# CS5785 Applied Machine Learning - Practice Prelim

Name: Answer Key NetID: sjb344

- 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 True Positive rate against the False Positive 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**? Why? See p. 9.

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

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

**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:**

(f) (**2 pts.**) A project team reports a low prediction error on their training set and claims their method is good. **Ok** or **Problematic**? This model could just be overfitting to the training data.

(g) (**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**? Accuracy is not a good metric for such an unbalanced dataset. Just always outputting "negative" would get 99% accuracy.

(h) (**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**?

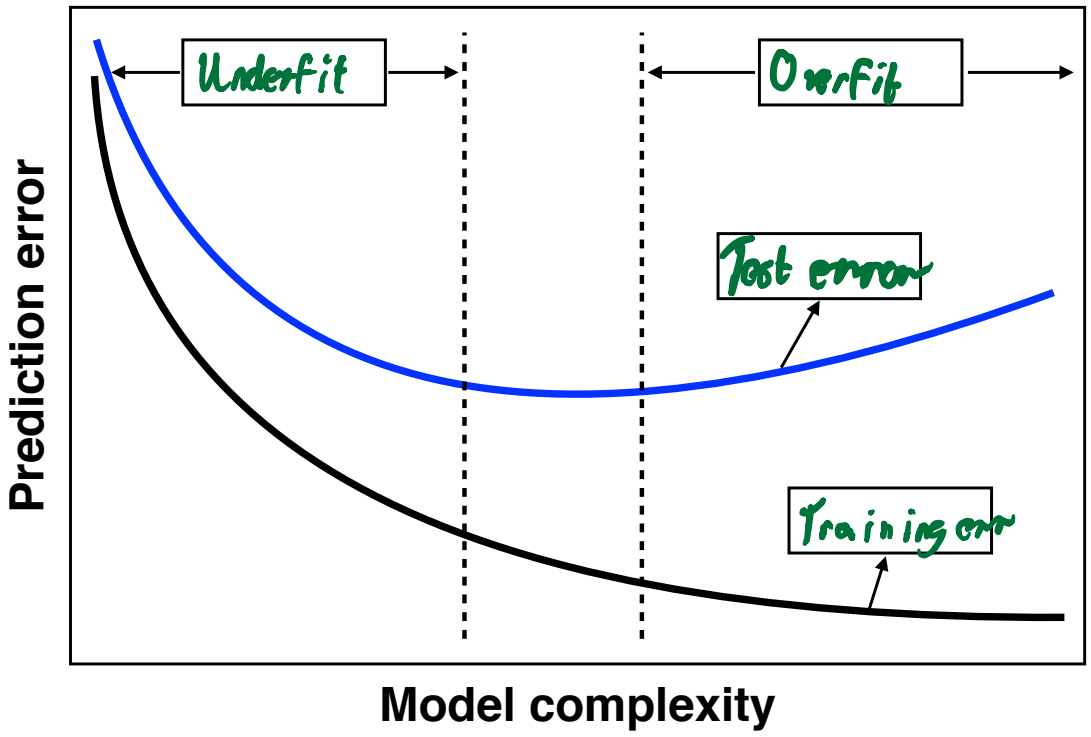
(i) (**2 pts.**) A project team did parameter selection on the full data set. Then they split the data into training and test sets. They built their model on the training set using several parameter model settings, and report the the best test error they achieved. **Ok** or **Problematic**?

This uses test data to select which version of a model to run.

Said another way, "which parameters to use" is now itself a parameter that is learned from the testing set.

## 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 $\theta$ probability of heads and coin 2 has a $2\theta$ 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.

"log likelihood" = log[P(seeing this configuration of coin flips)]

$= \log P(C_1 = H, C_2 = T, C_2 = T, C_2 = H, C_2 = T)$

$= \log[P(C_1 = H)P(C_2 = T)P(C_2 = T)P(C_2 = H)P(C_2 = T)]$ by independence assumption

$= \log[\theta\ (1-2\theta)\ (1-2\theta)(2\theta)(1-2\theta)]$

$= \log[\frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}]$

$= \log(\frac{1}{64}) = \log_2 2^{-6} = \boxed{-6}$

(b) (**10 pts.**) What is the maximum likelihood estimate for $\theta$ ?

