# CS5785 Applied Machine Learning - Prelim

April 12, 2017 2:30 - 3:45 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 11 numbered pages in this exam (including this cover sheet). The last 2 pages are used for scratch paper.
- There are 7 questions worth a total of 100 points, with a bonus question worth 25 points extra credit. 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 | 30 | |
| 2 | Regularized regression | 12 | |
| 3 | Decision boundaries | 17 | |
| 4 | Performance curves | 16 | |
| 5 | Bayes's law | 12 | |
| 6 | Naïve Bayes with Bag-of-Words | 13 | |
| | **Total** | 100 | |
| Take-home bonus | Fairness and causality in ML | 25 | |

# 1 SHORT QUESTIONS (30PT)

(a) (2pt) The individual trees in a random forest are all trained on all of the training data. **True** or **False**?

(b) (2pt) Fitting a Gaussian mixture model using expectation maximization is an example of soft clustering. **True** or **False**?

(c) (2pt) If $X$ and $Y$ are *independent random variables*, then $E[X-2Y] = E[X]-2E[Y]$ and $Var[X-2Y] = Var[X]+4Var[Y]$. **True** or **False**?

(d) (2pt) K-Medoids is an example of an *unsupervised* learning method. **True** or **False**?

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

(f) (2pt) Suppose $X \in \mathscr{R}^{n\times d}$ is a dataset with $n = 1000$ images as $d = (16\times 16)$-dimensional row vectors. This means $X_{ij}$ is the brightness level between 0 and 255 of pixel $j$ in image $i$. Let $X = U\Sigma V^T$ be the singular value decomposition of $X$. Are all entries of $V$ nonnegative? **Yes** or **No**?

(g) (2pt) Each time we run K-means on the same dataset, we will get the same clustering no matter the starting conditions. **True** or **False**?

(h) (2pt) Each time we run hierarchical clustering using Ward's method on the same dataset, we will get the same clusters no matter the starting conditions. **True** or **False**?

(i) (2pt) 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 $102^{\text{nd}}$ place. **What happened?** Please describe.

**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 what the problems are:**

(j) (4pt) A project team reports a low prediction error on their training set and claims their method is good. **Ok** or **Problematic**?

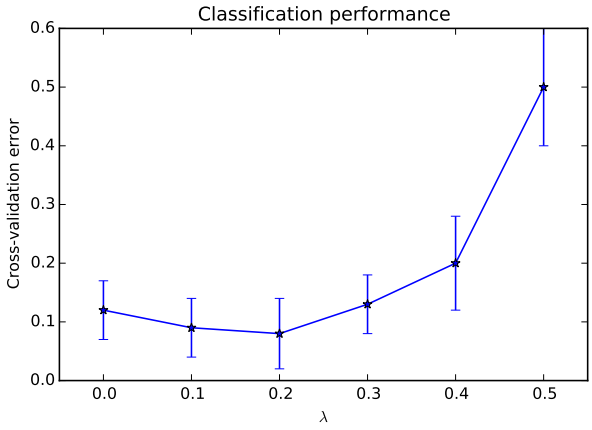
(k) (4pt) A project team noticed that when they added 5 features to their linear regression model, their training error went down. They chose this model because they said it was better. **Ok** or **Problematic**?

(l) (4pt) Suppose an intern at Google is trying to build a better spam classifier. They believe their model is good because every spam message in their test set was caught and correctly classified as spam. **Ok** or **Problematic**?

## 2 REGULARIZED REGRESSION (12PT)

Consider the following plot of the performance of a Lasso linear regression model. We plot cross-validation error with various choices of $\lambda$, 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) (2pt) 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** $\lambda$ or a **small** $\lambda$? _____

(b) (2pt) Which $\lambda$ should you pick using the one-standard rule of thumb? _____

(c) (2pt) Which $\lambda$ value is equivalent to OLS? _____

(d) (3pt) Using the $\lambda$ 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) (3pt) 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*.

## 3 Decision boundaries (17pt)

(a) (1.5pt each; 12pt total) Here are several two-class classifiers. Match the decision boundary with the classification method.

