可能相关的变量转换为一组线性不相关的变量，这组新的变量被称为主成分。PCA 的基本原理是找到一个新的坐标系统，将原始数据转换到这个坐标系统中，使得这些数据在新坐标轴上的投影具有最大的方差，从而捕获数据中最多的变异性。

PCA 确定主成分的过程包括几个步骤：首先，计算数据集的协方差矩阵，以反映变量间的相互关系。接着，计算协方差矩阵的特征值和特征向量。特征向量定义了新的坐标系统的方向，而特征值则给出了在这些方向上的方差大小。将特征值按降序排列，对应的特征向量就是主成分。通常，只选择前几个主成分，这些主成分的特征值较大，因而能够保留大部分数据的变异性。

四、计算题

1、（本小题 15 分）使用朴素贝叶斯分类器预测一个未知样本的分类。数据样本用属性“天气”、“温度”、“湿度”和“风力”描述。目标分类属性“是否适合打网球”具有两个不同值（即“是”和“否”）。设 C1 对应于分类“是否适合打网球”=“是”，而 C2 对应于分类“是否适合打网球”=“否”。训练数据如下表所示，预测样本为 X=（“天气”=“雨”，“温度”=“凉”，“湿度”=“高”，

“风力”=“弱”），根据朴素贝叶斯算法计算出 X 的属于不同分类目标值的概率，判断最终的预测结果。

天气	气温	湿度	风力	适合打网球吗？
晴	热	高	弱	否
晴	热	高	强	否
阴	热	高	弱	是
雨	适宜	高	弱	是
雨	凉	正常	弱	是
雨	凉	正常	强	否
阴	凉	正常	强	是
晴	适宜	高	弱	否
晴	凉	正常	弱	是
雨	适宜	正常	弱	是
晴	适宜	正常	强	是
阴	适宜	高	强	是
阴	热	正常	弱	是
雨	适宜	高	强	否

5.我们需要最大化 P(X|Ci)P(Ci)， i=1,2。每个类的先验概率 P(C)可以根据训练样本计算:
P(C1)=9/14=0.643　　P(C2)=5/14=0.357
（2 分，公式和计算各 1 分）
为计算.P(X/Ci)　i=1,2，我们计算下面的条件概率:
P(天气=″雨″ | C1)=3/9=0.333　　P(天气=″雨″ | C2)=2/5=0.400
（2 分，公式和计算各 1 分）
P(温度=″凉″ | C1)=3/9=0.333　　P(温度=″凉″ | C2)=1/5=0.200
（2 分，公式和计算各 1 分）
P(湿度=″高″ | C1)=3/9=0.333　　P(湿度=″高″ | C2)=4/5=0.800
（2 分，公式和计算各 1 分）
P(风力=″弱″ | C1)=6/9=0.667　　P(风力=″弱″ | C2)=2/5=0.400
（2 分，公式和计算各 1 分）
使用以上概率，我们得到:
P(X| C1)=0.333×0.333×0.333×0.667=0.0247
P(X| C2)=0.400×0.200×0.800×0.400=0.0256
（2 分，公式和计算各 1 分）
P(X|C1)P(C1)=0.0247×0.643=0.01588　　P(X|C2)P(C2)=0.0256×0.357=0.00914
或者
P(X|C1)P(C1)/P(X)=49/90=0.5444　　P(X|C2)P(C2)/P(X)=196/625=0.3136
（2 分，公式和计算各 1 分）
因此，对于样本 X，朴素贝叶斯分类预测 C1 （2 分）

2、样本特征为二维欧式空间点的两分类问题的训练集 (-1, 1)，(0, 1)，(0, 2)，(1, -1)，(1, 0)，(1, 2)，(2, 2)，(2, 3)，类别标签为“+”和“-”，如下表所示。用最近邻法给出测试样本点(1, 1)的类别，其中 K 取值为 3。

序号	1	2	3	4	5	6	7	8

x_1	-1	0	0	1	1	1	2	2
x_2	1	1	2	-1	0	2	2	3
Y	-	+	-	-	+	+	-	+

解：（1）计算距离

(x,y)　Distance--(1,1)

(-1,1)　√((-1-1)^2+(1-1)^2)=2 -

(0,1)　√((0-1)^2+(1-1)^2)=1 +

(0,2)　√((0-1)^2+(2-1)^2)=√2 -

(1,-1)　√((1-1)^2+(-1-1)^2)=2 -

(1,0)　√((1-1)^2+(0-1)^2)=1 +

(1,2)　√((1-1)^2+(2-1)^2)=1 +

(2,2)　√((2-1)^2+(2-1)^2)=√2 -

(2,3)　√((2-1)^2+(3-1)^2)=√5 +

最近邻法: (0,1) +, (1,0) +, (1,2) +　----->　+

3. 抛一枚硬币问题，观察数据情况是：一枚硬币包括正反两面，共抛了 30 次，其中 12 次是正面，18 次是反面。采用 Maximum Likelihood 方法，估计正面出现的概率和反面出现的概率。

答案：

设正面出现的概率为p，则反面出现的概率为1-p。

上述实验出现的概率为：

$$L(p)=C_{30}^{12}p^{12}(1-p)^{18}$$

对上式求偏导：

$$\frac{\partial L}{\partial p}=12C_{30}^{12}p^{11}(1-p)^{18}-18C_{30}^{12}p^{12}(1-p)^{17}$$

令偏导等于0，解得：$p=0.4$

所以，正面出现的概率为0.4，反面出现的概率为0.6。

4、用两个硬币玩抛硬币的游戏，硬币 1 得到正面的概率为 θ，硬币 2 得到正面的概率为 2θ，你一共抛了五次，得到的结果是这样的（硬币 1，正面）（硬币 2，反面）（硬币 2，反面）（硬币 2，反面）（硬币 2，正面），用极大似然法求参数 θ。

关键步骤：

1. 两个硬币的似然函数:
   $P_1(\mathbf{x}|\theta) \sim Ber(\theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$,
   $P_2(\mathbf{x}|\theta) \sim Ber(2\theta) = \prod_{i=1}^{N} (2\theta)^{x_i}(1-2\theta)^{1-x_i}$
2. 根据观测得到的数据，似然函数为
   $P(\mathbf{x}|\theta, 2\theta) = \theta * (1-2\theta)^3 * 2\theta = 2\theta^2(1-2\theta)^3$,对数似然为
   $logP(\mathrm{x}|\theta, 2\theta) = 2log\theta + 3log(1-2\theta) + log2$, 最大化对数似然，求导置零，得到$\theta^* = \frac{1}{5}$

9. 已知 $P(\omega_1)=0.2$， $P(\omega_2)=0.8$，
   $P(x=\text{阴天}|\omega_1)=0.6$， $P(x=\text{晴天}|\omega_1)=0.4$，
   $P(x=\text{阴天}|\omega_2)=0.1$， $P(x=\text{晴天}|\omega_2)=0.9$
   已知$x=$阴天，求 x 所属类别。

