Midterm

INST 737
April 1, 2013

Name: $\underbrace{\hspace{12cm}}_{\text{by writing my name I swear by the honor code}}$

Read all of the following information before starting the exam:

- For free response questions, show all work, clearly and in order, if you want to get full credit. I reserve the right to take off points if I cannot see how you arrived at your answer (even if your final answer is correct).
- This is exam is due April 1, 2013 at the end of class.
- This exam is open book, open notes. You may use a calculator and a one page (front and back) hand-written page of notes.
- If your native language is not English (or you think it would be useful for spelling), you may bring a paper dictionary. The dictionary may not be specialized (e.g. a "math" dictionary).
- You may use a calculator, but not any device that can access the Internet or store large amounts of data.
- Justify your answers algebraically whenever possible to ensure full credit. Be sure to have units for all answers that call for them.
- Please keep your written answers brief; be clear and to the point. I will take points off for rambling and for incorrect or irrelevant statements.
- Complete all problems; they are not worth equal points, and you cannot get partial credit if you don't try.
- Good luck!

Page	Points	Score
2	30	
3	18	
4	18	
5	18	
6	30	
7	30	
8	6	
9	50	
10	50	
11	50	
Total:	300	

1. (6 points) A die is loaded in such a way that the probability of each face turning up is proportional to the number of dots on that face. (For example, a six is three times as probable as a two.) What is the probability of getting an odd number in one throw?

   A. $\frac{9}{21}$
   B. $\frac{12}{21}$
   C. $\frac{1}{2}$
   D. $\frac{17}{36}$

1. ______________

2. (6 points) What is ggplot2?

   A. An R package that creates graphs and visualizations
   B. An R package that does text analysis
   C. A python package for reading csv files
   D. An open source package for printing files

2. ______________

3. (6 points) Which of these is a reasonable *categorical* representation of location?
   A. Zip / postal code B. Latitude C. Longitude D. Elevation

3. ______________

4. (6 points) One coin in a collection of 9 has two heads. The rest are fair. If a coin, chosen at random from the lot and then tossed, turns up heads 3 times in a row, what is the probability that it is the two-headed coin?
   A. $\frac{1}{2}$ B. $\frac{1}{3}$ C. $\frac{7}{8}$ D. 1

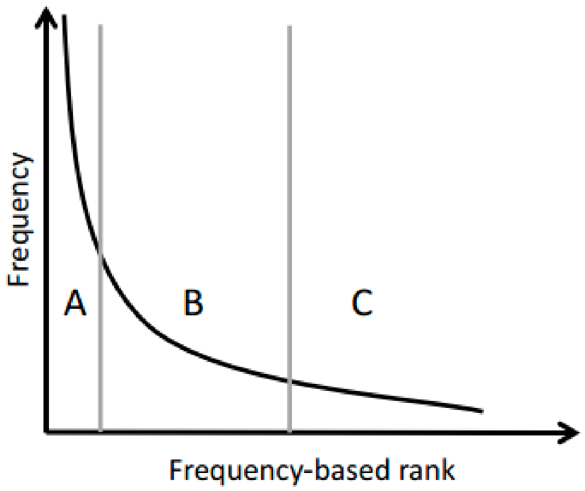
4. ______________

5. (6 points) Suppose you have a one-dimensional regression coefficients (β_0, β_1) learned using OLS from three observations, $\{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$, which give predictions $\hat{y}_1 = \beta_0 + \beta_1 x_1, \hat{y}_2 = \beta_0 + \beta_1 x_2$, and $\hat{y}_3 = \beta_0 + \beta_1 x_3$. Suppose that $\dot{y}_1 = \gamma_1 x_1$, $\dot{y}_2 = \gamma x_2$, and $\dot{y}_3 = \gamma x_3$. No matter what γ is and how it was chosen, which of the following must always be true?

   A. $\sum_i (y_i - \dot{y}_2)^2 \geq \sum_i (y_i - \hat{y}_i)^2$
   B. $\sum_i (y_i - \dot{y}_i) \geq \sum (y_i - \hat{y}_i)$
   C. $\sum_i (\dot{y}_i - y_i)^3 \geq \sum_i (y_i - \hat{y}_i)^3$
   D. $\beta_0 = x_1 / \gamma_1$

5. ______________

______________ **/ 30 points**



6. (6 points) Above is a plot of features and their frequency in the corpus. I have drawn three groups of features (A, B, C). Which of the following are words that are likely to come from these classes in normal English training data?