- Decision tree with depth 3

  _____

- Random forests with 10 trees of depth 3

  _____

- Kernel regression for classification with Gaussian kernel

  _____

- Forward stepwise linear regression, choosing exactly one feature

  _____

- $k$-NN, for $k = 1$

  _____

- $k$-NN, for $k = 3$

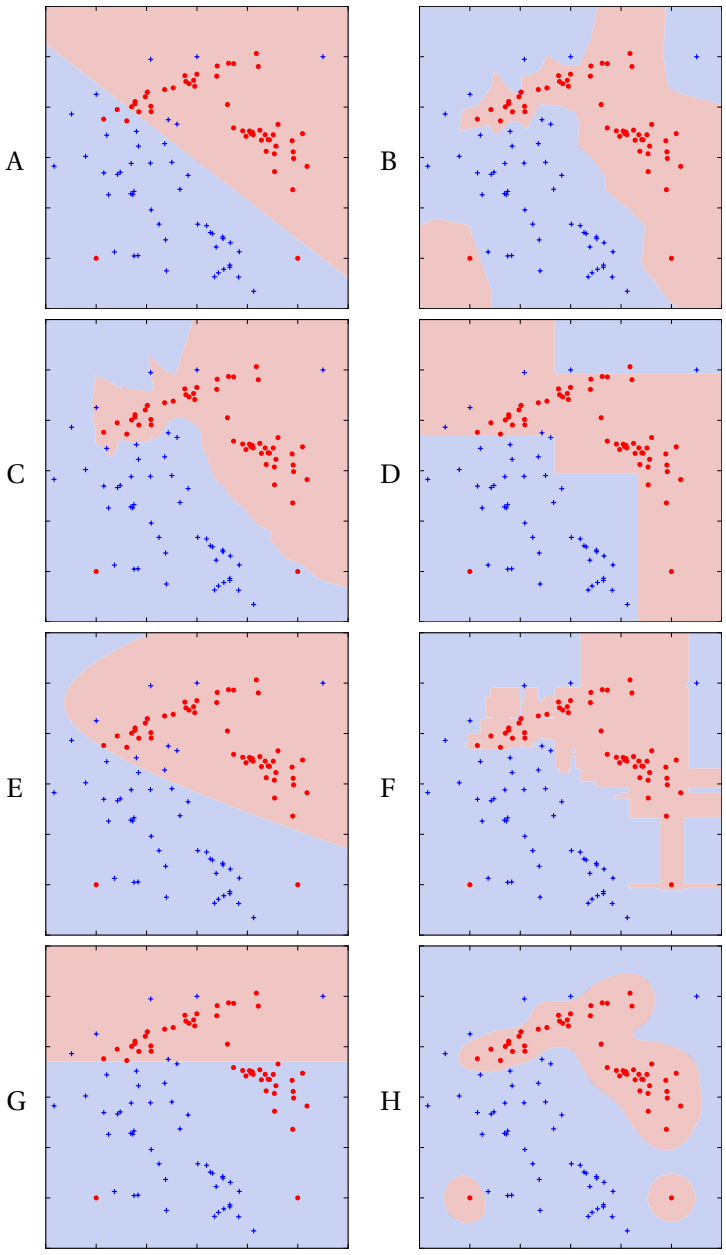
  _____

- Naïve Bayes using a univariate Gaussian to fit each of the conditional marginal distributions

