CS5785 Applied Machine Learning - Prelim

October 6, 2016 1:00 PM - 2:15 PM (75 minutes)

Name: _____ NetID: _____

- This exam is closed book. But you are allowed to use one cheat sheet (Letter size, two-sided).
- There should be in total 10 numbered pages in this exam (including this cover sheet). The last 2 pages are used for scratch paper.
- There are 6 questions worth a total of 100 points. Work efficiently. Carefully manage your time to focus on the easier questions first, and avoid getting stuck in the more difficult ones before you have answered the easier ones.
- You have 75 minutes. Good luck!

Question	**Topic**	**Max. Score**	**Score**
1	Short Questions	20	
2	Training and Validation	8	
3	Probability and MLE	20	
4	SVD and LDA	12	
5	ROC Curves and Score Distribution	25	
6	Decision Boundary	15	
	Total	100	

1 SHORT QUESTIONS [20 POINTS]:

(a) (**2 pts.**) Logistic Regression is an example of *supervised* learning method. **True** or **False**?

(b) (**2 pts.**) The ROC Curve plots the _____ rate against the _____ rate.

(c) (**2 pts.**) If X and Y are *independent random variables*, then $E[X+2Y] = E[X] + 2E[Y]$ and $Var[X+2Y] = Var[X] + 2Var[Y]$. **True** or **False**?

(d) (**2 pts.**) Given a rectangular matrix X, the matrix $(X^\top X)^{-1}X^\top$ is called the _____.

(e) (**2 pts.**) Stochastic Gradient Descent is an example of an *online learning method.* **True** or **False**?

(f) (**2 pts.**) Principal Components Analysis (PCA) and Fisher Linear Discriminant Analysis are two examples of unsupervised methods for dimensionality reduction. **True** or **False**?

Experimental design: for each of the listed descriptions below, choose whether the experimental set up is ok or problematic. If you think it is problematic, briefly state where the problems are:

(g) (**2 pts.**) A project team reports a low prediction error on their training set and claims their method is good. **Ok** or **Problematic**?

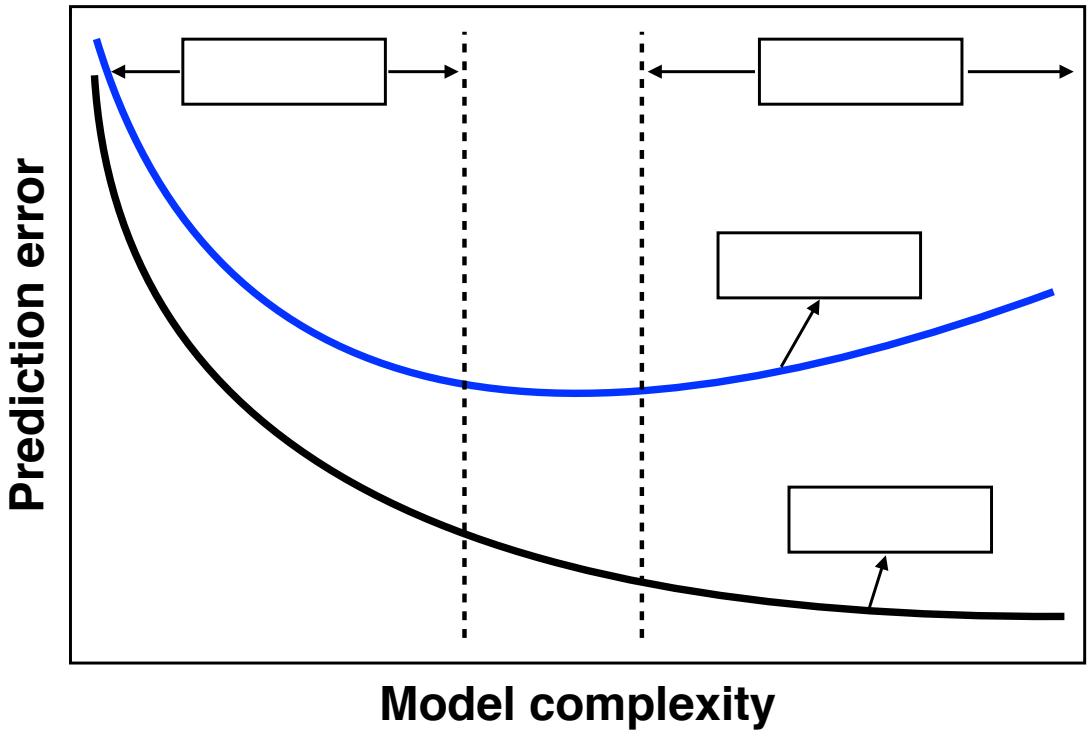
(h) (**2 pts.**) A project team claimed great success after achieving 98% classification accuracy on a spam email classification task where their data consisted of 50 positive examples and 5,000 negative examples. **Ok** or **Problematic**?

