CS5785 Applied Machine Learning - Practice Prelim

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 12 numbered pages in this exam (including this cover sheet). The last 3 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	22	
2	Training and Validation	8	
3	Regularized Regression	15	
4	ROC Curves and Score Distribution	18	
5	Bayes Law	16	
6	Naive Bayes with Bag-of-Words	21	
	Total	100	

1 SHORT QUESTIONS [22 POINTS]:

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

(b) (**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**?

(c) (**2 pts.**) K-means will terminate after a finite number of steps, no matter the input. **True** or **False**?

(d) (**4 pts.**) A Kaggle competition typically has two different test sets. One set is used to score the public leaderboard during the competition. The other set is used to create a private leaderboard after the competition ends. One team repeatedly submits their method on the public set, making small changes each time until they achieve first place on the public leaderboard. When the competition ends, teams are re-scored on the private set. The team finds they have dropped to 102^{nd} place. **What happened?** Please describe.

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:

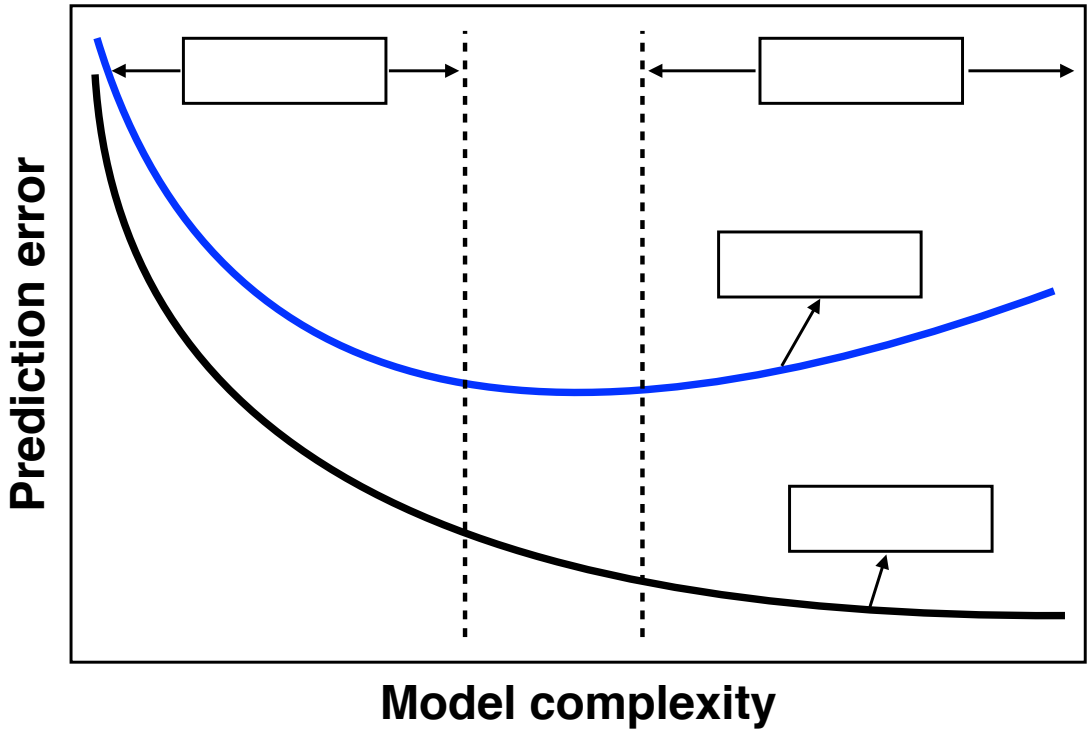
(e) (**4 pts.**) A project team noticed that when they added 5 features to their linear regression model, their training error went down. They chose this model because they said it was better. **Ok** or **Problematic**?

(f) (**4 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**?

(g) (**4 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):



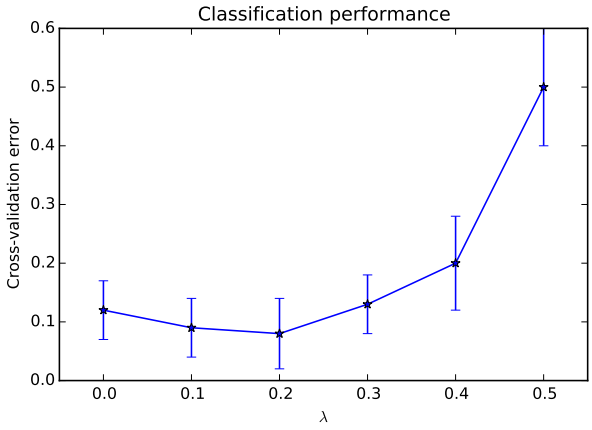
(a) (**4 pts.**) Which of the curves is more likely to be the training error and which is more likely to be the validation error? Indicate on the graph by filling the boxes with “training error" and “validation error".

(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 REGULARIZED REGRESSION (15PT)

Consider the following plot of the performance of a Lasso linear regression model. We plot cross-validation error with various choices of λ, the regularization coefficient. Recall that the Lasso problem can be formulated as

$$\hat{\beta}^{\text{Lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \; + \; \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$



(a) (**3 pts.**) A model with smaller coefficients and smaller support (number of nonzero coefficients) is considered simpler. With this in mind, which model is simpler: a **large** λ or a **small** λ? _

