Name:

USC ID:

Notes:

- Write your name and ID number in the spaces above.
- No books, cell phones or other notes are permitted. Only one letter size cheat sheet (back and front) and a calculator are allowed.
- Problems are not sorted in terms of difficulty. Please avoid guess work and long and irrelevant answers.
- Show all your work and your final answer. Simplify your answer as much as you can.
- Open your exam only when you are instructed to do so.

Problem	Score	Earned
1	25	
2	30	
3	25	
4	25	
5	15	
Total	120	

1. A statistician is working on the amount of funding that companies obtain on a crwod-sourcing website and has developed the following model. She used 26 companies to obtain the model

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5$$
$$\hat{y} = 964.8 + 700.2x_1 + 317.5x_2 - 200.2x_3 + 15.3x_4 + 17.1x_5$$

The standard errors are:

$$s_{b_1} = 12.$$
$$s_{b_2} = 22.5$$
$$s_{b_3} = 101.8$$
$$s_{b_4} = 45.3$$
$$s_{b_5} = 2.3$$

- $\hat{y}$: the amount of funding obtained by a company in 1000 dollars
- x_1: the average annual salary of the founders
- x_2: the number of employees the startup hired
- x_3: a dummy vaiable that is 1 when the company's field is information technology and 0 otherwise
- x_4: the age of the company
- x_5 is a dummy variable taking value 1 if the founders had previous failures and 0 otherwise

(a) Interpret the estimated coefficients $b_0 = 964.8$ and $b_5 = 17.1$.

(b) Test, at the 5% level, the null hypothesis that the true coefficient on the dummy variable x_3 is 0 against the alternative that it is not 0.

(c) Find and interpret a 80% confidence interval for the parameter β_1.

(d) If for the model, SSR=18147.5 (Regression Sum of Squares) and SSE = 17136.5 (Residual Sum of Squares), test the hypothesis that all the coefficients of the model are 0 (test overal significance of the model) using $\alpha = 1\%$.

(e) Calculate R^2 for the regression line.

2. Although in practice, we usually model conditional distributions of features in each class as Gauusians for Bayesian Discriminant Analysis, one must notice that Bayesian Discriminant Analysis is doable using any type of distribution. Assume that we wish to perform Discriminant Analysis with only one positive predictor $x \geq 0$ (i.e. $p = 1$), but instead of normal class pdfs, we have Lognormal class pdfs whose formula is: (40 pts)

$$f_k(x) = \frac{1}{x\sqrt{2\pi}\sigma_k} \exp\left(\frac{-(\ln x - \mu_k)^2}{2\sigma_k^2}\right), x \geq 0$$

We know that the posterior class probability for class k is:

$$\Pr(Y = k|X = x) = p_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

where π_l's are prior class probabilities.

(a) Determine the simplest form for the discriminant score $\delta_k(x)$ if we assume that $\sigma_1 = \sigma_2 = \ldots = \sigma_K = \sigma$; that is, there is a shared parameter $\sigma_k = \sigma$ across all K classes. Are the discriminant scores linear functions of x in this case? Hint: follow the same procedure as the one we used for Gaussian pdfs.

(b) For binay classification ($K = 2$), when classes are balanced ($\pi_1 = \pi_2$) and $\sigma_1 = \sigma_2 = \sigma$, show that the decision boundary between the two classes is the threshold

$$x^* = \sqrt{\exp(\mu_1)}\sqrt{\exp(\mu_2)}$$

3. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right) \text{ subject to } \sum_{\mathrm{j}=1}^{\mathrm{p}}|\beta_{\mathrm{j}}| \leq \mathrm{s}$$

for a particular value of s. Indicate which of the following is true for items in 3a-3e, as we increase s from zero to infinity, and justify your answers:

- Increase initially, and then eventually start decreasing in an inverted U shape.
- Decrease initially, and then eventually start increasing in a U shape.
- Steadily increase.
- Steadily decrease.
- Remain constant.

(a) Training RSS

(b) Test RSS.

(c) Variance.

(d) (Squared) bias.

(e) Irreducible error.