   A. A: the, B: apple, C: marzipan

   B. A: apple, B: the, C: marzipan

   C. A: the, B: marzipan, C:apple

   D. C: marzipan, B: apple, C: the

6. ______________

7. (6 points) (Also based on the above figure) Which bin is likely to have the most useful features?

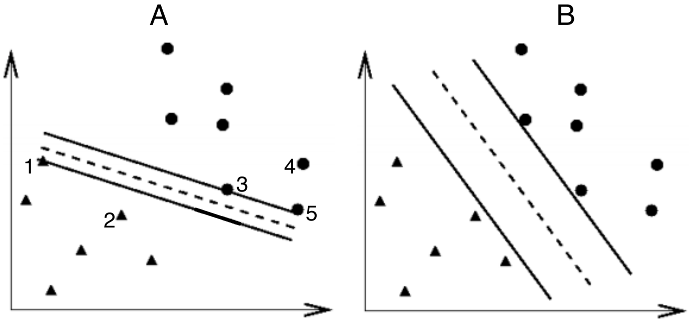
   A. B: These features don't appear in every document (and thus can discriminate) but appear in enough documents (and thus can generalize)

   B. C: You can be very sure of the class for these words because they don't appear in many documents

   C. A: You can get accurate statistics for these words

   D. C: These features will likely not appear in test data

7. ______________



8. (6 points) For the above figure, which classifier has the larger margin and why?

______________ **/ 18 points**

A. B: The distance between the separating hyperplane and the nearest examples is largest

B. **B**: The separating hyperplane has a more negative slope

C. **A**: The support vectors *in the same class* are further apart

D. **A**: There are more support vectors

8. ______________

9. (6 points) Also going by the above figure, which points are support vectors for Classifier A?

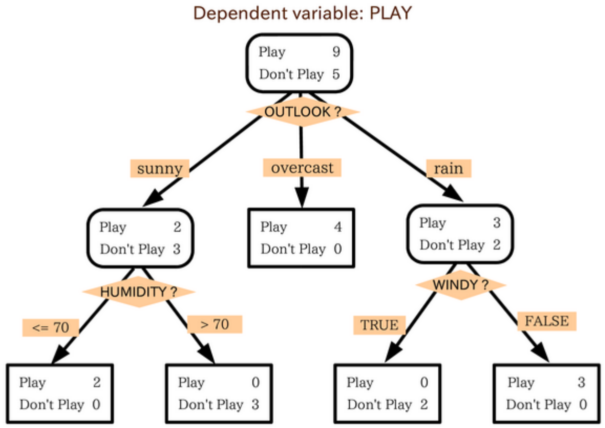
A. 1,3

B. 3,4,5

C. 1,2

D. 2,4

9. ______________



10. (6 points) Above is a decision tree for deciding whether to play tennis or not. If the wind is *WINDY*, the humidity is *grater than seventy*, and the outlook is *overcast*, then should you play or not?

A. Yes; there are 100% "play" examples in the corresponding node

B. No; there are 56% "don't play" examples in the corresponding nodes

C. No; there are 100% "don't play" examples in the corresponding nodes

D. Yes; there are 64% "play" examples in the corresponding nodes

10. ______________

Color	Probability
Red	0.1
Blue	0.4
Greed	0.5

11. (6 points) Given the distribution over colors given above, what is the entropy of that distribution?

A. $-0.1\lg(0.1) - 0.4\lg(0.4) - 0.5\lg(0.5) = 1.4$

B. $0.1\lg(0.1) + 0.4\lg(0.4) + 0.5\lg(0.5) = -1.4$

______________ **/ 18 points**

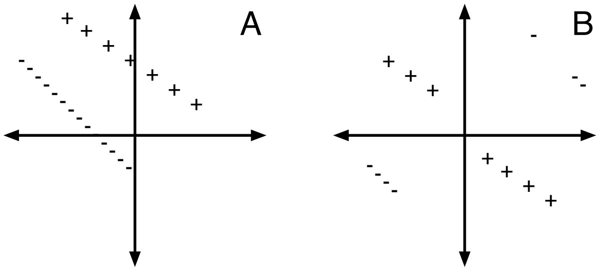
C. $(0.1 + 0.4 + 0.5) * \lg(0.1 + 0.4 + 0.5) = 0.0$
D. $-0.5 * \lg(0.5) = 0.5$

11. ______________

12. (6 points) What are reasons to use decision trees?

