**第五章作业：**（共20分）

（单选（0.5分）×6，多选（1分）×3，大题（14分））

1.【单选】最大似然估计中，使用对数似然函数是为了\_\_**D\_**\_。

A.提高似然函数灵敏度 B.将求极大值问题转化为求极小值问题

C.便于对似然函数求和 D.便于对似然函数求导

2.【单选】已知要估计的概率密度函数是多峰的，则选择\_**B**\_法能够得到一个相对较好的密度曲线估计。

A.贝叶斯估计 B.使用正态窗函数的parzen窗

C. Kn近邻法 D.最大似然估计

3.【单选】下面关于贝叶斯分类器的论述中，（ **C** ）是错误的。

A. 贝叶斯分类器的分类决策边界可以是线性的，也可以是非线性的

B. 贝叶斯分类器是最小化分类错误率的分类器

C. 最小化错误率的贝叶斯分类器等价于最小化风险的贝叶斯分类器

D. 最小化错误率的贝叶斯分类器没有考虑不同的错分类的代价

4.【单选】以下关于K近邻分类器的描述错误的是\_\_**D**\_\_\_.

A. K近邻分类器是一个非线性分类器。

B. K近邻分类器可以直接用于多类的分类问题。

C. SVM和贝叶斯分类器需要提前训练好分类器才能用于分类，而K近邻分类器不需要提前训练分类器。

D. K近邻分类器的分类结果不受距离函数的选择的影响。

5.【单选】在2维的坐标系中，距离原点的ℓ\_1范数距离(曼哈顿距离)等于1的点的轨迹是\_**A**\_。

A. 以原点为中心，以两个坐标轴为对角线，边长为根号2的正方形

B. 以原点为中心，边长为2，各个边与坐标轴平行的正方形

C. 以原点为中心，半径为1的圆

D. 以原点为中心，边长为根号2，各个边与坐标轴平行的正方形

6.【单选】以下关于特征归一化/标准化的描述错误的是\_**B**\_\_.

A.特征归一化的目的是近似地均衡每维特征的影响，使每维的特征在距离计算中有近似相同的作用。

B.特征归一化在任何的数据集上都是适用的。

C.对某一维度的特征进行z-score标准化后，该维度特征的均值为0，方差为1。

D.对数据集进行min-max标准化会将每个维度上的特征值变换到相同的取值范围内。

7.【多选】关于距离，下列说法正确的是\_\_**ABD**\_\_\_.

A. 欧式距离就是两点之间的直线距离。

B. 曼哈顿距离是两个点在所有维度上的距离之和，其中在每个维度上的距离为两个点在该维度上的坐标值之差的绝对值。

C. ℓ\_∞范数距离又称为城市街区距离。

D. ℓ\_∞范数距离是两个点在各个维度上距离的最大值。

8.【多选】贝叶斯分类器的训练，就是利用训练样本集估计出\_\_**BD**\_\_。

A.全概率 B.类条件概率/概率密度

C.后验概率 D.先验概率

9.【多选】下列描述正确的是\_\_**ABCD**\_\_。

A.非参数化的估计方法直接对概率密度函数p(x|ω\_j)本身进行估计，而不必假设p(x|ω\_j)具有某个确定的分布形式。

B.贝叶斯估计法把待估计的参数θ看作是随机量，服从某个已知的先验分布。

C.参数化方法假定p(x|ω\_j)具有某种确定的分布形式，但是分布的参数未知，然后利用训练集对分布的参数进行估计。

D.利用最大似然估计法对某类的类条件概率密度函数p(x|ω\_j)进行估计时，需要假设该类中的样本是独立同分布的。

10. 已知一个样本集是一维的，包含和两个类别的样本，每类样本出现的先验概率相同：，且类条件概率密度函数为：

(1)根据贝叶斯分类规则，判断应该将新样本预测为哪一类？（2分）

(2)记表示一个样本的真实类别为时，将其预测为类所产生的损失，且，，，根据最小化风险贝叶斯分类规则，判断将新样本预测为哪一类？（2分）

解答：(1)

因为，，所以将新样本预测为类。 **(2分)**

(2)