4. Consider the novel logistic regression method for binary classification ($Y = 0$ or $Y = 1$) with two features $\mathbf{X} = (X_1, X_2)$, formulated by

$$\mathrm{P}(Y = 1|\mathbf{X}) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_1^2+\beta_3 X_2}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_1^2+\beta_3 X_2}}$$

Assume that using a dataset of 200 observations, we obtained the following estimates:

	Coefficient	Standard Error
β_0	1	0.2
β_1	2	0.1
β_2	1	s
β_3	1	0.5

(a) Determine the equation for the decision boundary for this classifier.

(b) Sketch the decision boundary for this classifier and clearly show the regions for the positive class and the negative class.

(c) Find all values of s that makes the coefficient β_2 statistically insignificant. You can consider the significance level to be $\alpha = 0.05$.

5. The Federalist papers, authored by Alexander Hamilton, John Jay, and James Madison, consist a series of 85 papers published between October 1787 and April 1788 under the pseudonym PUBLIUS to convince the people of New York to ratify the US constitution. The authorship of some of the papers is in dispute. In particular, the authorship of 12 of the papers is disputed, while Hamilton and Madison later published their lists of authors of the rest, although even those lists have discrepancies. One can use Machine Learning algorithms to classify the disputed papers using papers with known authors.
.

   (a) Formulate the above problem as a multi-class problem. How many classes does their outcome Y have? What are those classes?

   (b) Some experts believe that some of the papers are collaborative efforts of two or sometimes all three of Hamilton, Jay, and Madison. Capturing multi-author papers can be done by formulating the problem as a multi-label problem. Explain what the labels are, if the labels are binary, and how the algorithm should label a paper that was solely written by Jay, a paper that was written by Hamilton and Madison, and a papert that was the collaborative work of Hamilton, Jay, and Madison.

Scratch paper

Name:

USC ID:

Scratch paper

Name:

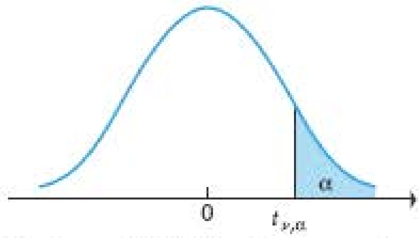
USC ID:

Cumulative Distribution Function, $F(z)$, of the Standard Normal Distribution Table

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09

Cumulative Distribution Function, $F(z)$, of the Standard Normal Distribution Table

Upper Critical Values of Student's t Distribution with ν Degrees of Freedom



For selected probabilities, α, the table shows the values $t_{\nu,\alpha}$ such that $P(t_\nu > t_{\nu,\alpha}) = \alpha$, where t_ν is a Student's t random variable with ν degress of freedom. For example, the probability is .10 that a Student's t random variable with 10 degrees of freedom exceeds 1.372.

	Probability of Exceeding the Critical Value					
ν	0.10	0.05	0.025	0.01	0.005	0.001
1	3.078	6.314	12.706	31.821	63.657	318.313
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.782
8	1.397	1.860	2.306	2.896	3.355	4.499
9	1.383	1.833	2.262	2.821	3.250	4.296
10	1.372	1.812	2.228	2.764	3.169	4.143
11	1.363	1.796	2.201	2.718	3.106	4.024
12	1.356	1.782	2.179	2.681	3.055	3.929
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
60	1.296	1.671	2.000	2.390	2.660	3.232
100	1.290	1.660	1.984	2.364	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.090
ν	0.10	0.05	0.025	0.01	0.005	0.001