★ Finding the $\theta$ that maximizes log-likelihood

$\frac{d}{d\theta} \log[2\theta^2(1-2\theta)^3] = 0$

$= \frac{d}{d\theta}\log 2 + \frac{d}{d\theta}\log[\theta^2] + \frac{d}{d\theta}\log[(1-2\theta)^3] = 0$

$= 0 + \frac{2\theta}{\theta^2} + \frac{3(1-2\theta)^2}{(1-2\theta)^3}(-2) = 0$

$= \frac{2}{\theta} + \frac{6}{2\theta-1} = 0$

$\rightarrow 2(2\theta-1) + 6\theta = 0$

$\rightarrow 4\theta - 2 + 6\theta = 0$

$10\theta = 2$

$\boxed{\theta = \frac{1}{5}}$

## 4 SVD AND LDA [12 POINTS]:

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

**PCA:**

- Doesn't require labels
- Finds components that maximize variance
- Used for dimensionality reduction
- Linear
- No assumptions about class distribution

**LDA:**

- Does require labels
- Finds components that maximize inter-class separation
- Also used for dim. reduction
- Also linear
- Assumes normally-distributed classes

(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.

$$M = U \Sigma V^T$$

where $\Sigma = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 & \cdots \\ 0 & \sigma_2 & 0 & 0 & \\ 0 & 0 & \sigma_3 & 0 & \\ 0 & 0 & 0 & \sigma_4 & \\ \vdots & & & & \ddots \end{bmatrix}$

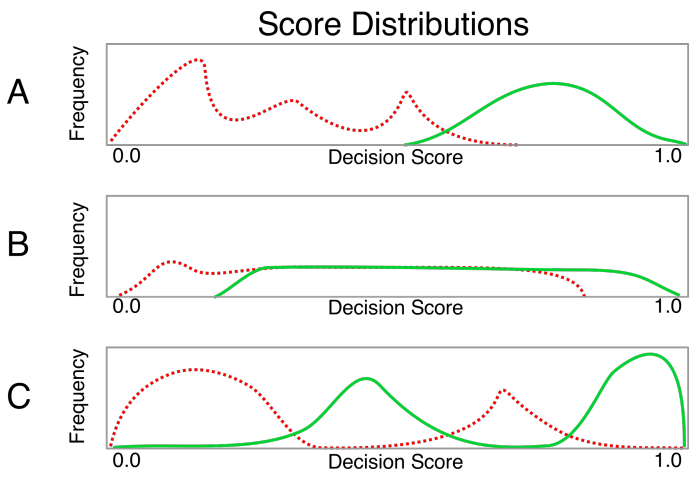
Let $\hat{\sigma}_i = \begin{cases} \sigma_i & \text{for } i \leq k \\ 0 & \text{for } i > k \end{cases}$

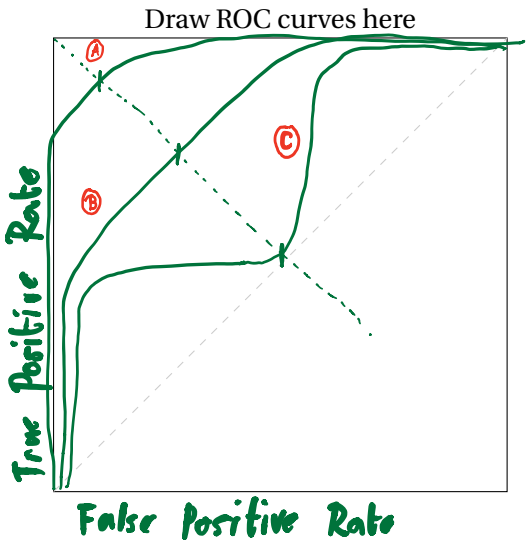
Set all but the first $k$ singular values to zero

Let $\hat{\Sigma} = \text{diag}(\hat{\sigma}_i)$

So $\hat{M} = U \hat{\Sigma} V^T$

## 5 ROC Curves and Score Distributions [25 points]:




(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.**) The equal error rate is the error at the point where the false negative rate equals the false positive rate. What is the approximate equal error rate for each classifier?

**Line A:** EER ≈ ~ 10%.
**Line B:** EER ≈ ~ 30%.
**Line C:** EER ≈ ~ 50%.

(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?

Optimize for low false positive rate, because keeping impostors out is most important.

(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?

Optimize for high recall, meaning we flag as much poisoned food as we can, throwing away some good food in the process.