答案：

$$P(\omega_1|x=\text{阴天}) = \frac{p(x=\text{阴天}|\omega_1)P(\omega_1)}{p(x=\text{阴天})}$$
$$= \frac{p(x=\text{阴天}|\omega_1)P(\omega_1)}{p(x=\text{阴天}|\omega_1)P(\omega_1)+p(x=\text{阴天}|\omega_2)P(\omega_2)}$$
$$= \frac{0.6\times0.2}{0.6\times0.2+0.1\times0.8} = 0.6$$
$$P(\omega_2|x=\text{阴天}) = \frac{p(x=\text{阴天}|\omega_2)P(\omega_2)}{p(x=\text{阴天})}$$
$$= \frac{p(x=\text{阴天}|\omega_2)P(\omega_2)}{p(x=\text{阴天}|\omega_1)P(\omega_1)+p(x=\text{阴天}|\omega_2)P(\omega_2)}$$
$$= \frac{0.1\times0.8}{0.6\times0.2+0.1\times0.8} = 0.4$$
$$\therefore x \in \omega_1$$

11. 以下为标注数据以及对应的特征，其中，A，B，C为两类特征，Y为类别标签，利用朴素贝叶斯分类器求A=0，B=1，C=1时，Y的分类标签。

A	1	0	0	1	0	1	0	0	1	1
B	0	1	1	0	1	0	0	1	0	1
C	0	0	1	0	1	1	0	1	0	0
Y	1	0	1	1	0	0	1	0	1	1

答案：

$$P(A=0|Y=0)=\frac{3}{4}\quad,\quad P(A=0|Y=1)=\frac{1}{3}$$

$$P(B=1|Y=0)=\frac{3}{4}\quad,\quad P(B=1|Y=1)=\frac{1}{3}$$

$$P(C=1|Y=0)=\frac{3}{4},\quad P(C=1|Y=1)=\frac{1}{6}$$

$$P(Y=0)=\frac{2}{5}\quad,\quad P(Y=1)=\frac{3}{5}$$

由贝叶斯公式得

$$P(Y=0|A=0,B=1,C=1)=\frac{P(A=0,B=1,C=1|Y=0)P(Y=0)}{P(A=0,B=1,C=1)}$$

$$=\frac{P(A=0|Y=0)P(B=1|Y=0)P(C=1|Y=0)P(Y=0)}{P(A=0,B=1,C=1)}$$

$$=\frac{\frac{3}{4}\times\frac{3}{4}\times\frac{3}{4}\times\frac{2}{5}}{P(A=0,B=1,C=1)}$$

$$=\frac{\frac{27}{160}}{P(A=0,B=1,C=1)}$$

同理

$$P(Y=1|A=0,B=1,C=1)=\frac{P(A=0,B=1,C=1|Y=1)P(Y=1)}{P(A=0,B=1,C=1)}$$

$$=\frac{P(A=0|Y=1)P(B=1|Y=1)P(C=1|Y=1)P(Y=1)}{P(A=0,B=1,C=1)}$$

$$=\frac{\frac{1}{3}\times\frac{1}{3}\times\frac{1}{6}\times\frac{3}{5}}{P(A=0,B=1,C=1)}$$

$$=\frac{\frac{1}{90}}{P(A=0,B=1,C=1)}$$

$$\because P(Y=0|A=0,B=1,C=1)>P(Y=1|A=0,B=1,C=1)$$

$$\therefore Y=0$$

4、（10 分）应用 PCA 算法将下列二维数据$X=\begin{pmatrix}-1 & -1 & 0 & 2 & 0\\ -2 & 0 & 0 & 1 & 1\end{pmatrix}$降为一维，要求写出具体的求解过程。

参考答案

协方差矩阵：$\sum = \frac{1}{5}\begin{pmatrix}-1 & -1 & 0 & 2 & 0\\ -2 & 0 & 0 & 1 & 1\end{pmatrix}\begin{pmatrix}-1 & -2\\ -1 & 0\\ 0 & 0\\ 2 & 1\\ 0 & 1\end{pmatrix} = \begin{pmatrix}\frac{6}{5} & \frac{4}{5}\\ \frac{4}{5} & \frac{6}{5}\end{pmatrix}$ **(4 分)**

求∑的特征值和特征向量：$|A-\lambda E| = \begin{vmatrix}\frac{6}{5}-\lambda & \frac{4}{5}\\ \frac{4}{5} & \frac{6}{5}-\lambda\end{vmatrix} = (\frac{6}{5}-\lambda)^2 - \frac{16}{25} = (\lambda-2)(\lambda-2/5) = 0$

求解得到特征值：$\lambda_1 = 2$，$\lambda_2 = 2/5$

其对应的特征向量分别是： $\sum_1 = \begin{pmatrix}1\\1\end{pmatrix}$，$\sum_2 = \begin{pmatrix}-1\\1\end{pmatrix}$

标准化后的特征向量为：$P = \begin{pmatrix}1/\sqrt{2}\\ 1/\sqrt{2}\end{pmatrix}, \begin{pmatrix}-1/\sqrt{2}\\ 1/\sqrt{2}\end{pmatrix}$ （4 分）

协方差矩阵∑的对角化：

$$P\sum P^T = \begin{pmatrix}1/\sqrt{2} & 1/\sqrt{2}\\ -1/\sqrt{2} & 1/\sqrt{2}\end{pmatrix}\begin{pmatrix}6/5 & 4/5\\ 4/5 & 6/5\end{pmatrix}\begin{pmatrix}1/\sqrt{2} & -1/\sqrt{2}\\ 1/\sqrt{2} & 1/\sqrt{2}\end{pmatrix} = \begin{pmatrix}2 & 0\\ 0 & 2/5\end{pmatrix}$$

用P的第一行乘以数据矩阵，得到降维后的数据表示：

$$Y = (1/\sqrt{2} \quad 1/\sqrt{2})\begin{pmatrix}-1 & -1 & 0 & 2 & 0\\ -2 & 0 & 0 & 1 & 1\end{pmatrix} = (-3/\sqrt{2} \quad -1/\sqrt{2} \quad 0 \quad 3/\sqrt{2} \quad 1/\sqrt{2})$$

（2 分）

5、有下列样本数据，要用决策树算法进行分类，请计算相关变量。

（1）计算经验熵 H(D)。

（2）分别以 A1、A2、A3 表示性别、车型、衬衣尺码这三个特征，试计算这三个特征各自对应的信息增益。

（3）三个特征中哪一个是最优特征？

顾客 ID	性别	车型	衬衣尺码	类
1	男	家用	小	C0
2	男	运动	中	C0
3	男	运动	中	C0
4	男	运动	大	C0
5	女	运动	小	C0
6	女	运动	小	C0
7	女	运动	中	C0
8	女	豪华	大	C0
9	男	家用	大	C1
10	男	家用	中	C1
11	女	豪华	小	C1
12	女	豪华	小	C1
13	女	豪华	中	C1
14	女	豪华	中	C1

