封

线

一、选择题（每小题 3 分,共 30 分）：

1、使用哪个库可以方便地进行数据的可视化？(　　　)

A. NumPy

B. Pandas

C. Matplotlib

D. Scikit-learn

2、考虑下面的 3×4 矩阵 A =np.array([[1,2,3,4], [5,6,7,8], [9,10,11,12]]), 表达式 A[2, :] 的结果是 (　　　).

A. array([[1, 2, 3, 4],
　　　　[5, 6, 7, 8]])

B. array([9, 10, 11, 12])

C. array([3, 7, 11])

D. array([5,6,7,8])

3、可以利用 numpy 库中的函数 array, arange, empty, linspace 等函数生成数组，以下能生成数组 array([-1. , -0.25, 0.5 , 1.25, 2.])的函数示例是 (　　　).

A. np.arange(4,dtype=float)　　B. np.arange(0,10,2,dtype=int)

C. np.empty((2,3),int)　　　　D. np.linspace(-1,2,5)

4、运算结果 np.arange(1,10).reshape(3,3).dot(np.eye(3)) = (　　).

A. array([[1, 2, 3], [4, 5, 6], [7, 8, 9]])

B. array([[1, 4, 7], [2, 5, 8], [3, 6, 9]])

C. array([[1, 0, 0], [0, 1, 0], [0, 0, 1]])

D. array([[3, 6, 9], [2, 5, 8], [1, 4, 7]])

5、线性回归模型的基本假设是什么？(　　)

A. 因变量与自变量之间是非线性关系

B. 因变量与自变量之间是线性关系

C. 因变量是离散的

D. 自变量之间必须相互独立

6、下面哪个 Pandas 函数可以用于读取 CSV 文件？(　　)

A) read.csv()　　B) load_csv()

C) read_csv()　　D) csv_read()

7、设 A 是一矩阵，则函数 np.linalg.eig(A)的返回值是 (　　).

A. A 的特征值　　B. A 的特征值

C. A 的特征值和特征向量　　D.无返回值

8、如果一个数据集有 p 个变量，那么通过 PCA 最多可以提取多少个主成分？(　　)

A. p-1

B. p

C. p+1

D. 无限多

9、如果你想通过 PCA(主成分分析)将数据集从高维降到二维以便可视化，你应该如何设置 n_components 参数？(　　)

A. n_components=2　　B. n_components='mle'

C. n_components=None　　D. n_components 设置为原数据集的维度数

10、在做距离判别分类的时，可以使用 sklearn.neighbors 模块的 KNeighborsClassifier 函数来处理，若需要创建一个指定分类类别数为 2，距离选择为马氏距离的对象的代码为 (　　).

A. KNeighborsClassifier(2, metric='euclidean')

B. KNeighborsClassifier(2, metric='manhattan')

C. KNeighborsClassifier(2, metric='minkowski')

D. KNeighborsClassifier(2, metric='mahalanobis')

二、填空题（每小题 3 分，共 30 分）：

1、样本 $x_1, x_2, \cdots, x_n$ 的平均数计算公式为: ________________________________

2、设 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ 为样本 $x_1, x_2, \cdots, x_n$ 的次序统计量，则该组样本的 p 分位数的计算公式是 $M_p =$

____

3、若 p 维总体 $\vec{X} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ 服从参数为 $N_p(\vec{\mu}, \Sigma)$ 的 p 维正态分布，则其概率密度函数为

____

4、两个维数为 p 的总体 $\vec{X}, \vec{Y}$ 协方差矩阵为________________________________

5、设变量 $\vec{X} = (X_1, X_2)$ 的均值为 $\vec{0}$，$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}$. 令 $Y = X_1 + X_2$，则 $\mathrm{Var}(Y) =$________

6、对于线性回归模型 $\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$, 设参数 $\vec{\beta}$ 的最小二乘估计为 $\hat{\vec{\beta}}$，则 $\mathrm{E}(\hat{\vec{\beta}}) =$________

7、设 $\vec{X} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ 的协方差矩阵为 $\Sigma = (\sigma_{ij})_{p \times p}, P = (\vec{e}_1, \vec{e}_2, \cdots, \vec{e}_p)$ 为由 Σ 的 p 个正交单位化特征向量为列所构成的矩阵，$\vec{Y} = (Y_1, Y_2, \cdots, Y_p)^{\mathrm{T}}$ 为 $\vec{X}$ 的 p 个主成分所构成的向量，则 $\vec{Y} =$________________________________

8、设 $\vec{x}, \vec{y}$ 是来自均值向量为 $\vec{\mu}$，协方差矩阵为 Σ 的总体 G 的两个样品,则 $\vec{x}, \vec{y}$ 之间的欧式平方距离是 $d^2(\vec{x}, \vec{y}) =$________________________________

9、已知 $\vec{X} = [x_1, x_2]^{\mathrm{T}}$ 服从二维正态分布 $N(\vec{\mu}, \Sigma)$，其中 $\vec{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$，$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$，试求点 $\alpha = [1,1]^{\mathrm{T}}$ 到总体均值的马氏距离平方________________________________

10、当一个判别准则提出以后,还要研究它的优良性,即考察它的误判率.以训练样本为

基础的误判率的估计思想如下:若属于G_1的样品被误判为属于G_2的个数为N_1个，属于G_2的样品被误判为属于G_1的个数为N_2个，两类总体的样品总数为N，则误判率P的估计为$\hat{P}=$________________________________

三、设三维随机向量$X\sim N_3(\mu,2I_3)$， 已知

$$\mu=\begin{bmatrix}1\\2\\3\end{bmatrix},\quad A=\begin{bmatrix}2&-2&1\\-1&0&-2\end{bmatrix},\quad d=\begin{bmatrix}1\\1\end{bmatrix}.$$

试求$Y=AX+d$的分布. （10 分）

四、设 $y_i=\beta_0+\beta_1x_i+\beta_2(3x_i^2-2)+\varepsilon_i(i=1,2,3), x_1=-1, x_2=0, x_3=1$，其中，$\varepsilon_1,\varepsilon_2,\varepsilon_3$ 相互独立，服从 $N(0,\sigma^2)$ 正态分布.

(1)写出矩阵 $Y=X\vec{\beta}+\vec{\varepsilon}$ 中的 X；（4 分）

(2) $\vec{\beta}=(\beta_0,\beta_1,\beta_2)^{\mathrm{T}}$ 的最小二乘估计. （6 分）

五、设随机向量 $\vec{X}=(X_1,X_2,X_3)^{\mathrm{T}}$ 的协方差矩阵为

$$\Sigma=\begin{bmatrix}1 & -3\\ -3 & 6\end{bmatrix},$$

求 $\vec{X}$ 的各主成分. （10 分）

六、设有 3 个组 G_1, G_2 和 G_3，欲判别某样品 x_0 属于何组，已知先验概率 $p_1 = 0.1, p_2 = 0.7, p_3 = 0.2$，且 $f_1(x_0) = 0.01, f_2(x_0) = 0.6, f_3(x_0) = 3$．求 x_0 属于各组的后验概率，并判别 x_0 应该属于哪一组．（10 分）