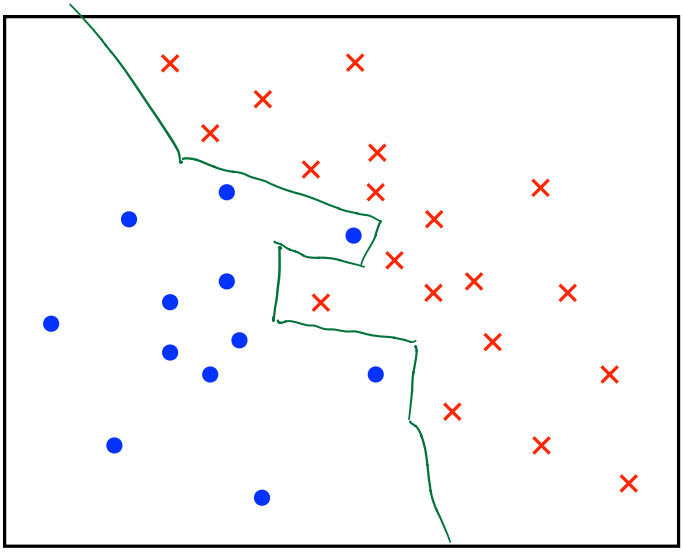
## 6 Decision Boundary [15 points]:

Consider learning a target function of the form $f : \mathbb{R}^2 \to \{-1, +1\}$. 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):

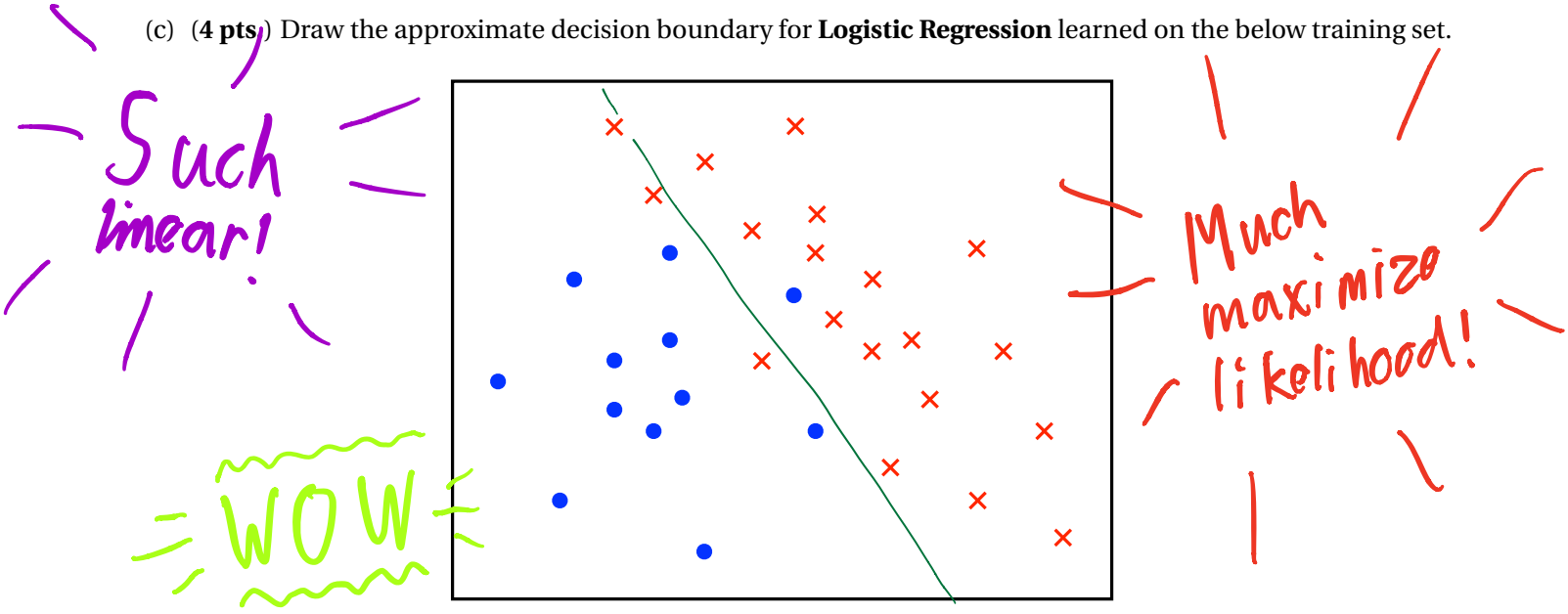
- 1-Nearest Neighbor Classifier (1-NN)
- Logistic Regression
- Gaussian Naive Bayes Classifier
  (i.e., fitting univariate Gaussian distributions to the conditional marginal distributions)