F Table for $\alpha = 0.01$

	DF1	α = 0.01																	
DF2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**Inf**
1	4052.2	4999.5	5403.4	5624.6	5763.7	5859	5928.4	5981.1	6022.5	6055.8	6106.3	6157.3	6208.7	6234.6	6260.6	6286.8	6313	6339.4	6365.
2	98.503	99	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399	99.416	99.433	99.449	99.458	99.466	99.474	99.482	99.491	99.49
3	34.116	30.817	29.457	28.71	28.237	27.911	27.672	27.489	27.345	27.229	27.052	26.872	26.69	26.598	26.505	26.411	26.316	26.221	26.12
4	21.198	18	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.374	14.198	14.02	13.929	13.838	13.745	13.652	13.558	13.46
5	16.258	13.274	12.06	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.888	9.722	9.553	9.466	9.379	9.291	9.202	9.112	9.0
6	13.745	10.925	9.78	9.148	8.746	8.466	8.26	8.102	7.976	7.874	7.718	7.559	7.396	7.313	7.229	7.143	7.057	6.969	6.8
7	12.246	9.547	8.451	7.847	7.46	7.191	6.993	6.84	6.719	6.62	6.469	6.314	6.155	6.074	5.992	5.908	5.824	5.737	5.6
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.667	5.515	5.359	5.279	5.198	5.116	5.032	4.946	4.85
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.111	4.962	4.808	4.729	4.649	4.567	4.483	4.398	4.31
10	10.044	7.559	6.552	5.994	5.636	5.386	5.2	5.057	4.942	4.849	4.706	4.558	4.405	4.327	4.247	4.165	4.082	3.996	3.90
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.397	4.251	4.099	4.021	3.941	3.86	3.776	3.69	3.60
12	9.33	6.927	5.953	5.412	5.064	4.821	4.64	4.499	4.388	4.296	4.155	4.01	3.858	3.78	3.701	3.619	3.535	3.449	3.36
13	9.074	6.701	5.739	5.205	4.862	4.62	4.441	4.302	4.191	4.1	3.96	3.815	3.665	3.587	3.507	3.425	3.341	3.255	3.16
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.14	4.03	3.939	3.8	3.656	3.505	3.427	3.348	3.266	3.181	3.094	3.00
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.666	3.522	3.372	3.294	3.214	3.132	3.047	2.959	2.86
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.89	3.78	3.691	3.553	3.409	3.259	3.181	3.101	3.018	2.933	2.845	2.75
17	8.4	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593	3.455	3.312	3.162	3.084	3.003	2.92	2.835	2.746	2.65
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.371	3.227	3.077	2.999	2.919	2.835	2.749	2.66	2.56
19	8.185	5.926	5.01	4.5	4.171	3.939	3.765	3.631	3.523	3.434	3.297	3.153	3.003	2.925	2.844	2.761	2.674	2.584	2.48
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.231	3.088	2.938	2.859	2.778	2.695	2.608	2.517	2.42
21	8.017	5.78	4.874	4.369	4.042	3.812	3.64	3.506	3.398	3.31	3.173	3.03	2.88	2.801	2.72	2.636	2.548	2.457	2.3
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258	3.121	2.978	2.827	2.749	2.667	2.583	2.495	2.403	2.30
23	7.881	5.664	4.765	4.264	3.939	3.71	3.539	3.406	3.299	3.211	3.074	2.931	2.781	2.702	2.62	2.535	2.447	2.354	2.25
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	3.032	2.889	2.738	2.659	2.577	2.492	2.403	2.31	2.21
25	7.77	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	2.993	2.85	2.699	2.62	2.538	2.453	2.364	2.27	2.16
26	7.721	5.526	4.637	4.14	3.818	3.591	3.421	3.288	3.182	3.094	2.958	2.815	2.664	2.585	2.503	2.417	2.327	2.233	2.13
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062	2.926	2.783	2.632	2.552	2.47	2.384	2.294	2.198	2.09
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.12	3.032	2.896	2.753	2.602	2.522	2.44	2.354	2.263	2.167	2.06
29	7.598	5.42	4.538	4.045	3.725	3.499	3.33	3.198	3.092	3.005	2.868	2.726	2.574	2.495	2.412	2.325	2.234	2.138	2.03
30	7.562	5.39	4.51	4.018	3.699	3.473	3.304	3.173	3.067	2.979	2.843	2.7	2.549	2.469	2.386	2.299	2.208	2.111	2.00
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.665	2.522	2.369	2.288	2.203	2.114	2.019	1.917	1.80
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.496	2.352	2.198	2.115	2.028	1.936	1.836	1.726	1.60
120	6.851	4.787	3.949	3.48	3.174	2.956	2.792	2.663	2.559	2.472	2.336	2.192	2.035	1.95	1.86	1.763	1.656	1.533	1.38
Inf	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321	2.185	2.039	1.878	1.791	1.696	1.592	1.473	1.325	1