6、假设某机器学习模型的原始类别和预测类别如下表所示，求它的混淆矩阵、准确率、精确率（也叫查准率）、召回率、F1 Score。

样本序号	1	2	3	4	5	6	7	8	9	10
原始类别	1	1	1	-1	-1	-1	1	1	-1	1
预测类别	1	1	-1	-1	-1	1	-1	1	-1	1

predict \ real	1	−1
1	4	1
−1	2	3

计算准确率：accuracy = (TP+TN)/(TP+FP+TN+FN)（4+3）/10=0.7

对于类别 1：

精确率：precision = TP/(TP+FP)=4/(4+1)=0.8

召回率：recall = TP/(TP+FN) = 4/(4+2) =0.67

F1 score:

$$F1_k = 2\frac{pecision_k \cdot recall_k}{pecision_k + recall_k} = 2\frac{0.8 \cdot 0.67}{0.8 + 0.67} = 0.73$$

对于类别 -1：

精确率：precision = TP/(TP+FP)=3/(3+2)=0.6

召回率：recall = TP/(TP+FN) = 3/(3+1) =0.75

F1 score:

$$F1_k = 2\frac{pecision_k \cdot recall_k}{pecision_k + recall_k} = 2\frac{0.6 \cdot 0.75}{0.6 + 0.75} = 0.67$$

整体 F1-score = (0.73+0.67)/2 = 0.7

7、给定 3 个数据点：正例点 x1=（3，3），x2=（4，3），负例点 x3=（1，1），求线性可分支持向量机。

解：（1）原问题：

$$例：\min_{w_1,w_2,b}\frac{1}{2}\left(w_1^2+w_2^2\right)$$

$$s.t.\begin{cases}3w_1+3w_2+b\geq 1\\4w_1+3w_2+b\geq 1\\w_1+w_2+b\leq -1\end{cases}$$

（2）转化为对偶问题：

$$\min_{\alpha_1,\alpha_2,\alpha_3} -\sum_{i=1}^{3}\alpha_i+\frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\alpha_i\alpha_j y_i y_j\left(x_i \square x_j\right)$$

$$s.t.\begin{cases}\alpha_1+\alpha_2-\alpha_3=0\\\alpha_1\geq 0,\alpha_2\geq 0,\alpha_3\geq 0\end{cases}$$

代入 x1、x2 和 x3 得：

$$\min_{\alpha_1,\alpha_2,\alpha_3}\left\{\frac{1}{2}\left(18\alpha_1^2+25\alpha_2^2+2\alpha_3^2+42\alpha_1\alpha_2-12\alpha_1\alpha_3-14\alpha_2\alpha_3\right)-\alpha_1-\alpha_2-\alpha_3\right\}$$

$$s.t.\ \begin{cases}\alpha_1+\alpha_2-\alpha_3=0\\ \alpha_1\geq 0,\alpha_2\geq 0,\alpha_3\geq 0\end{cases}$$

代入约束条件 $\alpha_3=\alpha_1+\alpha_2$,:

目标函数变形为:

$$s(\alpha_1,\alpha_2)=4\alpha_1^2+\frac{13}{2}\alpha_2^2+10\alpha_1\alpha_2-2\alpha_1-2\alpha_2$$

分别对 α 1 和 α 2 求导，得:

$$\Rightarrow\begin{cases}s_{\alpha_1}(\alpha_1,\alpha_2)=8\alpha_1+10\alpha_2-2\\ s_{\alpha_2}(\alpha_1,\alpha_2)=13\alpha_2+10\alpha_1-2\end{cases}$$

$$\Rightarrow\begin{cases}\alpha_1=\frac{3}{2}\\ \alpha_2=-1\end{cases}$$ 不符合要求，从而最小值在边界达到。

（3）利用 KKT 条件，求向量 W:

从而:当 $\alpha_1=\frac{1}{4},\ \alpha_2=0,$ 时，$s_{\min}=-\frac{1}{4}.$ 此时 $\alpha_3=\frac{1}{4}.$

$$故:w=\alpha_1y_1x_1+\alpha_3y_3x_3=\left(\frac{1}{2},\frac{1}{2}\right)$$

（4）利用 KKT 条件，求变量 b:

注意到 $\alpha_1>0,$ 从而 x_1 为支持向量。

从而 $y_1f(x_1)=1\overset{y_1^2=1}{\Longrightarrow}y_1^2f(x_1)=y_1\Rightarrow b=y_1-w'x_1=-2$

这样我们就得到了支持向量机(分离超平面)

$$\frac{1}{2}x_1+\frac{1}{2}x_2-2=0$$

对于新的样本点，我们使用的决策函数为

$$f(x)=\text{sign}\left(\frac{1}{2}x_1+\frac{1}{2}x_2-2\right)$$

8、KNN 算法

例 3.1 已知二维空间的 3 个点 $x_1=(1,1)^{\mathrm{T}}, x_2=(5,1)^{\mathrm{T}}, x_3=(4,4)^{\mathrm{T}}$，试求在 p 取不同值时，L_p 距离下 x_1 的最近邻点.

解 因为 x_1 和 x_2 只有第二维上值不同，所以 p 为任何值时，$L_p(x_1,x_2)=4$. 而

$$L_1(x_1,x_3)=6,\quad L_2(x_1,x_3)=4.24,\quad L_3(x_1,x_3)=3.78,\quad L_4(x_1,x_3)=3.57$$

于是得到：p 等于 1 或 2 时，x_2 是 x_1 的最近邻点；p 大于等于 3 时，x_3 是 x_1 的最近邻点. ■

例 3.2 给定一个二维空间的数据集：

$$T=\{(2,3)^{\mathrm{T}},(5,4)^{\mathrm{T}},(9,6)^{\mathrm{T}},(4,7)^{\mathrm{T}},(8,1)^{\mathrm{T}},(7,2)^{\mathrm{T}}\}$$

构造一个平衡 *kd* 树[3].

解 根结点对应包含数据集 T 的矩形，选择 $x^{(1)}$ 轴，6 个数据点的 $x^{(1)}$ 坐标的中位数是 7，以平面 $x^{(1)}=7$ 将空间分为左、右两个子矩形（子结点）；接着，左矩形以 $x^{(2)}=4$ 分为两个子矩形，右矩形以 $x^{(2)}=6$ 分为两个子矩形，如此递归，最后得到如图 3.3 所示的特征空间划分和如图 3.4 所示的 *kd* 树. ■

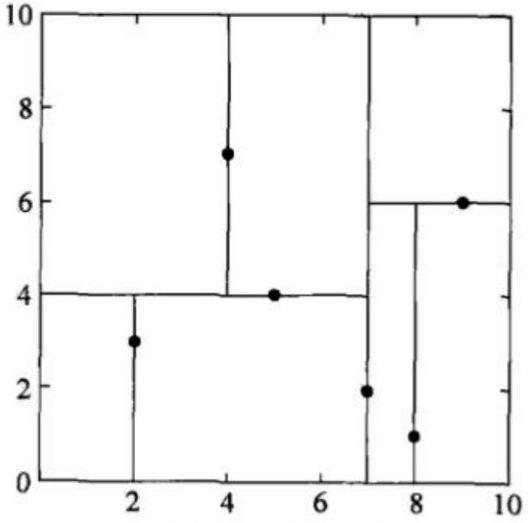


图 3.3 特征空间划分

9、朴素贝叶斯分类算法

例 4.1 试由表 4.1 的训练数据学习一个朴素贝叶斯分类器并确定 $x=(2,S)^{\mathrm{T}}$ 的类标记 y。表中 $X^{(1)}$，$X^{(2)}$ 为特征，取值的集合分别为 $A_1=\{1,2,3\}$，$A_2=\{S,M,L\}$，Y 为类标记，$Y\in C=\{1,-1\}$。