(i) (**2 pts.**) A project team split their data into training and test sets. Using cross-validation on the training set, they chose the best parameter setting. They built a model using these parameters on the entire training set, and then report their error on test set. **Ok** or **Problematic**?

(j) (**2 pts.**) A project team did parameter selection on the full data set. Then they split the data into training and test sets. They built their model on the training set using several parameter model settings, and report the the best test error they achieved. **Ok** or **Problematic**?

2 TRAINING AND VALIDATION [8 POINTS]:

The following figure depicts training and validation error curves of a machine learning algorithm with increasing model complexity (e.g., kNN classifier with decreasing k):



(a) (**4 pts.**) Which of the curves is more likely to be the training error and which is more likely to be the validation error? Indicate on the graph by filling the boxes with “training error" and “validation error".

(b) (**4 pts.**) In which regions does the model overfit or underfit? Indicate on the graph by filling the boxes with “overfit" and “underfit".

3 PROBABILITY AND MLE [20 POINTS]:

You are playing a game with two coins. Coin 1 has a θ probability of heads and coin 2 has a 2θ probability of heads. You flip these coins 5 times and observe the following results:

Coin	1	2	2	2	2
Result	Head	Tail	Tail	Head	Tail

(a) (**10 pts.**) What is the log-likelihood of your result data given $\theta = 0.25$? Use log with base 2.

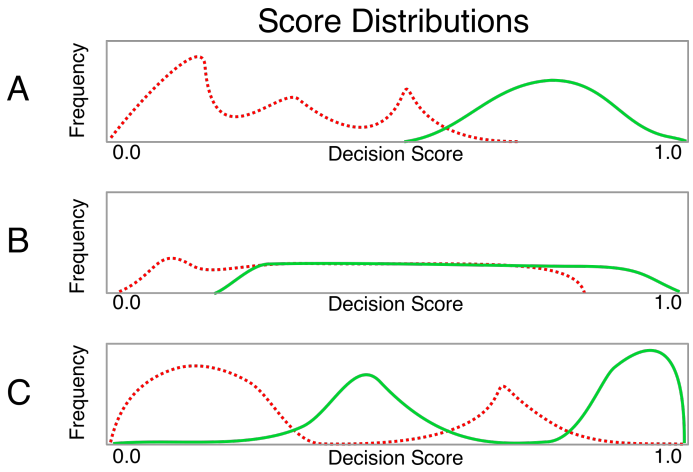
(b) (**10 pts.**) What is the maximum likelihood estimate for θ ?

4 SVD AND LDA [12 POINTS]:

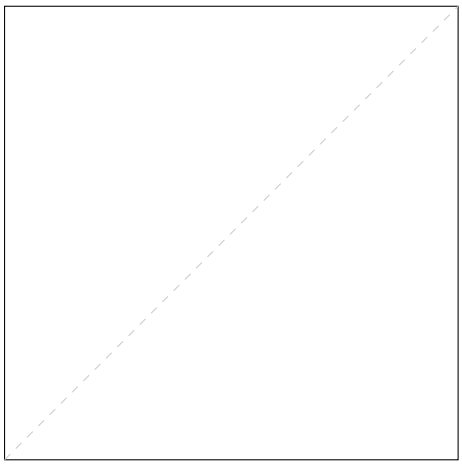
(a) (**6 pts.**) Compare and contrast Principal Components Analysis (PCA) and Fisher Linear Discriminant Analysis.

(b) (**6 pts.**) In homework 2 we have seen that the SVD can be used to provide a rank k approximation of a matrix X (a set of face images in homework 2). Show how to do rank k approximation using SVD step-by-step.

5 ROC Curves and Score Distributions [25 points]:



Draw ROC curves here



(a) (**9 pts.**) Your engineers are working on a classifier that can reliably detect whether a food contains chocolate, but something isn't quite right. Shown above are the *decision score distributions* for three potential classifiers. Decision scores for foods that contain chocolate are shown in solid green and scores for foods without chocolate are shown in dashed red. **Please draw an ROC curve** for each classifier in the space provided. Be sure to **label both axes** of the ROC curve, and **mark the three curves with A, B, C** so we know which is which.