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

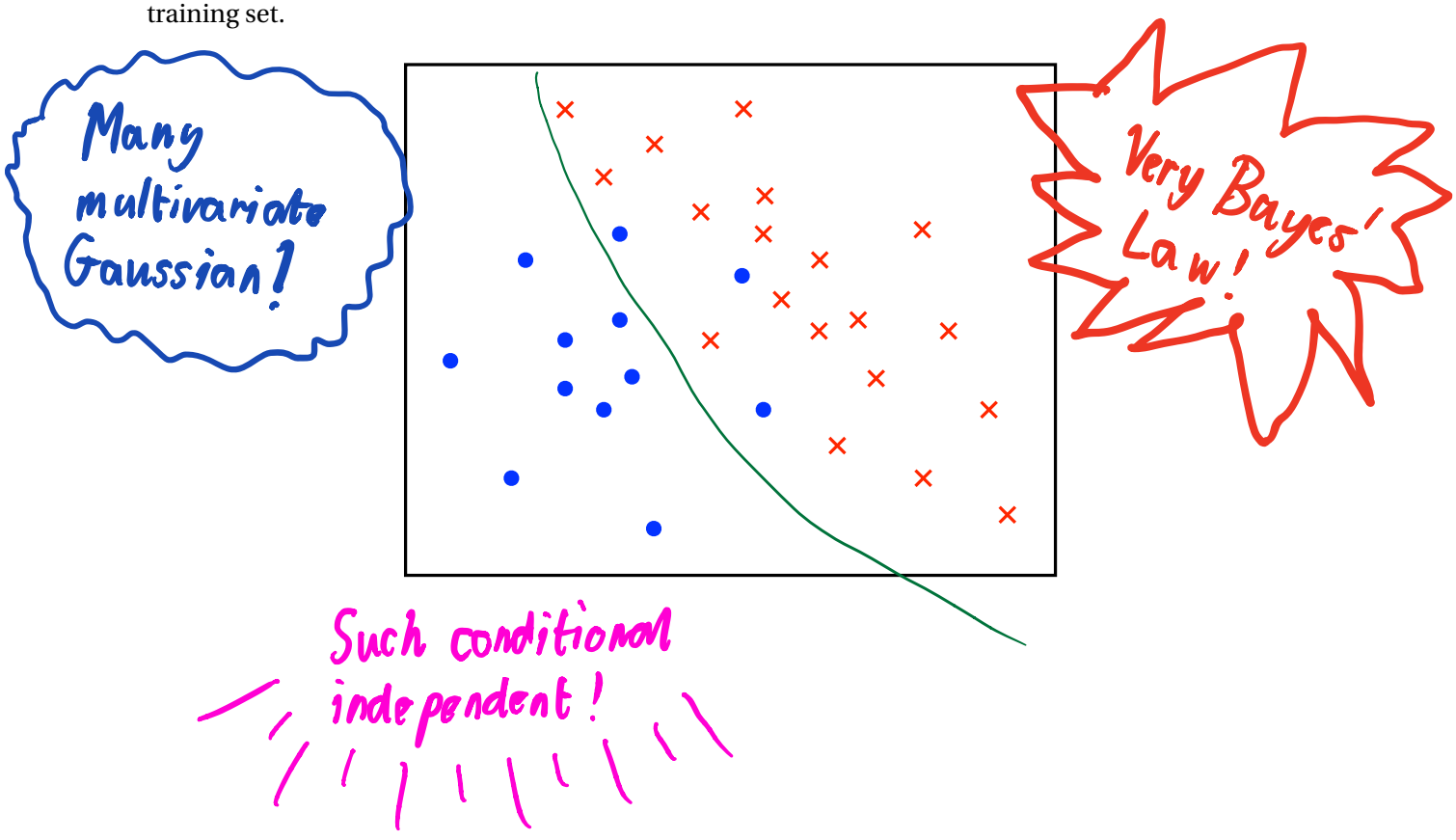
(b) (**4 pts.**) Draw the approximate decision boundary for **1-Nearest Neighbor Classifier (1-NN)** learned on the below training set.



(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.



1 (c): Why does $Var[X+2Y] \neq Var[X] + 2Var[Y]$?

$$Var[Z] = E[(Z - E[Z])^2]$$
$$= E[Z^2 - 2ZE[Z] + E[Z]^2]$$
$$= E[Z^2] - E[2ZE[Z]] + E[Z]^2$$
$$= E[Z^2] - 2E[Z]E[Z] + E[Z]^2$$
$$= E[Z^2] - E[Z]^2$$

So $Var[X+2Y] = E[(X+2Y)^2] - E[X+2Y]^2$
$$= E[X^2 + 4XY + 4Y^2] - (E[X] + 2E[Y])^2$$
$$= E[X^2] + 4E[XY] + 4E[Y^2] - (E[X]^2 + 4E[X]E[Y] + 4E[Y]^2)$$

Rearranging...

$$= E[X^2] - E[X]^2 + 4(E[Y^2] - E[Y]^2) + 4(E[XY] - E[X]E[Y])$$
$$= Var[X] + 4Var[Y] + 4Cov(X, Y)$$
$$= Var[X] + 4Var[Y] + 0 \quad \text{because } X, Y \text{ are independent}$$
$$\neq Var[X] + 2Var[Y]$$