表 4.1 训练数据

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	1	1	-1

解 根据算法 4.1，由表 4.1，容易计算下列概率：

$$P(Y=1)=\frac{9}{15},\quad P(Y=-1)=\frac{6}{15}$$

$$P(X^{(1)}=1|Y=1)=\frac{2}{9},\quad P(X^{(1)}=2|Y=1)=\frac{3}{9},\quad P(X^{(1)}=3|Y=1)=\frac{4}{9}$$

$$P(X^{(2)}=S|Y=1)=\frac{1}{9},\quad P(X^{(2)}=M|Y=1)=\frac{4}{9},\quad P(X^{(2)}=L|Y=1)=\frac{4}{9}$$

$$P(X^{(1)}=1|Y=-1)=\frac{3}{6},\quad P(X^{(1)}=2|Y=-1)=\frac{2}{6},\quad P(X^{(1)}=3|Y=-1)=\frac{1}{6}$$

$$P(X^{(2)}=S|Y=-1)=\frac{3}{6},\quad P(X^{(2)}=M|Y=-1)=\frac{2}{6},\quad P(X^{(2)}=L|Y=-1)=\frac{1}{6}$$

对于给定的 $x=(2,S)^{\mathrm{T}}$ 计算：

$$P(Y=1)P(X^{(1)}=2|Y=1)P(X^{(2)}=S|Y=1)=\frac{9}{15}\cdot\frac{3}{9}\cdot\frac{1}{9}=\frac{1}{45}$$

$$P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)=\frac{6}{15}\cdot\frac{2}{6}\cdot\frac{3}{6}=\frac{1}{15}$$

因为 $P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)$ 最大，所以 $y=-1$。

10、决策树算法

表 5.1　贷款申请样本数据表

ID	年龄	有工作	有自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

利用 ID3, CART, C4.5 算法构建决策树，要求写出计算过程并画出并决策树。

ID3算法

首先算经验熵 H(D)　　表数据集中有9个是 6个否

$$H(D) = -\frac{9}{15}\log_2\frac{9}{15} - \frac{6}{15}\log_2\frac{6}{15} = 0.971$$

分别用 A_1, A_2, A_3, A_4 表示年龄、有工作、有自己的房子和信贷情况4个特征，则有

① $g(D, A_1) = H(D) - H(D|A_1)$

A_1 有青年、中年、老年三种情况，将数据集分割为三个部分

$$H(D|A_1) = \frac{5}{15}H(D_1) + \frac{5}{15}H(D_2) + \frac{5}{15}H(D_3)$$

D_1 (青年)中 有 2是3否 ： $H(D_1) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$

D_2 (中年)有 3是 2否 ： $H(D_2) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971$

D_3 (老年)中有 4是1否 ： $H(D_3) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.722$

$$H(D|A_1) = \frac{5}{15}[H(D_1) + H(D_2) + H(D_3)]$$
$$= 0.888$$

$$g(D, A_1) = 0.971 - 0.888 = 0.083$$

---

② $g(D, A_2) = H(D) - H(D|A_2)$

A_2 有是与否两种情况，将D分为两部分

$$H(D|A_2) = \frac{5}{15}H(D_1) + \frac{10}{15}H(D_2)$$

D_1 (是)中有 5是 0否　　$H(D_1) = -\frac{5}{5}\log_2\frac{5}{5} = 0$

D_2 (否)中有 4是 6否　　$H(D_2) = -\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10} = 0.971$

$$g(D, A_2) = 0.971 - \frac{10}{15} \times 0.971$$
$$= 0.324$$

---

③ $g(D, A_3) = H(D) - H(D|A_3)$

$H(D|A_3) = \frac{6}{15} H(D_1) + \frac{9}{15} H(D_2)$

$= \frac{6}{15}[-\frac{6}{6}\log_2\frac{6}{6}] + \frac{9}{15}[-\frac{3}{9}\log_2\frac{3}{9} - \frac{6}{9}\log_2\frac{6}{9}]$

$= 0.551$

$g(D, A_3) = 0.971 - 0.551 = 0.420$

---

④ $g(D, A_4) = H(D) - H(D|A_4)$

$H(D|A_4) = \frac{5}{15} H(D_1) + \frac{6}{15} H(D_2) + \frac{4}{15} H(D_3)$

$= \frac{5}{15}[-\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}] + \frac{6}{15}[-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}]$

$+ \frac{4}{15}[-\frac{4}{4}\log_2\frac{4}{4}]$

$= 0.608$

$g(D, A_4) = 0.971 - 0.608 = 0.363$

ID3算法中选 $g(D, A)$ 最大的作为根节点

∴ A_3 为最优特征

A_3 作为根节点之后，D_2 需要从 A_1，A_2、A_4 之选择新特征

D_2 为 D中通过 A_3 判断为否的子集，因为有配自的房子（A_3 二是）类别全为是，也将成为一个叶节点，类别记为是

$g(D_2, A_1) = H(D_2) - H(D_2|A_1) = 0.251$

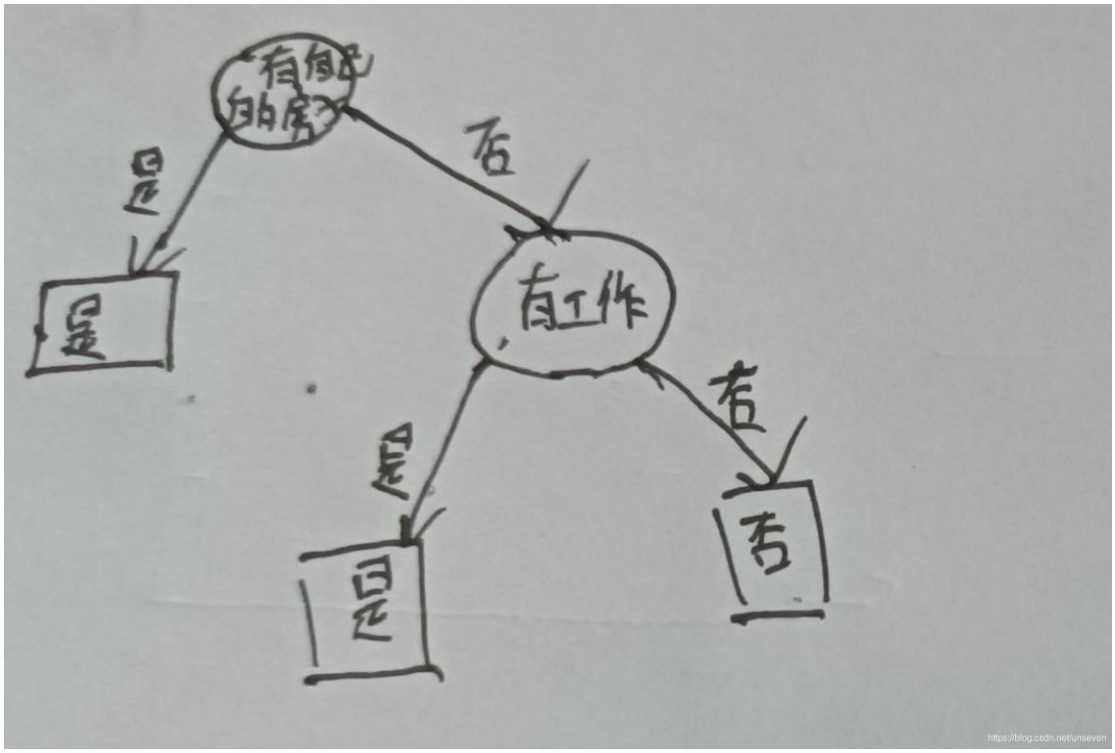
$g(D_2, A_2) = H(D_2) - H(D_2|A_2) = 0.981$

