上海交通大学 2015-2016 学年第二学期

《Linux 操作系统》小测 5：Python 科学计算

题号	一	二	三	总分	阅卷人
得分					

学号:_____ 姓名:_____ 班级:_____ 系:_____

装 订 线 内 不 要 答 题

一、填空题： (共 20 空，每空 2 分，共 40 分)

1. $\lim_{x\to\infty}(1+\frac{1}{x})^x =$ _____，$\lim_{x\to 0}\frac{3\sin^2 x}{x^2} =$ _____。

2. e^x 在原点处的泰勒展开式是_____，则相应的，$\lim_{n\to\infty} 1+\frac{1}{2!}+\cdots+\frac{1}{n!} =$ _____。

3. 多元偏微分 $\frac{\partial A^T x}{\partial x} =$_____，$\frac{\partial x^T A X}{\partial x} =$ _____。

4. 函数 $f(x,y) = 3x^2+2y^2+4xy+5$ 在点 $(1,2)$ 处的梯度函数值为_____，二阶梯度函数值为_____。

5. 函数 $f(x) = x^x, x > 0$ 的最小值是_____。

6. 矩阵 $A \in \mathbb{R}^{m\times n}$ 与向量 $x \in \mathbb{R}^n$ 的乘积的几何意义是_____。

7. 矩阵 A 的主成份分析（PCA）在数学上相当于_____，主成份的意义是_____。

8. 默认标准输入的文件描述符是_____，标准输出的文件描述符是_____，而标准错误的文件描述符是_____。

9. for 有两种用法，其中一种是_____，另一种是_____。

10. 对于字符串变量 a，`${#a}` 返回的是_____；而如果 a 为数组变量，返回的则是_____。

11. "declare -r" 实现的是_____，而"declare -i" 实现的是_____

12. $x:-5、$x:=5、$x:+5 分别实现的功能为_____、_____和

二、选择题： (共 10 小题，每小题 2 分，共 20 分)

1. 关于监督模型，下面哪些论断是正确的？

A. 模型越灵活，其 bias 越小，variance 越大 B. 交叉验证法可有效避免过拟合

C. 高阶多项式模型的准确度要高于低阶模型 D. LOO 是一类特殊的交叉验证方法

2. 下面哪些模型可以作为分类器？

A. logistic 回归　　B. 线性判别分析（LDA）

C. 支持向量机（SVM）　　D. 决策树

3. 下面哪些论断是错误的？

A. 由于所有模型都存在不可约简（irreducible）误差，所以不存在唯一的最优模型

B. 训练误差恒小于测试误差

C. 可约简（reducible）误差可被分解为 bias 和 ariance

D. 提高模型的灵活性会导致 bias 降低，但 bias 升高

4. 关于 k-近邻法，哪些论断是正确的？

A. k 必须为奇数　　B. k 越大，则判别边界越是接近于非线性

C. 最优的 k 可通过交叉验证法确定　　D. k 越大，模型复杂度越高，准确性也越高

5. 对于线性回归模型，哪些论断是正确的？

A. 线性回归模型的系数可通过最小化RSS计算得到

B. 最小二乘解的方程一定通过点$(\bar{x}, \bar{y})$

C. 最小二乘解与最大似然解完全相当

D. 线性回归模型的最小二乘估计是线性无偏的

6. 对线性回归模型中的 5 水平的分类解释变量，需要为其创建多少个伪变量（dummay variable）？

A. 1　　B. 4　　C. 5　　D. 4 or 5

7. 下面哪些是 pandas.Series 和 numpy.ndarray 的差别？

A. 前者只能是一维的，后者可以是任意多维的

B. 前者可以存储不同类型的数据，后者只能存储一种类型的数据

C. 前者的 index 可以是任意的 immutable，后者只能是数值

D. 前者比后者有更广泛的应用

8. 切比雪夫不等式：设随机变量 X 的期望为 μ，方差为 σ^2，对于任意正数 ϵ，有 $P\{|X-\mu| \geq \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}$，这说明：

A. 方差越小，X 越是集中于 μ 附近

B. 方差越大，X 越是集中于 μ 附近

C. 该不等式可证明大数定理

装 订 线 内 不 要 答 题

D. 方差越小，事件 $\{|X-\mu|<\epsilon\}$ 发生的概率越大

9. 若 **A** 为 n 阶实对称阵，则

A. **A** 的所有特征值都是正值

B. **A** 可逆

C. **A** 的奇异值全部为正

D. **A** 可以进行变换 $P^{-1}AP=P^TAP=\Lambda$，其中 Λ 是以 A 的特征值为对角元的对角阵

三、解答题： (共 20 小题, 每小题 3 分, 共 60 分)

1. 一种罕见病，其人群发病率为 5%；一种仍在临床试验阶段的诊断方法，对真正患者的误诊率为 8%，而对健康人的误诊率为 4%。A 被诊断为阴性，则 A 患该病的几率是多少？

2. 谈谈 numpy ndarray 与 list 的区别和联系。

3. 请问矩阵的奇异值分解（SVD）有哪些应用？

4. 什么是向量的线性组合（linear combination）？仿射组合 (affine combination)？凸组合 (convex combination)？

5. 怎么判断一个函数是否凸函数 (convex function)？一个凸函数是否一定存在一个全局最小值（global minimum）？

6. 如何判断一个方阵是否是正定或者半正定阵？

7. 说说 Naive Bayes 的原理，探讨一下其优缺点。

8. 请问 Logistic regression 是如何采用最大似然法进行求解的？

9. 谈谈最小二乘法（ordinary least squares）与正交投影的联系。

10. 梯度下降法和牛顿法在求解最优化问题时各有什么样的优缺点？请举例说明。

11. pandas 模块有哪几种常见的数据类，在处理数据方面有哪些优势？

12. 用 Poisson 分布拟合 RNA-seq 数据时存在哪些不足？如何去解决这些问题？

13. 谈谈重抽样（resampling）在数据分析中的作用，有哪几种重抽样方法？

14. 写出信息熵（entropy）和信息增量（Information gain）的计算公式，谈谈其在决策树分析中的作用。

15. 决策树的剪枝选择算法中，剪枝系数 α 是如何计算的？如何选择需要去除的分枝，选择最优的决策树？

装 订 线 内 不 要 答 题