  _____

- Logistic regression

  _____

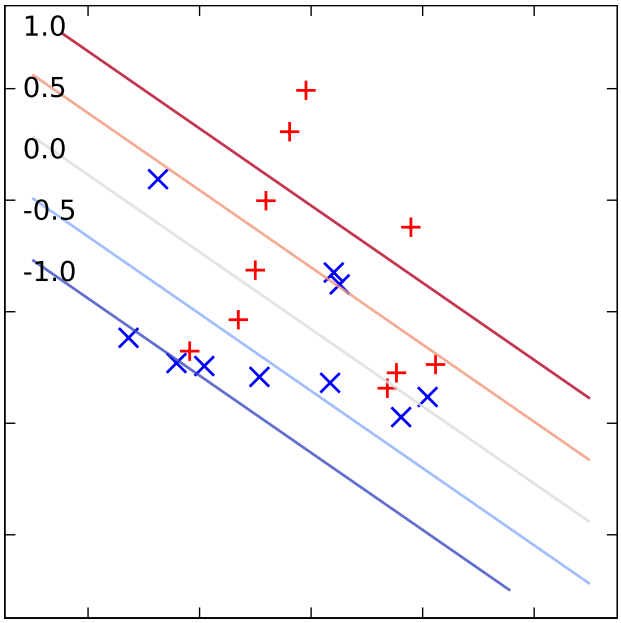


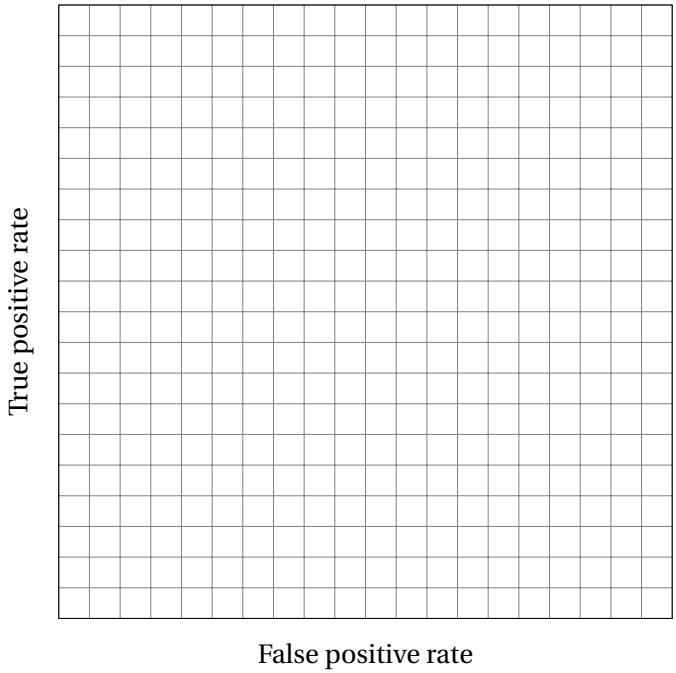

(b) (2pt) Which classifier(s) have the smallest training error?

(c) (3pt) Which classifier do you think would probably work best on a different test set? Please explain.

## 4 PERFORMANCE CURVES (16PT)

On the left is a two-class testing dataset with 20 points, ten of one class and ten of the other. We fit a Logistic Regression classifier on another training dataset and are plotting the decision boundaries at five different classification thresholds.




(a) (5pt) For each of the five thresholds, draw a point for the true positive rate and the false positive rate on the right graph. Please be exact and label each point with its TPR and FPR. (To help you, there are exactly 20 grid lines in the graph.)

(b) (2pt) Connect the dots. What is this graph called? _____ (Acronym is ok.)

(c) (3pt) Suppose this is a spam email classifier. The blue × are the spam class and the red + are the non-spam class. It's very bad to send a legitimate email to the spam folder. Which threshold would you pick? Please explain why.

(d) (3pt) Now suppose this is a model that detects cancer. The blue × represent patients with cancer and the red + are the patients that do not have cancer. Patients flagged by this model are sent through additional expensive tests. It's better to flag a healthy patient as cancerous than it is to miss a cancer case. Which threshold would you pick? Please explain why.

(e) (3pt) Someone from another team created a model with the following rates: at FPR=0.3, TPR=0.2; at FPR=0.5, TPR=0.6; at FPR=0.9, TPR=0.85. Which model is better: theirs or ours? Please explain.

# 5 BAYES'S LAW (12PT)

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) (3pt) Please draw a banana confusion matrix for Bananalr's classification model. (Suppose 1,000,000 bananas are tested.)

(b) (3pt) What is the probability that a banana marked bad by Bananalr's model is actually bad?

(c) (3pt) How much profit per banana can an artisinal banana farmer expect if they **don't** use the service?

(d) (3pt) 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 (13PT)

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* |
|---|---|
| Red Sox **win** big last week, **score** 36-35 against Yankees | **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) (5pt) 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**.

