CS 443 – Introduction to Data Science
Spring 2015 – Homework 3 – 100 points
Due 11:59pm (Eastern) April 13, 2015

This homework consists of problems covering logistic regression, linear regression, ensemble methods, SVM, model evaluation, hypothesis testing, and clustering.

A few instructions to make life easier for all of us:

- Turn in your answers in a single file, called <Last-Name><First-Name>HW3.pdf, by uploading it into Sakai.
- Please write concisely and clearly. There are points for intermediate steps, but not in "talking problems to death."
- Have the following at the top of every page: <Last-name>, <First-name> [<page-number>/<number-of-pages>] Example: Smith, John [1/10]
- Your solution should have a cover-page that provides the following information:

<First-name> <Last-name>

Problem Number	Solution Page
1	
2	
3a	
3b	
…	

  If you did not answer a problem, then enter "not done" instead of the solution page.
- It is highly recommended that you turn-in a typed copy of your homework (as opposed to scanned hand-written copy). This is especially true for equations and plots.
- As always, our TA is here to help you with any doubt or confusion that you may have.

Q1. [6 points] Suppose you train a logistic classifier $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$. Suppose $\theta_0 = -6$, $\theta_1 = 1$, and $\theta_2 = 0$. Draw the decision boundary found by your classifier. (Hint: You have a two-dimensional input, so your drawing should be a two-dimensional plot.)

Q2. [6 points] Suppose you have $m = 14$ training examples with $n = 3$ features (excluding the additional all-ones feature for the intercept term, which you should add). The normal equation is $\theta = (X^T X)^{-1} X^T y$. For the given values of m and n, what are the dimensions of θ, X, and y in this equation?

Q3. [6 points] Suppose $m = 4$ students have taken some class, and the class had a midterm exam and a final exam. You have collected a dataset of their scores on the two exams, which is as follows:

student ID	midterm exam	(midterm exam)^2	final exam
1	89	7921	96
2	72	5184	74
3	94	8836	87
4	69	4761	78

You would like to use polynomial regression to predict a student's final exam score from their midterm exam score. Concretely, suppose you want to fit a model of the form $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$, where x_1 is the midterm score and x_2 is (midterm exam)^2.

3a. [2 points] Is feature scaling necessary here? If so, provide one strategy for scaling the features.

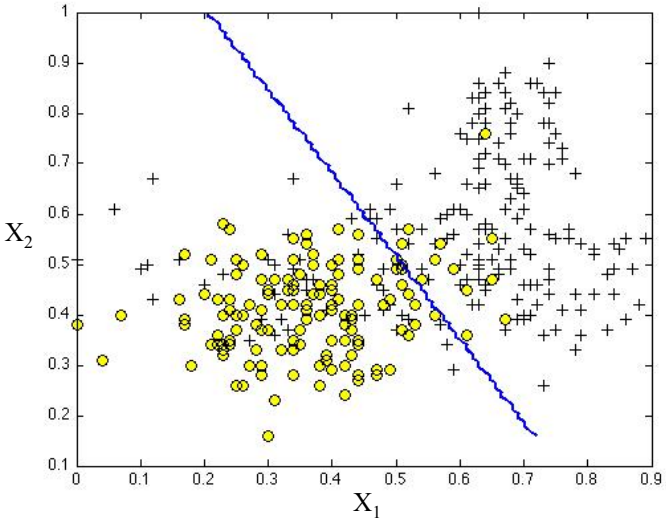
3b. [4 points] You run gradient descent for 15 iterations with the learning rate $\eta = 0.3$ and compute the cost function $J(\theta)$ after each iteration. You find that the value of $J(\theta)$ increases over time. How would you fix this problem so $J(\theta)$ decreases over time?

Q4. [5 points] Suppose the classification performance of your Adaboost on a test set is poor.

4a. [2 points] What is the most likely explanation for this poor performance?

4b. [3 points] Name a solution to fix this poor performance.

Q5. [9 points] Suppose you have trained an SVM classifier with a Gaussian kernel, and it learned the following decision boundary on the training set:



When you measure the SVM's performance on a cross-validation set, it does poorly.

5a. [3 point] What is the problem here in terms of bias and variance?

5b. [3 points] Should you try increasing or decreasing C?

5c. [3 points] Should you try increasing or decreasing σ^2?

Q6. [3 points] Bagging methods (such as random forest) can be used to increase overall accuracy by learning and combining a collection of individual classifiers. Name one major shortcoming of bagging.

Q7. [4 points] Why is naïve Bayesian classification called "naïve"?

Q8. [4 points] You are working on a spam classification system using naïve Bayes. "Spam" is the positive class ($Y = 1$) and "not spam" is the negative class ($Y = 0$). You have trained your classifier, and there are $n = 1000$ examples in the cross-validation set. The chart of predicted class vs. actual class is as follows: (Recall from lecture that this table is often called the *Confusion Matrix*.)

		Actual Class	
		1	**0**
Predicted Class	**1**	85	890
	0	15	10

What are the classifier's accuracy, precision, recall, and F1-score? Give your answers to three decimal places.

Q9. [6 points] Suppose you are working on a fraud classifier, where fraud cases are positive examples ($Y = 1$) and non-fraud cases are negative examples ($Y = 0$). You have a training set of cases in which 99% of the examples are non-fraud and the other 1% is fraud.

9a. [3 points] What is the classifier's accuracy if it always predicts non-fraud? Explain briefly why this is not a good fraud detection system.

