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.
- The exam has 5 questions, 9 pages, and 20 points extra credit. However, your grade cannot exceed 100/100.

Problem	Score	Earned
1	25	
2	30	
3	25	
4	20	
5	20	
Total	120	

1. For each Major League Baseball team we have the number of wins (`Wins`) and the total player salary in millions of dollars (`Salary`) for 2006. (You don't need to know anything about baseball for this question.) The total league payroll was $2,326.707 million. For each team i, define

$$\texttt{SalaryShare}_i = \frac{\texttt{Salary}_i}{\sum_{j=1}^{n} \texttt{Salary}_j} = \frac{\texttt{Salary}_i}{2,326.707}$$

Now consider the following summary.

```
Call:
lm(formula = Wins ~ SalaryShare)

Residuals:
     Min       1Q   Median       3Q      Max
-17.7907  -4.5503   0.3654   4.5352  17.4042

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   67.982      4.178  16.271  8.4e-16 ***
SalaryShare  389.540    116.013   3.358  0.00228 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.665 on 28 degrees of freedom
Multiple R-squared:  0.2871,    Adjusted R-squared:  0.2616
F-statistic: 11.27 on 1 and 28 DF,  p-value: 0.002277
```

(a) But suppose that instead of regressing `Wins` on `SalaryShare` we used `Salary` itself as the input. Use the summary above to compute the estimates of the intercept $\hat{\beta}_0$, the slope $\hat{\beta}_1$, and the R^2 value for this hypothetical regression.

(b) Do we have reason to believe in a linear relationship between `Wins` and `Salary`, in the hypothetical regression in part 1a? State a formal hypothesis test, the value of the test statistic, and the conclusion. Use $\alpha = 0.05$

2. In a classification problem with two classes and two features, the joint distribution of the features in each class is:

$$f_k(x_1, x_2) = \frac{1}{2\pi\sqrt{(1-k/4)}} \exp\left[-\frac{z}{2(1-k/4)}\right], \ k = 1, 2$$

where

$$z = (x_1 - k)^2 - \sqrt{k}(x_1 - k)(x_2 - k^2) + (x_2 - k^2)^2$$

(a) Assuming that the prior probability of class one is twice the prior probability of class two, in what class is the point $(X_1, X_2) = (1, 5)$ is classified?

(b) The marginal distributions of features in each class can be calculated from the joint distributions, and are:

$$f_k(x_1) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x_1 - k)^2}{2}\right]$$
$$f_k(x_2) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x_2 - k^2)^2}{2}\right]$$

The Naïve Bayes assumption clearly does not hold in this problem. However, classify $(X_1, X_2) = (1, 5)$ pretending the Naïve Bayes assumption holds and compare the results with part 2a.

3. Consider a logistic regression problem in which there are no features, which means that:

$$\Pr(Y = 1) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

Assume that we have m data points with label $Y = 1$ and n data points with label $Y = 0$ (remember that features are irrelevant).

(a) Write down the likelihood function $l(\beta_0)$.

(b) Find the Maximum Likelihood estimate $\hat{\beta}_0$ for this data set. [**Hint**: maximize $\log_e l(\beta_0)$].

(c) Determine conditions under which this simple classifier classifies data points into $Y = 1$ or $Y = 0$.

4. Let us consider a data set containing 50 positive and 50 negative instances, where the attributes contain no information about the class labels. Hence, the generalization error rate of any classification model learned over this data is expected to be 0.5. Let us consider a classifier that assigns the majority class label of training instances (ties resolved by using the positive label as the default class) to any test instance, irrespective of its attribute values. We can call this approach *the majority inducer* classifier. Determine the error rate of this classifier using the following methods.

   (a) Leave-one-out cross validation.

   (b) 2-fold stratified cross-validation, where the proportion of class labels at every fold is kept same as that of the overall data.

   (c) From the results above, which method provides a more reliable evaluation of the classifier's generalization error rate?

5. Consider the dataset presented in the following table for classification of loan defaults. Given a pair of categorical attribute values, V_1 and V_2, the distance between them is defined as follows:

$$d_M(V_1, V_2) = \sum_{i=1}^{k} \left| \frac{n_{i1}}{n_1} - \frac{n_{i2}}{n_2} \right|$$

where n_{ij} is the number of examples from class i with attribute value V_j and n_j is the number of examples with attribute value V_j.

The distance between a test point $X^* = (HO^*, MS^*, An^*)$ and training point $X = (HO, MS, An)$ is defined as $d(X^*, X) = d_M(HO^*, HO) + d_M(MS^*, MS) + |An^* - An|$, where HO, MS, An respectively stand for Home Owner. Marital Status, and Annual Income in \$1K. How is the test point $(Yes, Single, 110K)$ classified using 3-, 5-, 7-nearest neighbors and $d(X^*, X)$?

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Scratch paper

Name:

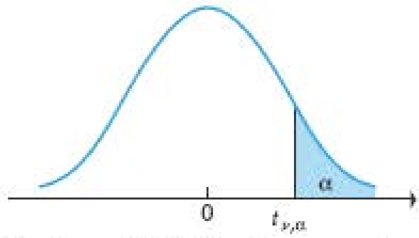
USC ID:

Scratch paper

Name:

USC ID:

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