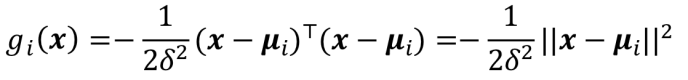
因为，，所以将新样本预测为类。 **(2分)**

11. 在两类三维空间的分类问题中，第一类的特征向量样本服从三维正态分布，第二类特征向量服从三维正态分布，其中，，，且两类特征向量出现的概率相等。

(1)试写出将贝叶斯分类规则通过判别函数表示法表示时，相应的两个判别函数的方程和分类决策面的方程各是什么。（2分）

(2)根据贝叶斯分类规则判断应该将一个样本分为哪一类？（2分）

解答：(1)由题目可得：**，**代入下面的判别函数方程：



可得：， **(1分)**

分类决策面方程为:，其中，。

**(1分)**

1. 因为, ,

，所以应该将分为第二类。**(2分)**

12. 假设一个一维样本集D={,}是根据参数θ未知的伯努利分布Bernoulli(θ)采样得到的，即， 其中*x*=0或1，0≤θ≤1，请采用最大似然估计法给出θ的估计，写出详细的推导步骤。（3分）

解答：似然函数可以表示为：

**(1分)**

对数似然函数可以表示为：

**(1分)**

对关于求导，令导数为0，得到：

求解上述方程得到： **(1分)**

13. 已知二维空间中的3个点，，，试求在分别取1，2和时，在范数距离下的最近邻点。（3分）

解答：

4，，所以，在范数距离下的最近邻点是。 **(1分)**

4，，所以在范数距离下的最近邻点是。**(1分)**

4，，所以，在范数距离下的最近邻点是。 **(1分)**

**第六章作业：**（共10分）（多选（1分）×1，大题（9分））

1.【多选】下列描述正确的有\_\_**BD**\_\_\_.

A. PCA寻找的是能够有效分类的方向，而FDA寻找的是能够有效表示数据的方向。

B. FDA寻找到的最优投影方向满足：不同类的投影点的均值之间的差距最大，同类的投影点的分散程度(方差)最小。

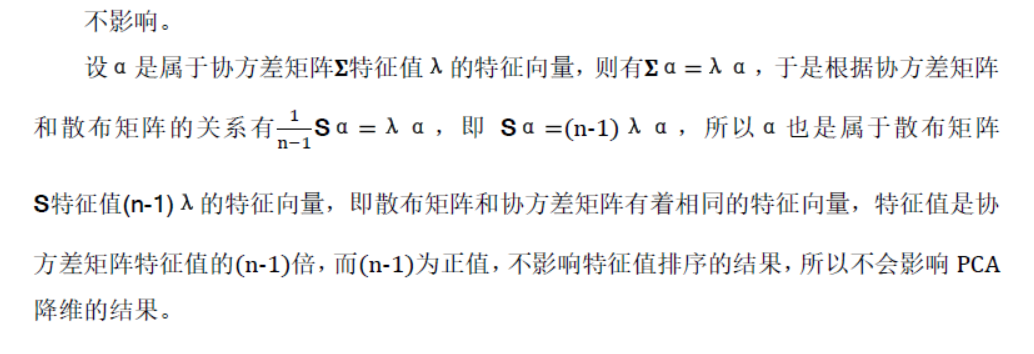
C. 数据集的本征维度是表示数据集所需要的最小变量数，与数据集的原始维度不同，一般地，本征维度大于原始维度。

D. PCA是无监督的降维方法，不考虑样本集的类标信息，而FDA则是有监督的降维方法，需要知道样本集的类标信息。

2. PCA算法的主成分e\_1, e\_2,..., e\_k之间有什么关系? 每个主成分的模是多少? (1分)

**解答：**任意两个主成分都互相垂直，且每个主成分的模为**1.**

3. 给定d维的样本集D={x\_1, x\_2,⋯, x\_n}，μ为样本集的均值向量，已知样本的**协方差矩阵**定义为, 这里符号表示矩阵的转置运算。在PCA中，**散布矩阵**定义为，可以看到散布矩阵是协方差矩阵的倍。试分析在PCA用协方差矩阵来代替散布矩阵会有什么样的影响，会影响PCA降维的结果吗？(2分)