$g(D_2, A_4) = H(D_2) - H(D_2|A_4) = 0.474$

A_2（有工作）作为最优特征

A_2 有两个取值，A_2 二是时，三个样本全为是，这是一个叶结点，类别为是

A_2 二否时，6个样子全为否，这也是一个叶节点，记为否



例 5.2 对表 5.1 所给的训练数据集 D，根据信息增益准则选择最优特征.

解 首先计算经验熵 $H(D)$.

$$H(D)=-\frac{9}{15}\log_2\frac{9}{15}-\frac{6}{15}\log_2\frac{6}{15}=0.971$$

然后计算各特征对数据集 D 的信息增益. 分别以 A_1，A_2，A_3，A_4 表示年龄、有工作、有自己的房子和信贷情况 4 个特征，则

(1)

$$\begin{aligned}g(D,A_1)&=H(D)-\left[\frac{5}{15}H(D_1)+\frac{5}{15}H(D_2)+\frac{5}{15}H(D_3)\right]\\&=0.971-\left[\frac{5}{15}\left(-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}\right)\right.\\&\quad\left.+\frac{5}{15}\left(-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5}\right)+\frac{5}{15}\left(-\frac{4}{5}\log_2\frac{4}{5}-\frac{1}{5}\log_2\frac{1}{5}\right)\right]\\&=0.971-0.888=0.083\end{aligned}$$

这里 D_1，D_2，D_3 分别是 D 中 A_1（年龄）取值为青年、中年和老年的样本子集. 类似地，

（2）

$$g(D,A_2)=H(D)-\left[\frac{5}{15}H(D_1)+\frac{10}{15}H(D_2)\right]$$

$$=0.971-\left[\frac{5}{15}\times 0+\frac{10}{15}\left(-\frac{4}{10}\log_2\frac{4}{10}-\frac{6}{10}\log_2\frac{6}{10}\right)\right]=0.324$$

（3）

$$g(D,A_3)=0.971-\left[\frac{6}{15}\times 0+\frac{9}{15}\left(-\frac{3}{9}\log_2\frac{3}{9}-\frac{6}{9}\log_2\frac{6}{9}\right)\right]$$

$$=0.971-0.551=0.420$$

（4）

$$g(D,A_4)=0.971-0.608=0.363$$

最后，比较各特征的信息增益值．由于特征 A_3（有自己的房子）的信息增益值最大，所以选择特征 A_3 作为最优特征． ■

例 5.3 对表 5.1 的训练数据集，利用 ID3 算法建立决策树．

解 利用例 5.2 的结果，由于特征 A_3（有自己的房子）的信息增益值最大，所以选择特征 A_3 作为根结点的特征．它将训练数据集 D 划分为两个子集 D_1（A_3 取值为“是”）和 D_2（A_3 取值为“否”）．由于 D_1 只有同一类的样本点，所以它成为一个叶结点，结点的类标记为“是”．

对 D_2 则需从特征 A_1（年龄），A_2（有工作）和 A_4（信贷情况）中选择新的特征．计算各个特征的信息增益：

$$g(D_2,A_1)=H(D_2)-H(D_2\mid A_1)=0.918-0.667=0.251$$

$$g(D_2,A_2)=H(D_2)-H(D_2\mid A_2)=0.918$$

$$g(D_2,A_4)=H(D_2)-H(D_2\mid A_4)=0.474$$

选择信息增益最大的特征 A_2（有工作）作为结点的特征．由于 A_2 有两个可能取值，从这一结点引出两个子结点：一个对应“是”（有工作）的子结点，包含 3 个样本，它们属于同一类，所以这是一个叶结点，类标记为“是”；另一个是对应“否”（无工作）的子结点，包含 6 个样本，它们也属于同一类，所以这也是一个叶结点，类标记为“否”．

这样生成一个如图 5.5 所示的决策树．该决策树只用了两个特征（有两个内部结点）．

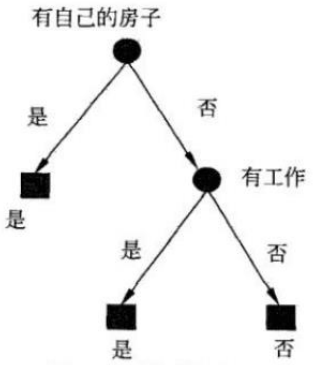


图 5.5　决策树的生成

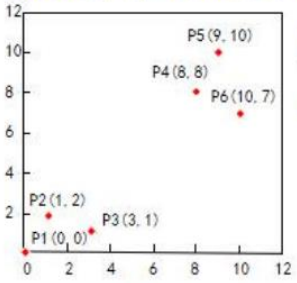
ID3 算法只有树的生成，所以该算法生成的树容易产生过拟合.