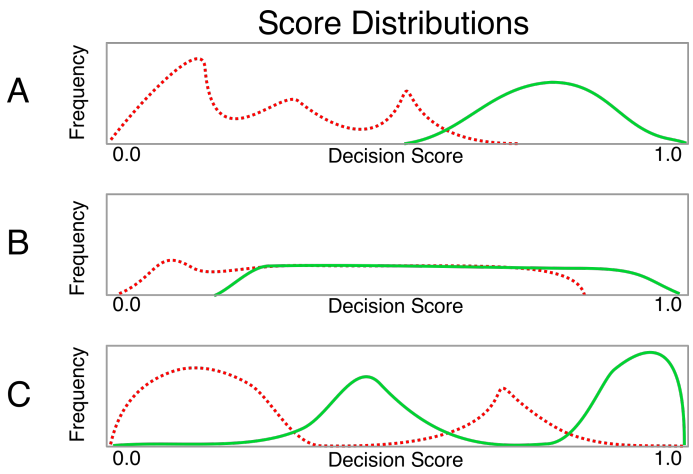
(b) (**3 pts.**) Which λ should you pick using the one-standard rule of thumb? _

(c) (**3 pts.**) Which λ value is equivalent to OLS? _

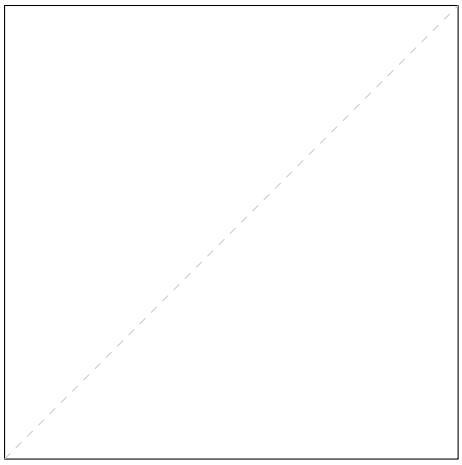
(d) (**3 pts.**) Using the λ from part (b), is $\sum_{i=1}^{p} |\beta_i^{\text{Lasso}}|$ **greater** than or **less** than $\sum_{i=1}^{p} |\beta_i^{\text{OLS}}|$?

(e) (**3 pts.**) Now assume this plot is showing training error. Judging from the plot, $\lambda \approx$_____ are likely to *underfit*, and $\lambda \approx$_____ are likely to *overfit*.

4 ROC Curves and Score Distributions [18 points]:



Draw ROC curves here



(a) (**6 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.**) The equal error rate is the error at the point where the false negative rate equals the false positive rate. What is the misclassification rate (1 − accuracy) for each classifier at the point where it obtains its EER?

Line A: misclassification rate ≈ _____.

Line B: misclassification rate ≈ _____.

Line C: misclassification rate ≈ _____.

(c) (**3 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) (**3 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?

5 BAYES'S LAW (16PT)

A Cornell Tech student invites you to be part of her enterprising startup, "Bananalr," providing consulting services to farmers.

That target clientelle are farmers growing premium free range artisinal bananas. Each banana sells at a profit of **$1**. However, **one in every thousand** bananas is infested with bugs and at present no solution is available to detect these bad bananas before selling them. If any customer ends up eating one of these, it will cost the farmer **$500** on average in bad publicity and lawsuits. Yuck!

Bananalr's value proposition is a machine learning model that decides whether a banana is infested or not before it is sold. **99%** of rotten bananas are detected as rotten, and **95%** of good bananas are detected as good. Bananalr runs on a Bananas-as-a-Service model and charges **$0.20** per banana per test.

(a) (**4 pts.**) Please draw a banana confusion matrix for Bananalr's classification model. (Suppose 1,000,000 bananas are tested.)

(b) (**4 pts.**) What is the probability that a banana marked bad by Bananalr's model is actually bad?

(c) (**4 pts.**) How much profit per banana can an artisinal banana farmer expect if they **don't** use the service?

(d) (**4 pts.**) How much profit per banana can an artisinal banana farmer expect if they **do** use the service?

Please show your work.

6 NAÏVE BAYES WITH BAG-OF-WORDS (21PT)

We are building a twitter robot that can tell the difference between tweets about sports versus academic tweets about machine learning conferences. Our students have scoured twitter, collecting a dataset as follows. Each tweet is labeled with either *sports* or *machine learning*.

Sports	*Machine learning*
Yankees **win** big last week, **score** 36-15 against Red Sox	**Deep Learning** models using rectified linear **loss** functions **score** better than hand-selected baselines in latest competition!
Learning their place on the new food chain, patriots suffer another **deep loss**	More **deep** model magic! Yann LeCunn for the **win**!
Giants **score** 42 points for the **win**, averting another **loss** in the semifinals	Reinforcement **learning** is the wave of the future

(a) (**6 pts.**) Convert the above tweets into a bag of words representation, a binary vector with five elements. Calculate P(word i present|class), for each of the important bolded words: **Win**, **Loss**, **Learning**, **Score**, **Deep**.

(b) (**6 pts.**) Consider the following tweet:

> Latest ResNet is crushing the ImageNet competition, with lower **loss** value than other models! Another big **win** for **deep learning**!

Let x be its bag-of-words vector. What probability $P(\text{class}_{\text{ML}}|x)$ and $P(\text{class}_{\text{sparts}}|x)$ would a Naive Bayes classifier assign to this tweet? What label would it assign?

(c) (**3 pts.**) Suppose now we move to a large dataset with millions of tweets, and we construct a matrix of bag of word representations X. Assume that the SVD decomposition of X is $X = UDV^T$. We would like to represent each word and each tweet in the dataset using a 300-dimensional vector for efficiency reasons. Describe how this could be done using the SVD decomposition of X.

(d) (**3 pts.**) It is possible to approximately reconstruct the original matrix X using the 300 dimensional vectors you described how to create in the previous part. Explain with a mathematical equation how to do this reconstruction.

(e) (**3 pts.**) The SVD decomposition solves the mathematical problem of providing encodings of the words and documents such that the original matrix X can be reconstructed with minimal error, for some error function. Write a mathematical equation for this reconstruction error.