A. They give interpretable decisions and can fit non-linear functions.
B. They are linear classifiers that are easy to fit.
C. They find max-margin classifiers for a dataset.
D. The optimize the RSS.

12. ______________



13. (6 points) Based on the above image, for which dataset would a linear kernel SVM work best?

A. A, a hyperplane through the origin would completely divide the data
B. A, there are an unbalanced number of positive and negative examples
C. B, there are a balanced number of positive and negative examples
D. B, the data are non-linear

13. ______________

14. (6 points) Consider the following statements about what will happen when I play tennis champion Roger Federer in a tennis set:

1. Federer will win the match.
2. Federer will lose the first set.
3. Federer will lose the first set but win the match.
4. Federer will win the first set but lose the match.

With associated probabilities P_1, P_2, P_3, P_4. Which of the following is a mathematically possible ordering of the probabilities?

A. $P_2 > P_4 > P_1 > P_3$
B. $P_1 > P_3 > P_2 > P_4$
C. $P_2 > P_3 > P_4 > P_1$
D. $P_3 > P_2 > P_1 > P_4$

14. ______________

Consider the following naïve Bayes training set to distinguish spam from non-spam. We have observed two spam e-mails:

______________ **/ 18 points**

- buy pharmacy
- earn money

and one non-spam e-mail:

- overdue book

Assume that we use Laplace (add one) smoothing and that the vocabulary size is six (these words).

15. (6 points) For the above dataset, what label would we apply to an empty test document?

    A. Spam
    B. Non-spam
    C. Both classes are equally likely
    D. Naïve Bayes cannot handle empty documents

15. ______________

16. (6 points) For the above dataset, which of the following is true?

    A. $p(\mathbf{buy}|\mathbf{SPAM}) < p(\mathbf{book}|\mathbf{NOTSPAM})$
    B. $p(\text{buy}|\text{SPAM}) > p(\text{book}|\text{NOTSPAM})$
    C. $p(\text{buy}|\text{SPAM}) > p(\text{earn}|\text{SPAM})$
    D. $p(\text{buy}|\text{NOTSPAM}) > p(\text{book}|\text{NOTSPAM})$

16. ______________

17. (6 points) For the above dataset, what is the probability $p(\text{buy}|\text{NOTSPAM})$?
    A. $\frac{1}{8}$ B. $\frac{1}{3}$ C. $\frac{7}{36}$ D. $\frac{1}{6}$

17. ______________

For the next two questions, assume that you learn a linear regression on the resale value of a car. You model it as

$$\text{resale} = \beta_0 + \beta_1 \text{years}, \tag{1}$$

where "years" represents the number of years the car has been owned.

18. (6 points) Suppose that β_1 is a large negative number. What does this mean?

    A. Cars lose value the longer you own them.
    B. Classic cars grow in value.
    C. You shouldn't own a car longer than five years.
    D. Cars don't change value the longer you own them.

18. ______________

19. (6 points) What does β_0 represent?

    A. The resale value of a new car
    B. The resale value of the oldest car
    C. The average resale value of all cars
    D. The maximum resale value of all cars

______________ **/ 30 points**

19. ______________

The next two questions are about the following object in R.

```
> d
  foo bar
1   4  10
2   5  12
3   6  14
```

20. (6 points) What is $d\$bar[2]$?
A. 12 B. 10 C. 14 D. 5

20. ______________

21. (6 points) What is $sum(d\$foo)$?
A. 15 B. 51 C. 36 D. 17

21. ______________

The next two question are about the following logistic regression classification model that distinguishes American ($Y = 1$) from British documents ($Y = 0$). For reference, the logistic regression equations are

$$P(Y = 0|X) = \frac{1}{1 + \exp\left[w_0 + \sum_i w_i X_i\right]} \tag{2}$$

$$P(Y = 1|X) = \frac{\exp\left[w_0 + \sum_i w_i X_i\right]}{1 + \exp\left[w_0 + \sum_i w_i X_i\right]} \tag{3}$$

feature	symbol	weight
bias	w_0	1.0
"trousers"	w_1	-1.0
"color"	w_2	2.0
"kerb"	w_3	-0.5
"elevator"	w_4	1.0

22. (6 points) For which of the following documents are the class probabilities perfectly balanced?
A. {color, trousers, trousers, kerb, kerb} B. {elevator, kerb, kerb} C. {} D. {kerb}

22. ______________

23. (6 points) Which of the following documents has the highest probability of being from an American document?
A. {} B. {trousers} C. {kerb} D. {trousers, kerb, elevator}

23. ______________

The next two questions are based on the following confusion matrix.