11、KMeans 算法

K-Means

K-Means举例

已知如下数据，对表中数据点运用K-means方法进行聚类分析。设k=2，初始选择 P1,P2。



1.选择初始聚类点：

选P1和P2

2.计算其它各点和初始点的距离：

	P1	P2
P3	3.16	2.24
P4	11.3	9.22
P5	13.5	11.3
P6	12.2	10.3

P3到P6都跟P2更近，所以第一次站队的结果是：

组A：P1

组B：P2、P3、P4、P5、P6

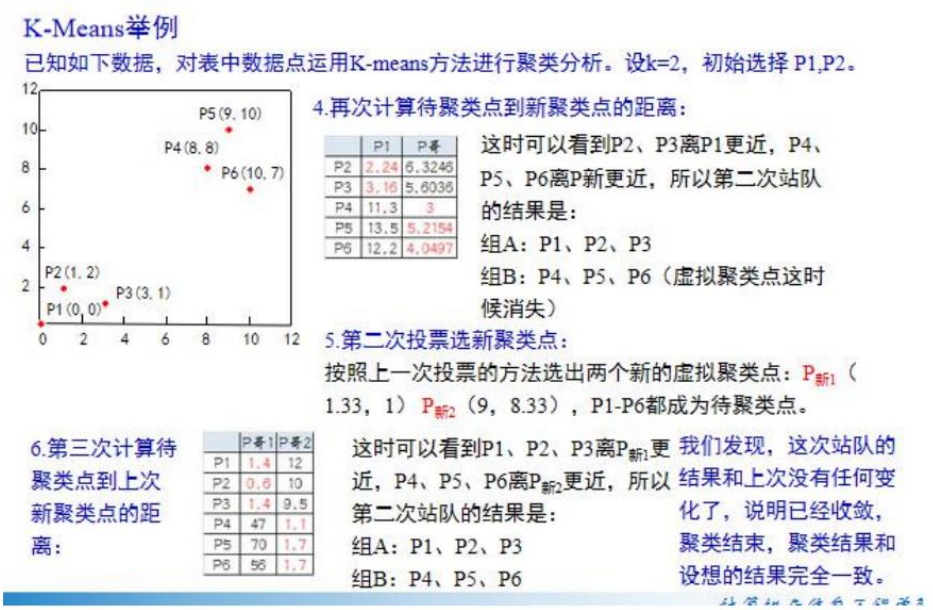
3.投票选新聚类点：

组A没啥可选的，聚类点还是P1自己

组B有五个点，需要选新聚类，这里要注意选新聚类点的方法是每个点 X 坐标的平均值和 Y 坐标的平均值组成的新的点，为新聚类点，也就是说这是“虚拟的”。

因此，B组选出新聚类点的坐标为：P新 ((1+3+8+9+10)/5，(2+1+8+10+7)/5)=(6.2，5.6)

综合两组，新聚类点为P1（0，0），$P_{新}$（6.2，5.6），而P2-P6重新成为待聚类点。



12.使用k-means算法，给出下列数据每一轮的聚类结果和最终的聚类结果。

点	x_1	x_2
P1	0	1
P2	1	2
P3	2	2
P4	8	8
P5	9	10
P6	10	10

注：初始化聚类中心为P1和P2。

答案：

第一轮：

{P1}，{P2}

{P1}，{P2，P3}

{P1}，{P2，P3，P4}

{P1}，{P2，P3，P4，P5}

{P1}，{P2，P3，P4，P5，P6}

新的质心：(0，1)，(6，6.4)

第二轮：

{P1}，{}

{P1，P2}，{}

{P1，P2，P3}，{}

{P1，P2，P3}，{P4}

{P1，P2，P3}，{P4，P5}

{P1，P2，P3}，{P4，P5，P6}

新的质心：(1，5/3)，(9，28/3)

第三轮：

{P1}，{}

{P1，P2}，{}

{P1，P2，P3}，{}

{P1，P2，P3}，{P4}

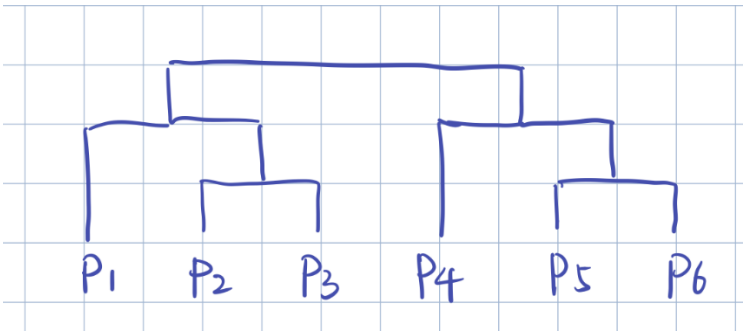
{P1，P2，P3}，{P4，P5}

{P1，P2，P3}，{P4，P5，P6}

新的质心：(1，5/3)，(9，28/3)

质心不再改变，得出最终的聚类结果：

{P1，P2，P3}，{P4，P5，P6}



13.假设数据挖掘的任务是将8 个点聚类成3 个簇，A1(2,10)，A2(2,5)，A3(8,4)，B1(5,8)，B2(7,5)，B3(6,4)，C1(1,2)，C2(4,9)，距离函数是欧几里得距离。假设初始选择A1，B1，C1 分别作为每个聚类的中心，用k—均值算法来给出：

（1）第一次循环执行后的三个聚类中心；

（2）最后的三个簇。

答案：

（1）第一轮

A1(2,10)

B1（5,8）,A3（8,4）,B2（7,5）,B3（6,4）,C2（4,9）

C1（1,2）,A2（2,5）

对应中心分别是（2,10),（6,6）,(1.5,3.5)

（2）最后三个簇｛A1(2,10),B1（5,8）,C2（4,9）｝｛A3（8,4）,B2（7,5）,B3（6,4）｝｛C1（1,2）,A2（2,5）｝

例 4.1 试由表 4.1 的训练数据学习一个朴素贝叶斯分类器并确定 $x=(2,S)^{\mathrm{T}}$ 的类标记 y。表中 $X^{(1)}$，$X^{(2)}$ 为特征，取值的集合分别为 $A_1=\{1,2,3\}$，$A_2=\{S,M,L\}$，Y 为类标记，$Y\in C=\{1,-1\}$。

表 4.1 训练数据

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	−1	−1	1	1	−1	−1	−1	1	1	1	1	1	1	1	−1

答案见《统计学习方法》和 PPT

8.1.3 AdaBoost 的例子①

例 8.1 给定如表 8.1 所示训练数据。假设弱分类器由 $x<v$ 或 $x>v$ 产生，其阈值 v 使该分类器在训练数据集上分类误差率最低。试用 AdaBoost 算法学习一个强分类器。

① 例题来源于 http://www.csie.edu.tw。

表 8.1 训练数据表

序号	1	2	3	4	5	6	7	8	9	10
x	0	1	2	3	4	5	6	7	8	9
y	1	1	1	−1	−1	−1	1	1	1	−1

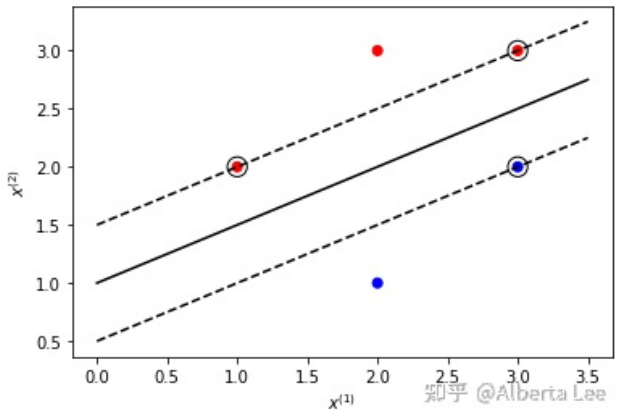
答案见《统计学习方法》和 PPPT

已知正例点 $x_1=(1,2)^T, x_2=(2,3)^T, x_3=(3,3)^T$，负例点 $x_4=(2,1)^T, x_5=(3,2)^T$，试求最大间隔分离平面和分类决策函数，并在图中挂出分离超平面、间隔边界及支持向量。