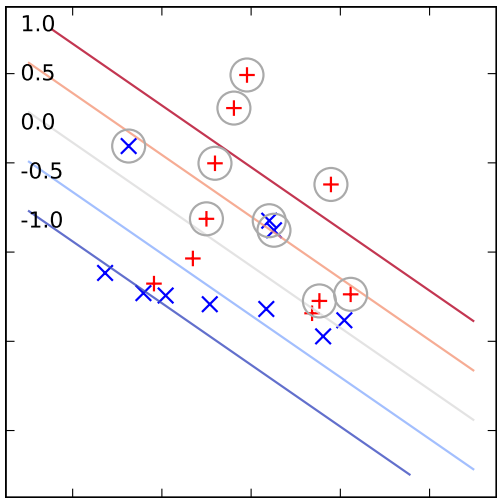
(b) (8pt) Consider the following tweet:

> 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?

# TAKE HOME, OPEN BOOK BONUS QUESTION: FAIRNESS AND CAUSALITY IN ML (25PT EXTRA CREDIT, APPLIED POST CURVE)

**Don't feel confident in your answers? Feel confident but have nothing planned for the weekend? Take these two pages home and return your *individual* solution in class on Monday 4/24/17 for a chance at 25 extra points (only to be applied after any and all grade curving).**




Displayed on the left are past applicants for jobs at a federal deperatment. Suppose that + are *qualified* job applicants and × are *unqualified* applicants from historical data. The features are performance in two competency tests. Circled points correspond to *androids* and non-circled points are *humans*. The true population is 50% human and 50% android.

There will be 10 new applicants applying for federal jobs in our department tomorrow. We want to build a classifier that will tell us who to hire. The new applicants will be drawn from the same population that past applicants were drawn from.

(a) (3pt) We want to hire approximately half of these applicants; *i.e.* our goal is $P(\text{hire}) \approx 0.5$. Choose a single threshold that is closest to this target. _____

   (a) How many qualified applicants do we expect to hire? _____

   (b) What percent of qualified androids do we expect to hire? _____

   (c) What percent of qualified humans do we expect to hire? _____

(b) (3pt) The Human Nondiscrimination Act of 2036 maintains that all federal employers must provide equal opportunities to androids and humans. What two thresholds should you choose that work for our targets so that the fraction of qualified humans that we hire is the same as for andoirds (on average)?
Human: _____ Android: _____ (Please try to get as close to the target as you can.)

   How many qualified applicants do we expect to hire? _____

(c) (3pt) The Human Non-Irrelevancy Act of 2045 amends the law to say that all federal employers must (on average) employ the same number of androids and humans, to reflect that 50-50 split in the US combined population. What two thresholds should you choose that work for our targets and hire on average the same number of humans and androids? _____ and _____

   How many qualified applicants do we expect to hire? _____

(d) (6pt) The android C-3PO didn't a get a job even though it tested better than a human who ended up being hired instead. C-3PO sued the government and won! Federal employers are no longer allowed to have two different thresholds for hiring. At the same time, the equal-proportion quotas of the Human Non-Irrelevancy Act of 2045 remain the law of the land. Damned if we do and damned if we don't?! How can we make sure not to get in trouble and still hire $\approx 50\%$ of applicants? Note that: **A.** Our classifier has to hire $\approx 50\%$ of applicants; **B.** $\approx 50\%$ of them must be human; **C.** We cannot use humanness as input to decide upon. Come up with a solution and report how many qualified applicants you expect to hire.

(e) (10pt) Applicants are deemed "qualified" if, were they hired, they would end up successfully completing 95% of tasks given to them within the first 6 months. We are not getting enough qualified candidates to keep our department running! What can we do?

Nathan and Mike offer their expert consulting help (for a significant fee). They build the above classification model to predict qualification based on test scores and come up with a truly inspired solution: to get more qualified candidates, we should provide free prep classes for our competency tests to every applicant. The prep classes are *sure* to increase any applicant's score. Then, more of our applicants will have higher test scores and therefore more of our applicants will be qualified, as predicted by the model, which was based on *very solid* ML.

**Will Nathan and Mike's plan work? Why might it or why might it not?** Try to come up with made up scenarios for both cases.