Homework 1 Solutions
Probability, Naive Bayes, and Logistic Regression

CMU 10-601: Machine Learning (Fall 2015)
http://www.cs.cmu.edu/~10601b/
OUT: Sep 10, 2015
DUE: Sep 22, 2015, 10:20 AM

1 Linear Algebra and Probability Review [15 points]

1.1 Linear Algebra review [7 pts]

Please choose T/F for questions 1-12 . Recall that if a set of vectors spans $\mathbb{R}^n$, than any n-dimensional vector can be written as a weighted sum of that set of vectors. Also, recall that the null space of a matrix A is $\{\mathbf{x} : A\mathbf{x} = \mathbf{0}\}$.

1. The matrix $\begin{bmatrix} I & A \\ 0 & I \end{bmatrix}$ is always invertible for any A. **Solution:** True, it is $\begin{bmatrix} I & -A \\ 0I & \end{bmatrix}$

   For questions 2-7, select True if the statement holds for any invertible $n \times n$ matrix A.

2. The columns of A span R^n. **Solution:** True
3. $A\mathbf{x} = \mathbf{0}$ has only the trivial solution $\mathbf{x} = \mathbf{0}$. **Solution:** True
4. A^T is invertible. **Solution:** True
5. The rank of A is n. **Solution:** True
6. $\det(A) = 0$. **Solution:** False
7. 0 is an eigenvalue of A. **Solution:** False
8. Let $A = \begin{bmatrix} 1 & -1 & 5 \\ 2 & 0 & 7 \\ -3 & -5 & -3 \end{bmatrix}$ and $u = \begin{bmatrix} -7 \\ 3 \\ 2 \end{bmatrix}$. True or False: $\mathbf{u}$ is in the null space of A. **Solution:** True
9. $Q = \left\{ \begin{bmatrix} -1 \\ 2 \\ -3 \end{bmatrix}, \begin{bmatrix} 2 \\ -7 \\ 9 \end{bmatrix} \right\}$ spans $\mathbb{R}^3$. **Solution:** False
10. Q spans R^2. **Solution:** False
11. Vector $\begin{bmatrix} 6 \\ -5 \end{bmatrix}$ is an eigenvector of matrix $\begin{bmatrix} 1 & 6 \\ 5 & 2 \end{bmatrix}$. **Solution:** True
12. Vector $\begin{bmatrix} 3 \\ -2 \end{bmatrix}$ is an eigenvector of matrix $\begin{bmatrix} 1 & 6 \\ 5 & 2 \end{bmatrix}$. **Solution:** False For questions 13 and 14, write your answer to 3 decimal place accuracy.
13. If $\mathbf{u}^T\mathbf{v} = 0$, and $\|\mathbf{u}\| = 2$ and $\|\mathbf{v}\| = 3$, what is $\|\mathbf{u} + \mathbf{v}\|$? (Hint: think about what $\mathbf{u}^T\mathbf{v} = 0$ means geometrically.) **Solution:** $\sqrt{13} \approx 3.606$
14. Suppose $n \times n$ positive definite diagonal matrix $A = \begin{bmatrix} d^2 & 0 & \cdots & 0 \\ 0 & d^2 & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & d^2 \end{bmatrix}$, and $\text{tr}(A) = 9n$. What is d? **Solution:** 3 (or -3)

1.2 Marginal, Joint, & Conditional Probabilities [3 pts]

Consider the following joint probability table:

A	B	$P(A,B)$
0	0	0.2
0	1	0.3
1	0	0.1
1	1	0.4

1. What is $P(A = 0, B = 0)$?

   Solution: 0.2

2. What is $P(A = 0 \mid B = 1)$?

   Solution: $0.3/0.7 = 0.429$

3. What is $P(A = 1 \vee B = 1)$?

   Solution: $0.5 + 0.7 - 0.4 = 0.8$

1.3 Short answer questions [5 pts]

1. A bag contains 2 unbiased coins. One coin has heads and tails on opposing sides, while the other has heads on both the sides. You put your hand in the bag and take out a coin. Without looking at the coin, you flip it. When the coin lands on the floor, you observe that the side facing you reads heads. What is the probability that the opposite side (the one you cannot see) is also heads? (Hint: Consider A to be the event that the HH coin is picked, and B to be event that the side facing you is heads. What is P($A \mid B$) ?)

   Solution: $\frac{2}{3}$

2. Two friends take turns to hit the bull's-eye while playing a game of darts. The probability of the first friend succeeding in a particular turn is 1/3, while the probability of the second friend succeeding is 1/4. The game is limited to 5 turns. Both the friends keep playing the game till one of them hits the target and wins the game, or till they exhaust all five turns. What is the probability of the first friend winning the game?

   Solution: $\frac{1}{3} + \frac{1}{6} + \frac{1}{12} + \frac{1}{24} + \frac{1}{48} \approx 0.64583$. The first player could win in the first turn with probability $\frac{1}{3}$. To win in the second turn, both player need to fail in their first turn and the probability is $\frac{2}{3} * \frac{3}{4} * \frac{1}{3}$. The same generalization can be made for the first player to win on an arbitrary turn. The sum of these probabilities forms a geometric progression and is $\frac{2}{3}$.