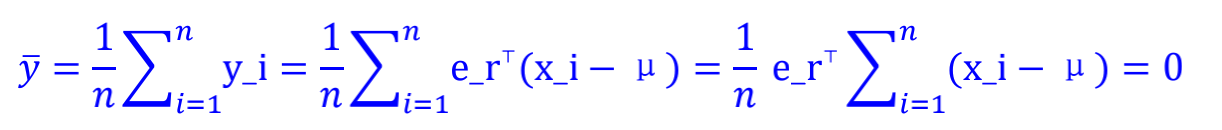
4. 设原始数据集D={x\_1, x\_2,⋯, x\_n}是d维的，采用PCA降维后，变成k维的，其中kd。设PCA找到的k个主成分依次为e\_1, e\_2,..., e\_k，则任意的d维样本x进行PCA降维后，在第r个主成分上的坐标为：，其中。因此，第i个样本在第r个主成分上的投影点可以表示为，所以，所有样本在第r个主成分上的投影点依次为y\_1, y\_2, ...y\_n，共n个投影点。

①计算这n个投影点的均值; (1分)

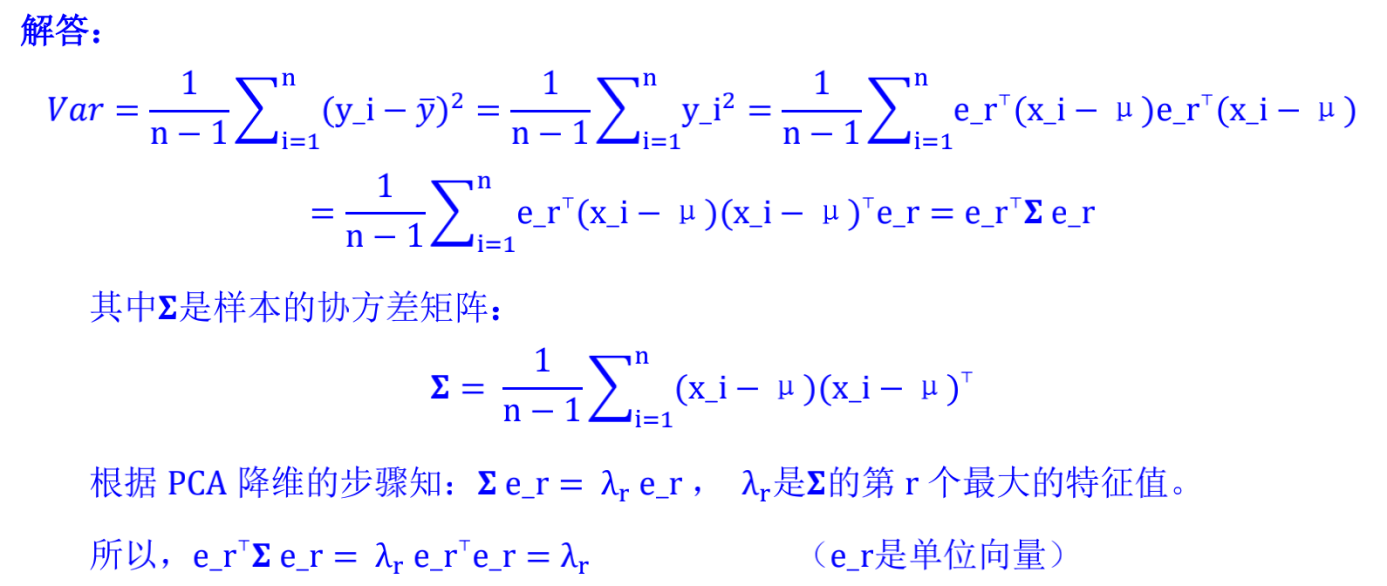
②计算这n个投影点的方差；（注意使用样本方差公式）（4分）

③利用上面的结果来说明，PCA的第一主成分就是所有样本投影到该方向上的投影点的方差最大的方向。（1分）

①**解答：**



②



（1分）

（2分）

（1分）



③**解答：**所有样本在第一个主成分上的投影点的方差为λ1，而λ1是所有特征值中最大的那个。