最大间隔分离超平面： $-x^{(1)}+2x^{(2)}-2=0$

分类决策函数： $f(x)=\mathrm{sign}(-x^{(1)}+2x^{(2)}-2)$

支持向量： $x_1=(3,2)^T, x_2=(1,2)^T, x_3=(3,3)^T$



14、将表4-1所示的样本集合按年龄（大于等于29）进行切分，试计算切分后的信息增益和基尼指数。

表4-1 某人相亲数据

编号	年龄（岁）	身高（cm）	学历	月薪（元）	是否相亲
1	35	176	本科	20000	否
2	28	178	硕士	10000	是
3	26	172	本科	25000	否
4	29	173	博士	20000	是
5	28	174	本科	15000	是

解：切分前信息熵 H(A)=0.971

按年龄（大于等于 29）进行切分：两个样本 A1 和 A2，A1 中两个样本，A2 三个样本。

$H(A1)= -1/2*log_2 1/2 - 1/2*log_2 1/2=1$

$H(A2)= - 2/3*log_2 2/3 - 1/3*log_2 1/3=0.918$

所以切分后信息熵为 $H(A')=2/5H(A1)+3/5H(A2)=0.951$

∴切分后的信息增益为 0.02

和基尼指数 $Gini(\{A1,A2\})=|A1|/|A|Gini(A1)+|A2|/|A|Gini(A2)$

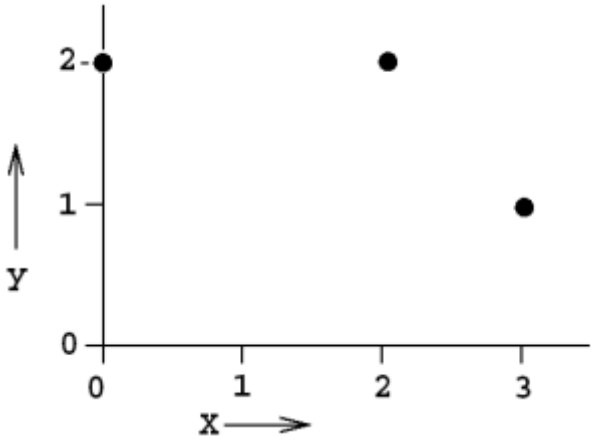
$=2/5*(1-[(1/2)^2+(1/2)^2]+3/5*(1-[(2/3)^2+(1/3)^2])=0.467$

15、假设你有以下数据：输入和输出都只有一个变量。使用线性回归模型（y=wx+b）来拟合数据。那么使用留一法（Leave-One Out）交叉验证得到的均方误差是多少？

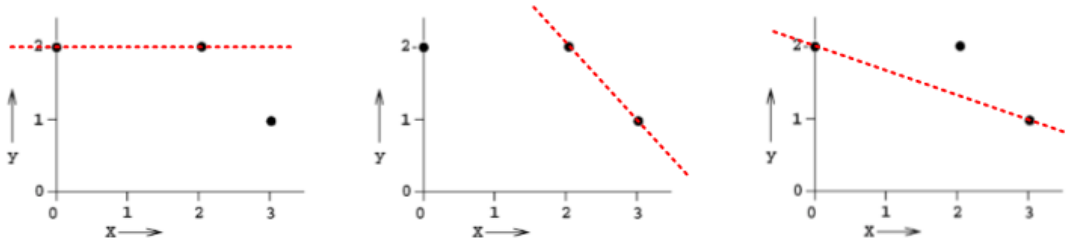
X(independent variable)	Y(dependent variable)
0	2
2	2
3	1

解析：留一法，简单来说就是假设有 N 个样本，将每一个样本作为测试样本，其它 N-1 个样本作为训练样本。这样得到 N 个分类器，N 个测试结果。用这 N 个结果的平均值来衡量模型的性能。

对于该题，我们先画出 3 个样本点的坐标：



使用两个点进行线性拟合，分成三种情况，如下图所示：



第一种情况下，回归模型是 y = 2，误差 E1 = 1。

第二种情况下，回归模型是 y = -x + 4，误差 E2 = 2。

第三种情况下，回归模型是 y = -1/3x + 2,误差 E3 = 2/3。

则总的均方误差为：

$$MSE = \frac{1}{3}(E_1^2 + E_2^2 + E_3^2) = \frac{1}{3}(1^2 + 2^2 + (\frac{2}{3})^2) = \frac{49}{27}$$

二．训练数据集正实例点是 $x_1=(2,3)^T$，$x_2=(4,3)^T$，负实例点是 $x_3=(0.5,0.5)^T$，要求：

（1）建立最大间隔分离超平面的原问题模型；　（2）建立最大间隔分离超平面的对偶问题模型；

（3）假设对偶问题的最优解 $\alpha_1=0.2353,\alpha_2=0.0000,\alpha_3=0.2353$，请写出分离超平面和判别超平面方程。（20 分）

提示：

支持向量机原问题数学模型：

$$\min \frac{\|w\|}{2}$$
$$s.t.\quad y_i(x_i\cdot w+b)-1\geq 0, i=1,2,\cdots,n$$

支持向量机对偶问题数学模型：

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{l}\alpha_i$$
$$s.t.\sum_i a_i y_i=0; a_i\geq 0, i=1,\cdots,l$$

超平面参数：$w^*=\sum_{i=1}^{l}a_i y_i x_i$，$b^*=y_j-\sum_{i=1}^{l}a_i^* y_i(x_i\cdot x_j)$。

Linear Regression

Consider fitting the linear regression model for these data

x	-1	0	2
y	1	-1	1

(b) Fit $Y_i=\beta_0+\epsilon_i$ (degenerated linear regression), find β_0.

β_0 =argmin $\sum(Y_i-\beta_0)^2$

$\beta_0=1/3$

(b) Fit $Y_i=\beta_1 X_i+\epsilon_i$ (linear regression without the constant term), find β_0 and β_1.

β_1 =argmin $\sum(Y_i-\beta_1 X_i)^2$

$\beta_1=\sum X_i Y_i/\sum X_i^2=1/5$

16、计算变量［0,0,1,1,1］的信息熵。A

A -(3/5 log(3/5) + 2/5 log(2/5))

B 3/5 log(3/5) + 2/5 log(2/5)

C 2/5 log(3/5) + 3/5 log(2/5)

D 3/5 log(2/5) - 2/5 log(3/5)

17、1、假设系统 A 内包含 4 个事件 A、B、C、D 发生的概率分别为 P(A)=0.25, P(B)=0.25, P(C)=0.25, P(D)=0.25，系统 B 内包含 2 个事件 E、F 发生的概率分别为 P(E)=0.5, P(F)=0.5，两个系统的熵分别是多少？哪个系统更加混乱，在分类器的设计过程中希望系统的熵沿着哪个方向进行变化？（7 分）

参考答案：

H_A = -(0.25*log$_2$0.25+0.25*log$_2$0.25+0.25*log$_2$0.25+0.25*log$_2$0.25)=2　（2 分）

H_B =-(0.5*log$_2$0.5+0.5*log$_2$0.5) = 1　（2 分）

系统 A 更混乱　（2 分）

希望系统的熵越小越好 （1 分）

18、给定 3 个数据点：正例点 x_1=（3，3），x_2=（4，3），负例点 x_3=（1，1），要求（1）建立最大间隔分离超平面的原问题模型；（2）建立最大间隔分离超平面的对偶问题模型；（3）假设对偶问题的最优解$\alpha_1 = \frac{1}{4}$，$\alpha_2 = 0$，$\alpha_3 = \frac{1}{4}$，请写出分离超平面和判别超平面方程。(10 分)