9b. [3 points] What are the classifier's recall and precision if it always predicts fraud?

Q10. [10 points] Consider the following approach for testing whether a classifier *A* beats another classifier *B*. Let *N* be the size of a given data set, p_A be the accuracy of classifier *A*, p_B be the accuracy of classifier *B*, and $p = \frac{p_A + p_B}{2}$ be the average accuracy for both classifiers. To test whether classifier *A* is significantly better than *B*, the following *Z*-statistic is used:

$$Z = \frac{p_A - p_B}{\sqrt{\frac{2p(1-p)}{N}}}$$

Classifier *A* is assumed to be better than classifier *B* if $Z > 1.96$.

The following table lists the accuracies of decision tree and naïve Bayes on various data sets.

Data Set	Size (*N*)	Decision Tree Accuracy (%)	Naïve Bayes Accuracy (%)
Anneal	898	92.09	79.62
Australia	690	85.51	76.81
Auto	205	81.95	58.05
Breast	699	95.14	95.99
Cleve	303	76.24	83.5
Credit	690	85.8	77.54
Diabetes	768	72.4	75.91
German	1000	70.9	74.7
Glass	214	67.29	48.59
Heart	270	80	84.07
Hepatitis	155	81.94	83.23
Horse	368	85.33	78.8
Ionosphere	351	89.17	82.34
Iris	150	94.67	95.33
Labor	57	78.95	94.74
Led7	3200	73.34	73.16
Lymphography	148	77.03	83.11
Pima	768	74.35	76.04
Sonar	208	78.85	69.71
Tic-tac-toe	958	83.72	70.04
Vehicle	846	71.04	45.04
Wine	178	94.38	96.63

Identify the data set in the above table where decision tree's accuracy is greater than naïve Bayes' accuracy, but this difference is not significant according to the *Z*-statistic. Report the *Z*-statistic. Give your answer for the Z-statistic to two decimal places.

Q11. [7 points] Given a database of information about your users, you want to automatically group them into *k* different market segments. You are using SSE (short for Sum of Squared Error) to measure the quality of your clustering. Recall that SSE is the sum of the squares of the distances between each of the points of the cluster and the centroid.

11a. [4 points] Briefly describe how you would pick the appropriate *k*.

11b. [3 points] Why do you expect your method to give the appropriate *k*?

Q12. [4 points] Suppose you have an unlabeled dataset $\{x^{(1)}, \ldots, x^{(m)}\}$. You run K-means with 50 different random initializations, and obtain 50 different clusterings of the data. What is the recommended way for choosing which one of these 50 clusterings to use?

Q13. [12 points] The SSE (short for Sum of Squared Error) is a common measure of the quality of a cluster. It is the sum of the squares of the distances between each of the points of the cluster and the centroid. Suppose a cluster, C_1, consists of the following three points: (3, 2), (0, 8), and (6, 5).

13a. [3 points] Compute the SSE for cluster C_1.

13b. [6 points] Sometimes, we decide to split a cluster in order to reduce the SSE. Find the optimal partitioning (in terms of reduction in SSE) that produces two clusters from C_1. List the two clusters.

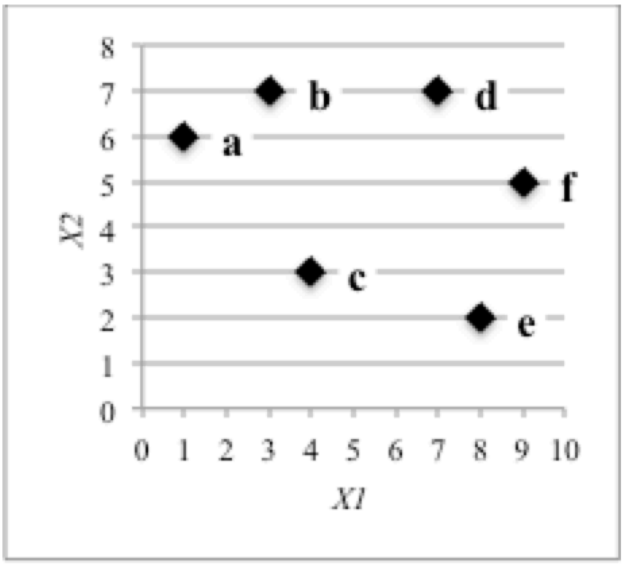
13c. [3 points] What is the reduction in the SSE if use the above optimal partitioning?

Q14. [6 points] Suppose we want to assign points to one of two cluster centroids, either (0, 0) or (100, 40). Depending on whether we use the L1 or L2 norm, a point (x, y) could be clustered with a different one of these two centroids. Which one of these points will be clustered with the centroid (0, 0) when the L1 norm is used, but clustered with the centroid (100, 40) when the L2 norm is used?

Point 1: (52, 13)
Point 2: (51, 15)
Point 3: (61, 8)
Point 4: (63, 8)

Q15. [12 points] Suppose the *clustroid* of a cluster is taken to be the point in the set that has the minimum sum of the squares of the distances to all other points in the cluster. We are given the following points in a two-dimensional Euclidean space:

a = (1, 6)
b = (3, 7)
c = (4, 3)
d = (7, 7)
e = (8, 2)
f = (9, 5)



Assuming the usual L2 norm as our distance measure, compute the clustroids of the following sets: {a, b, c, d}, {b, c, d, e}, {c, d, e, f}, and {a, b, e, f}.