3. Let X be a normally distributed random variable with zero mean and standard deviation of 1. What is P(X = 0.6)?

   Solution: 0. The probability at any point on the pd curve is zero.

4. You have a game with a fair die. The only way to win the game is to roll a 6 on the die. You get 3 attempts. The game stops if you get a 6, or if you have exhausted your attempts. What is the probability of you winning the game?

   Solution: $\frac{91}{216}$.

2 MLE and MAP estimation [3 points]

Your friend gives you a coin and asks you to estimate the probability of the coin showing heads when flipped. You assume, a-priori, that the most probable value for heads is 0.5. To estimate the probability, you then flip the coin 3 times, and it comes up heads twice. Which will be higher, your maximum likelihood estimate (MLE), or your maximum a posteriori probability? Please write ”MLE” or ”MAP” as the answer.

Solution: MLE

Table 1: P(Y)

	$P(Y)$
Y = 0	0.7
Y = 1	0.3

Table 2: P($X_1|Y$)

	$X_1 = 0$	$X_1 = 1$
Y = 0	0.6	0.4
Y = 1	0.4	0.6

3 Naive Bayes and Logistic Regression [22 points]

3.1 [2 points] True or False

For questions 1-4, write either T or F.

1. The boundary of a Gaussian naive Bayes classifier is always linear. **Solution:** False
2. Provided enough training data, Naive Bayes will achieve zero classification error over training examples. **Solution:** False
3. A Naive Bayes classifier assumes that training features are independent of each other. **Solution:** False, only independent after conditioning on class.
4. Maximizing the likelihood of the logistic regression model leads to multiple local optima. **Solution:** False

3.2 [20 points] Short Answer

1. Consider three random variables X_1, X_2 and Y. X_1 and Y are generated according to Tables 1 and 2. Once X_1 has been generated, X_2 takes on the same value as X_1. Consider a Naive Bayes classifier trained on features X_1 and X_2 and class label Y, using a very large number of training samples generated from Tables 1 and 2. What is the prediction of the classifier for $X_1 = 0$, $X_2 = 0$? **Solution:** Y=0
2. What is the probability of the class from 3.2.1? **Solution:** $\frac{0.7*0.6*0.6}{0.7*0.6*0.6+0.3*0.4*0.4} = 0.840$
3. What is the prediction for $X_1 = 0$, $X_2 = 1$? **Solution:** $Y = 0$
4. What is the probability of the class from 3.2.3? **Solution:** $\frac{0.7*0.6*0.4}{0.7*0.6*0.4+0.3*0.4*0.6} = 0.700$
5. What is the prediction for $X_1 = 1$, $X_2 = 0$? **Solution:** $Y = 0$
6. What is the probability of the class from 3.2.5? **Solution:** $\frac{0.7*0.4*0.6}{0.7*0.4*0.6+0.3*0.6*0.4} = 0.700$
7. What is the prediction for $X_1 = 1$, $X_2 = 1$? **Solution:** Y=0
8. What is the probability of the class from 3.2.7? **Solution:** $\frac{0.7*0.4*0.4}{0.7*0.4*0.4+0.3*0.6*0.6} \approx 0.509$
9. [4 pts] Consider the classifier trained in the previous question. What is the expected error rate on any test examples generated using Tables 1 and 2? **Solution:** $P(Y = 1, X_1 = 0, X_2 = 0) + P(Y = 1, X_1 = 1, X_2 = 1) = 0.3 * 0.4 + 0.3 * 0.6 = 0.300$

10. [4 pts] Consider the same scenario as the previous two questions, but without the duplicate feature X_2. What is the expected error rate of the classifier in this case? **Solution:** The classifier makes the same predictions, so 0.300

11. [2 pts] Consider a Gaussian Naive Bayes classifier. The number of training features is 20 and the number of possible class labels is 2. Once done with training, how many independent parameters do you need to store for the classifier? Assume we estimate the variance separately over each of the classes and features.

    Solution: 81. This is because each feature, conditioned on class, needs a mean and a std dev parameter. And there are 2 classes. Thus, each feature contributes 2x2 = 4 parameters. For 20 features, you need 20x4 = 80 parameters. You also need one parameter to store the class density.

12. [2 pts] Repeat the above analysis for a Gaussian Bayes Classifier. How many parameters do you need to specify the classifier now? (Hint: In a Gaussian Bayes Classifier, features are NOT conditionally independent of each other given the class label.)

    Solution: 461. 20 parameters for the mean plus 210 parameters to store the symmetric 20x20 covariance matrix, multiplied by 2 for the two classes. And add 1 parameter to store class density

4 Evaluation.txt

Each worth 0.8 points:

- 0.83486
- 0.95122
- 0.97500
- 0.87500
- 0.77778