(b) (**6 pts.**) What is the approximate equal error rate for each classifier? Recall that the equal error rate is the error at the point where the false negative rate equals the false positive rate.

Line A: EER ≈ _____.

Line B: EER ≈ _____.

Line C: EER ≈ _____.

(c) (**5 pts.**) You are building a fingerprint lock that protects a vault full of classified documents. Only the owner's fingerprint should match. If your classifier accepts a fingerprint belonging to anyone else, the documents will get stolen. Should you optimize for the highest/lowest *accuracy*, *true positive rate*, or *false positive rate*? Why?

(d) (**5 pts.**) You work for the Food and Drug Administration. You are building a binary classifier that detects whether food sold in a store contains salmonella. If your system doesn't catch infected food, somebody will get sick. Should you optimize for the highest/lowest *accuracy*, *precision*, or *recall*? Why?

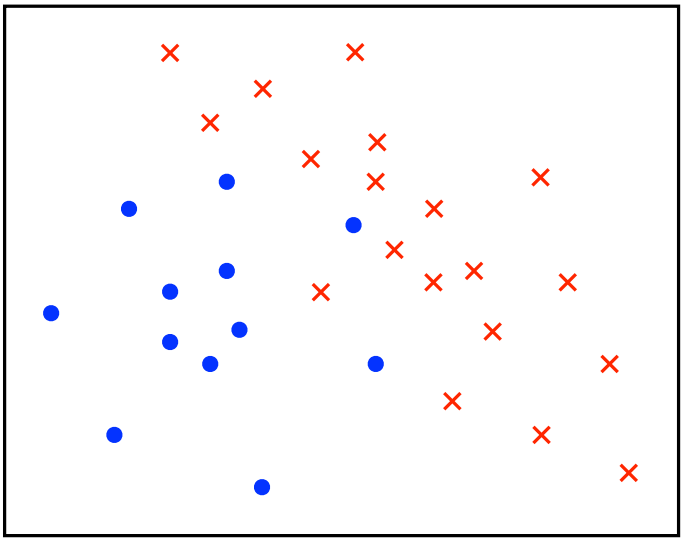
6 Decision Boundary [15 points]:

Consider learning a target function of the form $f : \mathbb{R}^2 \rightarrow \{-1, +1\}$. That is, a function to classify 2-dimensional data into 2 categories. Consider the following learning algorithms on the given training set (marked by blue dots and red cross):

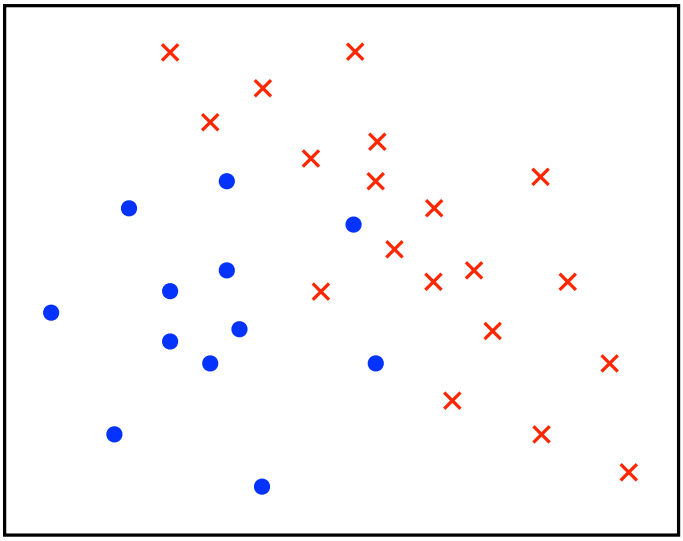
- 1-Nearest Neighbor Classifier (1-NN)
- Logistic Regression
- Gaussian Naive Bayes Classifier

(a) (**3 pts.**) Among these 3 algorithms, which one can be guaranteed to give 0 error rate on any arbitrary training set?

(b) (**4 pts.**) Draw the approximate decision boundary for **1-Nearest Neighbor Classifier (1-NN)** learned on the below training set.



(c) (**4 pts.**) Draw the approximate decision boundary for **Logistic Regression** learned on the below training set.



(d) (**4 pts.**) Draw the approximate decision boundary for **Gaussian Naive Bayes Classifier** learned on the below training set.