华东师范大学期末试卷（A）

2021-2022 学年第一学期

课程名称：统计方法与机器学习

学生姓名：_____　　学　　号：_____

专　　业：_____　　年级/班级：_____

课程性质：专业必修课

一	二	三	四	五	六	总分	阅卷人签名

一、（本题共 20 分）

表 1 是一个不完整的双因素方差分析表。

表 1 不完整的双因素方差分析表

来源	自由度	平方和	均方	F统计量	p 值
因素 A	/	/	0.0833	0.05	0.952
因素 B	/	96.333	96.333	57.80	<0.001
交互效应 AB	2	12.167	6.0833	3.65	/
误差	6	10	/		
汇总	11	118.667			

请根据表 1 回答以下问题：

1.（2 分） 因素 A 的平方和SS_A是___ 。

2.（2 分） 因素 A 的自由度为___ 。

3.（2 分） 在实验中，因素 B 的水平数为___ 。

4.（2 分） 均方误差为___ 。

5.（2 分） 在这个实验中，每种组合的重复次数为___ 。

6.（5 分） 如何计算交互效应的p值？并给定显著性水平$\alpha = 0.05$时，简述如何判断交互效应的显著性。

7.（5 分） 证明：在双因素方差分析中，$SS_T = SS_A + SS_B + SS_{AB} + SS_E$ 。

二、（本题共 20 分）

现有一个数据集，其中包含 400 条观测，每条观测有 1 个因变量y以及 20 个中心化后的特征$x_1, x_2, \cdots, x_{20}$。前 5 行数据如图 2 所示。

	y	X1	X2	X3	X4	X5	X6	X7	X8	X9	...	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
0	-1.24	0.31	-0.95	-0.99	-0.58	-0.47	0.70	-0.88	1.05	0.03	...	-1.42	-0.85	-0.37	-0.52	-0.19	-1.75	0.78	-0.78	0.23	0.05
1	-0.72	-0.04	-2.58	1.32	-0.75	-0.92	1.43	-1.85	-0.83	1.36	...	0.31	-0.01	-0.49	-1.20	0.08	-0.83	-0.98	2.89	-0.79	-0.82
2	6.40	-0.89	0.91	-0.07	0.14	1.31	0.60	0.34	-0.96	1.67	...	1.58	0.44	1.80	-0.04	1.65	-0.06	-1.01	-0.87	-0.41	-1.03
3	1.10	1.81	0.20	-0.70	-1.03	0.72	-0.89	1.56	-0.03	-0.28	...	0.61	0.01	-1.39	-0.78	-1.20	-2.09	-0.70	-0.73	-1.96	0.48
4	2.33	0.15	-0.27	-0.82	-0.21	0.42	-0.15	-0.04	0.80	2.55	...	-0.44	-0.47	-1.08	-2.31	0.87	-0.62	1.10	1.02	1.26	0.58

图 2 前 5 条数据的示意图

取显著性水平 $\alpha = 0.05$，现回答以下问题：

1.（5 分）同学 A 想构建利用X_1来预测y，从而构建了一个一元线性回归模型。请根据图 3 中 Python 运行的结果，写出一元线性回归模型，并从一个角度阐述该模型是否显著。

Dep. Variable:	y	**R-squared:**	0.127
Model:	OLS	**Adj. R-squared:**	0.125
Method:	Least Squares	**F-statistic:**	58.04
Date:	Fri, 02 Dec 2022	**Prob (F-statistic):**	1.90e-13
Time:	15:55:38	**Log-Likelihood:**	-1120.5
No. Observations:	400	**AIC:**	2245.
Df Residuals:	398	**BIC:**	2253.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.6954	0.200	3.482	0.001	0.303	1.088
X1	1.6034	0.210	7.618	0.000	1.190	2.017

Omnibus:	2.449	**Durbin-Watson:**	2.120
Prob(Omnibus):	0.294	**Jarque-Bera (JB):**	2.200
Skew:	-0.093	**Prob(JB):**	0.333
Kurtosis:	2.688	**Cond. No.**	1.05

图 3 Python 的运行结果（一个特征）

2.（5 分）根据图 3 中 Python 运行的结果，请给出当X_1的取值为 0.5 时，y的点预测。同时，阐述如何计算其$1-\alpha$的预测区间。

3.（5 分）同学 B 将特征X_1和X_2同时纳入线性回归模型，并利用 Python 得到结果，如图 4 所示。将图 3 和图 4 进行比较，发现在线性回归模型中R^2从 0.127 提升到了 0.320，即结果为$R^2_{mod_1} = 0.127 \leq R^2_{mod_2} = 0.320$。请问这个结论是否普

遍存在？如果是，请证明它；如果不是，请举出反例。

Dep. Variable:	y	**R-squared:**	0.320
Model:	OLS	**Adj. R-squared:**	0.317
Method:	Least Squares	**F-statistic:**	93.54
Date:	Fri, 02 Dec 2022	**Prob (F-statistic):**	5.19e-34
Time:	17:04:30	**Log-Likelihood:**	-1070.5
No. Observations:	400	**AIC:**	2147.
Df Residuals:	397	**BIC:**	2159.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	0.6500	0.177	3.682	0.000	0.303	0.997
X1	1.4421	0.187	7.728	0.000	1.075	1.809
X2	1.8526	0.174	10.618	0.000	1.510	2.196

Omnibus:	2.166	**Durbin-Watson:**	2.004
Prob(Omnibus):	0.339	**Jarque-Bera (JB):**	2.246
Skew:	-0.169	**Prob(JB):**	0.325
Kurtosis:	2.859	**Cond. No.**	1.12

图 4 Python 的运行结果（两个特征）

4.（5 分）经验所知，R^2越大表明特征的拟合效果越好。于是，同学 C 逐一将特征放入线性回归模型中。具体方案是，第一个模型的特征是X_1；第二个模型的特征是X_1和X_2；第三个模型的特征是X_1, X_2和X_3，以此类推。结果发现R^2的数值如表 1 所示。

表 1 20 个模型中不同特征维度下的R^2值

维度	1	2	3	4	5	6	7	8	9	10
R^2	0.127	0.320	0.495	0.568	0.637	0.707	0.779	0.841	0.902	0.948
维度	11	12	13	14	15	16	17	18	19	20
R^2	0.949	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.950	0.952

请问，R^2是否适合作为模型选择的指标？并请说明理由。如果不是，请给出一个改进方案。

三、（本题共 10 分）

请阐述一下，如何诊断出数据中存在多重共线性？（提示：只需要提供一种完整的方案）。

四、（本题共 15 分）

比较感知机和线性 SVM 的损失函数。

五、（本题共 10 分）

1.（5 分）解释生成式模型和判别式模型，并分析二者的不同点；

2. 列出三种判别式模型（3 分）和两种生成式模型（2 分）。

六、（本题共 25 分）

考虑利用线性支持向量机对如下两类可分数据进行分类：

+1：(1,1), (2,2), (2,0)

-1：(0,0), (1,0), (0,1)

1.（8 分）在图中做出这 6 个训练点，构造具有最优超平面和最优间隔的权重向量；

2.（4 分）哪些是支撑向量？

3.（13 分）通过寻找拉格朗日乘子来构造在对偶空间的解，并将它与第一小问中的结果比较。