	Truth	
Prediction	True	False
True	40	30
False	10	20

24. (6 points) What is the accuracy of this classifier?
A. 60% B. 30% C. 20% D. 40%

24. ______________

______________ **/ 30 points**

25. (6 points) How many documents with a “True” label were in our test set?
A. 50 B. 100 C. 40 D. 70

25. ______________

______________ / **6 points**

Answer only two of the next three questions. If you answer all three, I will grade the first two. If you start to answer one of the questions, but a BIG, DARK cross through it so I know not to grade it. If it isn't obvious that I shouldn't grade a problem (by it being blank or not crossed out), you risk me grading it.

26. (50 points) You want to find a linear SVM that has the maximum margin on the following dataset on the 2D plane:[1]

TRUE	FALSE
$(-4, 5)$	$(4, -4)$
$(-3, 4)$	$(3, -5)$
$(-3, 5)$	$(3, -4)$

Report the margin as the sum of the distance between the separating hyperplane and closest "TRUE" point and the distance between the separating hyperplane and the closest "FALSE" point (I'm not trying to trick you; this is the same definition we used in class).

1. For the hyperplane $y = 0$, what is the margin?
2. For the hyperplane $x = 0$, what is the margin?
3. Find a hyperplane with the best (maximum) margin. (Hint: the correct hyperplane should have have a margin that's an integer.)

bf ANSWER:

1. 6
2. 8
3. $y = \frac{3}{4}x$; the margin is 10

[1] Points are represented in (x, y) format

27. (50 points) (Question borrowed from Charles Elkan at UCSD) UCSD has about 25,000 students. Every year, 1.4% students attempt suicide, and there are about two actual suicides. Students who have attempted suicide are 500 times more likely to commit suicide successfully in the next year.

    1. What percentage of all suicide attempts are successful?

       Now, suppose we have a database with detailed information about each UCSD student, with two labels for each person: whether or not s/he actually committed suicide, and whether or not s/he attempted suicide.

    2. Explain why it would or would *not* be useful to train a classifier to distinguish the actual suicides from all other students.

    3. Suppose you train a classifier that gets 95% accuracy. Is this good? What baseline performance should you compare it against?

    bf ANSWER:

    1. $25000\frac{1.4}{100} = 350$; $\frac{2}{350} = 0.6\%$
    2. No; with only two positive examples, there are not enough data to train a reliable classifier
    3. It's hard to say without seeing a confusion matrix. If it has high recall, this is good. However, a classifier that always said "no"; it would get over 99% accuracy.

______________ / **50 points**

28. (50 points) As part of an inheritance, you and your cousins have to split an inheritance of valuable, old dimes mixed in with some duds. The will is oddly specific about the rules of how you divide the dimes. After spending the night in a haunted house, you can take ten dimes to an appraiser to find out their value, each inheritor must submit a decision tree to decide which dimes are valuable and which are not. (Compute MLE probability estimates, do not use smoothing, and assume that all variables are binary.)

Face	Mint	During WWI	Silver	**Valuable**
Mercury	D	N	Y	Y
Mercury	D	Y	N	Y
Mercury	D	Y	Y	Y
Mercury	D	Y	Y	Y
Liberty	P	Y	Y	Y
Mercury	P	Y	Y	N
Liberty	D	N	Y	N
Liberty	D	Y	N	N

You want good dimes, so let's build a decision tree.

1. What is the entropy of $P(\text{Valuable})$?
2. What is the conditional entropy of $P(\text{Valuable}|\text{Face} = M)$?
3. What is the conditional entropy of $P(\text{Valuable}|\text{Face} = L)$?
4. What is the conditional entropy of $P(\text{Valuable}|\text{WWI} = Y)$?
5. What is the conditional entropy of $P(\text{Valuable}|\text{WWI} = N)$?
6. Which would make a more informative rule for the decision tree, asking the "Face" of a coin or whether it was made during "WWI"? Be sure to justify your answer using information theory.

bf ANSWER:

1. $-(5/8)lg(5/8) - (3/8)lg(3/8) = 0.95$
2. $-(4/5)lg(4/5) - (1/5)lg(1/5) = 0.72$
3. $-(4/6)lg(4/6) - (2/6)lg(2/6) = 0.92$
4. $-(4/6)lg(4/6) - (2/6)lg(2/6) = 0.92$
5. $-(1/2)lg(1/2) - (1/2)lg(1/2) = 1.0$
6. Face has a higher information gain