提示：

支持向量机原问题数学模型：

$$\min \frac{1}{2}\|w\|^2$$

$$s.t. \quad y_i(x_i \cdot w + b) - 1 \geq 0, i = 1,2,3,\cdots,n$$

支持向量机对偶问题数学模型：

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{l}\alpha_i$$

$$s.t. \quad \sum_{i}\alpha_i y_i = 0; \ \alpha_i \geq 0, i = 1,2,3,\cdots,l$$

超平面参数：$w^* = \sum_{i=1}^{l}\alpha_i y_i x_i$，$b^* = y_j - \sum_{i=1}^{l}\alpha_i^* y_i (x_i \cdot x_j)$

参考答案：

解：（1）原问题：（3 分）

$$例：\min_{w_1,w_2,b}\frac{1}{2}\left(w_1^2 + w_2^2\right)$$

$$s.t. \begin{cases} 3w_1 + 3w_2 + b \geq 1 \\ 4w_1 + 3w_2 + b \geq 1 \\ w_1 + w_2 + b \leq -1 \end{cases}$$

（2）对偶问题：（3 分）

$$\min_{\alpha_1,\alpha_2,\alpha_3} -\sum_{i=1}^{3}\alpha_i + \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\alpha_i\alpha_j y_i y_j \left(x_i \square x_j\right)$$

$$s.t. \begin{cases} \alpha_1 + \alpha_2 - \alpha_3 = 0 \\ \alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0 \end{cases}$$

代入 x1、x2 和 x3 得：

$$\min_{\alpha_1,\alpha_2,\alpha_3} \left\{\frac{1}{2}\left(18\alpha_1^2 + 25\alpha_2^2 + 2\alpha_3^2 + 42\alpha_1\alpha_2 - 12\alpha_1\alpha_3 - 14\alpha_2\alpha_3\right) - \alpha_1 - \alpha_2 - \alpha_3\right\}$$

$$s.t. \begin{cases} \alpha_1 + \alpha_2 - \alpha_3 = 0 \\ \alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0 \end{cases}$$

(3) 分离超平面：（3 分）

当 $\alpha_1 = \frac{1}{4}$，$\alpha_2 = 0$，$\alpha_3 = \frac{1}{4}$时，

$$故：w=\alpha_1 y_1 x_1+\alpha_3 y_3 x_3=\left(\frac{1}{2},\frac{1}{2}\right)$$

注意到$\alpha_1>0$,从而 x_1 为支持向量。

$$从而\ y_1 f(x_1)=1 \overset{y_1^2=1}{\Longrightarrow} y_1^2 f(x_1)=y_1 \Rightarrow b=y_1-w'x_1=-2$$

这样我们就得到了支持向量机(分离超平面)

$$\frac{1}{2}x_1+\frac{1}{2}x_2-2=0$$

(4) 判别超平面方程为：(1 分)

对于新的样本点，我们使用的决策函数为

$$f(x)=\text{sign}\left(\frac{1}{2}x_1+\frac{1}{2}x_2-2\right)$$

19、1、基于一个学生在大学一年级的表现，预测他在大学二年级表现。令 x 等于学生在大学第一年得到的“A”的个数（包括 A-，A 和 A+成绩），以此表示学生在大学第一年得到的成绩。预测 y 的值：第二年获得的“A”级的数量。下表中每一行是一个训练数据。在线性回归中，我们的假设$h_\theta(x)=\theta_0+\theta_1 x$，并且我们使用 m 来表示训练示例的数量。

x	y
3	2
1	2
0	1
4	3

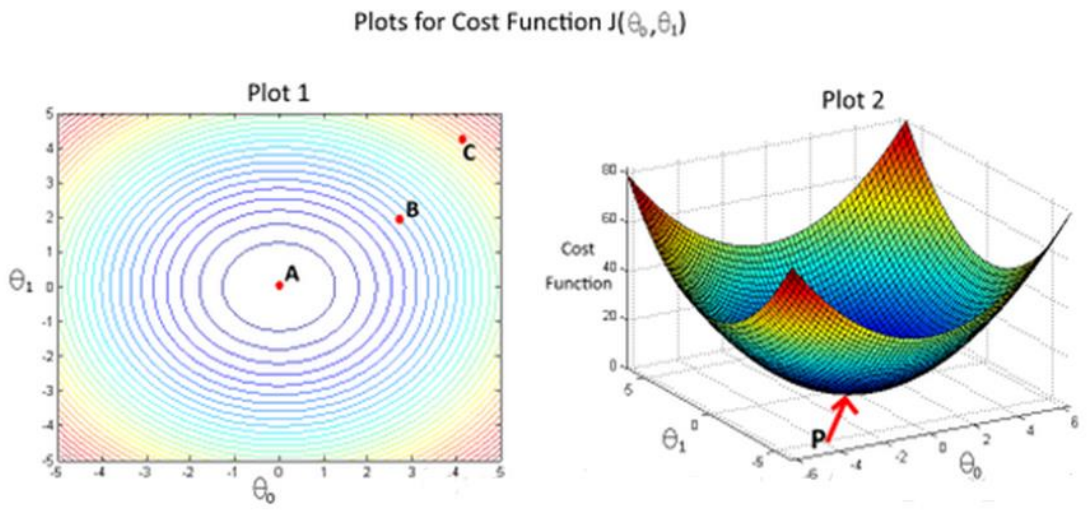
（1） 代价函数的定义是$J(\theta_0,\theta_1)=\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i)-y^{(i)})^2$, 求 J(0, 1)。

解析：$J(0,1)=\frac{1}{2*4}\left((3-2)^2+(1-2)^2+(0-1)^2+(4-3)^2\right)=0.5$

（2） 线性回归中，假设$\theta_0=-1,\theta_1=2$，求$h_\theta(6)$的值。

解析：$h_\theta(6)=-1+2*6=11$

（3） 代价函数$J(\theta_0,\theta_1)$与θ_0,θ_1的关系如下图 plot2 所示。下图 plot1 中给出了相同代价函数的等高线图。根据图示，选择正确的选项（选出所有正确项）。



A. 从 B 点开始，学习率合适的梯度下降算法会最终帮助我们到达或者接近 A 点，即代价函数$J(\theta_0,\theta_1)$在 A 点有最小值

B. 点 P（图 plot2 的全局最小值）对应于图 plot1 的点 C

C. 从 B 点开始，学习率合适的梯度下降算法会最终帮助我们到达或者接近 C 点，即代价函数$J(\theta_0,\theta_1)$在 C 点有最小值

D. 从 B 点开始，学习率合适的梯度下降算法会最终帮助我们到达或者接近 A 点，即代价函数$J(\theta_0,\theta_1)$在 A 点有最大值

E. 点 P（图 plot2 的全局最小值）对应于图 plot1 的点 A

解析：AE，P 是全局最小值，对应的是 A。