______________ **/ 50 points**

1 Entropy Table (Base 2)

$$H(x) \equiv -x\lg(x) \tag{4}$$

$H(0.010)$	$= 0.066$	$H(0.020)$	$= 0.113$	$H(0.030)$	$= 0.152$	$H(0.040)$	$= 0.186$
$H(0.050)$	$= 0.216$	$H(0.060)$	$= 0.244$	$H(0.070)$	$= 0.269$	$H(0.080)$	$= 0.292$
$H(0.090)$	$= 0.313$	$H(0.100)$	$= 0.332$	$H(0.110)$	$= 0.350$	$H(0.120)$	$= 0.367$
$H(0.130)$	$= 0.383$	$H(0.140)$	$= 0.397$	$H(0.150)$	$= 0.411$	$H(0.160)$	$= 0.423$
$H(0.170)$	$= 0.435$	$H(0.180)$	$= 0.445$	$H(0.190)$	$= 0.455$	$H(0.200)$	$= 0.464$
$H(0.210)$	$= 0.473$	$H(0.220)$	$= 0.481$	$H(0.230)$	$= 0.488$	$H(0.240)$	$= 0.494$
$H(0.250)$	$= 0.500$	$H(0.260)$	$= 0.505$	$H(0.270)$	$= 0.510$	$H(0.280)$	$= 0.514$
$H(0.290)$	$= 0.518$	$H(0.300)$	$= 0.521$	$H(0.310)$	$= 0.524$	$H(0.320)$	$= 0.526$
$H(0.330)$	$= 0.528$	$H(0.340)$	$= 0.529$	$H(0.350)$	$= 0.530$	$H(0.360)$	$= 0.531$
$H(0.370)$	$= 0.531$	$H(0.380)$	$= 0.530$	$H(0.390)$	$= 0.530$	$H(0.400)$	$= 0.529$
$H(0.410)$	$= 0.527$	$H(0.420)$	$= 0.526$	$H(0.430)$	$= 0.524$	$H(0.440)$	$= 0.521$
$H(0.450)$	$= 0.518$	$H(0.460)$	$= 0.515$	$H(0.470)$	$= 0.512$	$H(0.480)$	$= 0.508$
$H(0.490)$	$= 0.504$	$H(0.500)$	$= 0.500$	$H(0.510)$	$= 0.495$	$H(0.520)$	$= 0.491$
$H(0.530)$	$= 0.485$	$H(0.540)$	$= 0.480$	$H(0.550)$	$= 0.474$	$H(0.560)$	$= 0.468$
$H(0.570)$	$= 0.462$	$H(0.580)$	$= 0.456$	$H(0.590)$	$= 0.449$	$H(0.600)$	$= 0.442$
$H(0.610)$	$= 0.435$	$H(0.620)$	$= 0.428$	$H(0.630)$	$= 0.420$	$H(0.640)$	$= 0.412$
$H(0.650)$	$= 0.404$	$H(0.660)$	$= 0.396$	$H(0.670)$	$= 0.387$	$H(0.680)$	$= 0.378$
$H(0.690)$	$= 0.369$	$H(0.700)$	$= 0.360$	$H(0.710)$	$= 0.351$	$H(0.720)$	$= 0.341$
$H(0.730)$	$= 0.331$	$H(0.740)$	$= 0.321$	$H(0.750)$	$= 0.311$	$H(0.760)$	$= 0.301$
$H(0.770)$	$= 0.290$	$H(0.780)$	$= 0.280$	$H(0.790)$	$= 0.269$	$H(0.800)$	$= 0.258$
$H(0.810)$	$= 0.246$	$H(0.820)$	$= 0.235$	$H(0.830)$	$= 0.223$	$H(0.840)$	$= 0.211$
$H(0.850)$	$= 0.199$	$H(0.860)$	$= 0.187$	$H(0.870)$	$= 0.175$	$H(0.880)$	$= 0.162$
$H(0.890)$	$= 0.150$	$H(0.900)$	$= 0.137$	$H(0.910)$	$= 0.124$	$H(0.920)$	$= 0.111$
$H(0.930)$	$= 0.097$	$H(0.940)$	$= 0.084$	$H(0.950)$	$= 0.070$	$H(0.960)$	$= 0.057$
$H(0.970)$	$= 0.043$	$H(0.980)$	$= 0.029$	$H(0.990)$	$= 0.014$	$H(1.000)$	$= 0.000$

______________ **/ 0 points**