STA 9890 - Statistical Learning for Data Mining

In-Class Test 2

This is a closed-note, closed-book exam.
You may not use any external resources other than a (non-phone) calculator.

Name: _____

Instructions

This exam will be graded out of **100 points**.

This exam is divided into three sections:

- Multiple Choice (30 points; 10 questions at three points each)
- Short Answer (40 points; 8 questions at five points each)
- Mathematics of Machine Learning (30 points)

You have one hour to complete this exam from the time the instructor says to begin. The instructor will give time warnings at: 30 minutes, 15 minutes, 5 minutes, and 1 minute.

When the instructor announces the end of the exam, you must stop **immediately**. Continuing to work past the time limit may be considered an academic integrity violation.

Write your name on the line above *now* before the exam begins.

Each question has a dedicated answer space. Place all answers in the relevant spot. Answers that are not clearly marked in the correct location **will not** receive full credit. Partial credit may be given at the instructor's discretion.

Mark and write your answers clearly: if I cannot easily identify and read your intended answer, you may not get credit for it.

Additional pages for scratch work are included at the end of the exam packet.

This is a closed-note, closed-book exam.
You may not use any external resources other than a (non-phone) calculator.

Multiple Choice: 30 points total at 3 points each

For each question, **CIRCLE** or **CHECK** your answer(s) as appropriate.

MC1. True/False: In classification, a *false negative* refers to an observation which is in the negative (0) class, but is falsely predicted as a positive (1) instead.

TRUE FALSE

MC2. True/False: *Boosting* is the practice of building an ensemble by sub-sampling features.

TRUE FALSE

MC3. True/False: Under a suitable generative model (*i.e.*, $y \sim \text{Bern}(\text{Logit}^{-1}(\boldsymbol{x}^\top \boldsymbol{\beta}_*))$), logistic regression is BLUE.

TRUE FALSE

MC4. True/False: Poisson or log-linear regression is suitable for predicting count-valued responses, such as the number of goals scored in a soccer match.

TRUE FALSE

MC5. Multiple Choice: Which of the following are properties of support vector classifiers?

☐ Use of Hinge Loss ☐ Automatic Identification of Kernel Points ☐ Lack of Tuning Parameters ☐ Probabilistic Output ☐ Insensitivity to training data far from the margin

Select *all* that apply.

MC6. Multiple Choice: Which of the following are discriminative classifiers?

☐ LDA ☐ SVM ☐ Random Forest ☐ Decision Trees ☐ Boosting ☐ QDA ☐ Bayes' Rule

Select *all* that apply.

MC7. True/False: A *maximum likelihood* estimator is one which sets the unknown parameters in order to maximize the negative log PDF/PMF of the sampling distribution on the observed data.

TRUE FALSE

MC8. Multiple Choice: Which of the following **ARE NOT** convex approximations to 0/1 Accuracy loss in classification?

☐ Hinge Loss ☐ Smoothed Hinge Loss ☐ Gini Coefficient ☐ False Negative Rate ☐ Binomial Deviance Loss ☐ Tree Loss

Select *all* that apply.

MC9. True/False: The main purpose of *boosting* is to iteratively refine our predictor by compensating for previous prediction errors.

TRUE FALSE

MC10. True/False: We cannot use cross-validation to tune the regularization parameter (λ) of logistic ridge regression for maximum 0/1-Accuracy because 0/1-Accuracy loss function is nonconvex.

TRUE FALSE

Short Answer: 40 points total at 5 points each

SA1. Apple devices support FaceID as an alternative to traditional password-based authentication. (In this context a 'positive' refers to an authorized user.) Label each of the following scenarios as a false positive (FP), false negative (FN), true positive (TP), or true negative (TN).

_____ I am able to successfully authenticate on my phone.

_____ My phone refuses to authenticate me because I just woke up and my hair is a mess ('bed head').

_____ My phone is stolen by my evil twin who is then able to access it because his face matches mine.

_____ My phone cannot be accessed by my kids when they take it without permission.

_____ My wife is added as a second user profile on my phone and she is able to use it to check my emails for me while I am driving.

SA2. Given a binary classifier, how can you use it to perform to multi-class classification? *(Hint: "One- vs-rest" approaches may be easier to describe.)*

SA3. Compare and contrast *bagging* and *stacking*. Give at least 2 similarities and two differences.

SA4. Describe the three parts of a *generalized linear model*, noting their general role in GLM specification and precisely identifying in logistic regression:

I. Name: _____

- Purpose: _____
- In logistic regression:_____

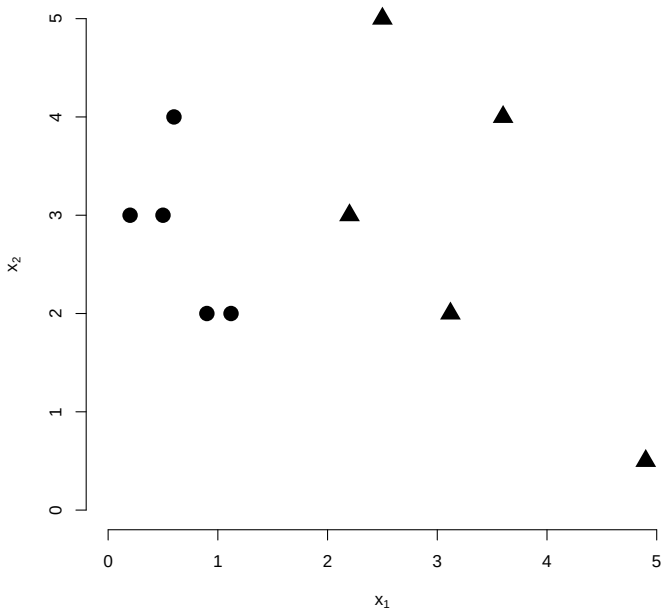
II. Name: _____

- Purpose:_____
- In logistic regression:_____

III. Name: _____

- Purpose:_____
- In logistic regression:_____

SA5. Given the following data, draw the decision boundary estimated by a maximum margin classifier and mark the support points by circling them.



SA6. Given the following set of classification outcomes, compute the Folwkes-Mallows (FM) Index:

		Ground Truth +	Ground Truth −
Prediction	+	50	10
	−	10	1000

Recall that

$$\text{FM} = \sqrt{\text{PPV} \times \text{TPR}} \text{ where PPV} = 1 - \text{FDR} = \frac{\text{TP}}{\text{TP} + \text{FP}} \text{ and TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

FM = _____

SA7. Give an example of an *ordinal* classification problem and explain *in a concrete problem-specific sense* why it cannot be approached as a binary or multiclass classification problem (*i.e.*, your answer needs to be more substantial than "because it is ordinal.").

SA8. Compare and contrast *Naive Bayes* and *Quadratic Discriminant Analysis*. Give at least 2 similarities and two differences.

Mathematics of Machine Learning: 30 points total

In this section, you will develop your own *multinomial generative classifier* to determine whether a given email is valid ("ham") or spam.

Before we get into the mathematics, recall that a *multinomial* distribution is a generalization of the binomial distribution (with a categorical sampling scheme replacing the Bernoulli). Specifically, a K-class multinomial is characterized by a sample size $n \in \mathbb{N}$ and a probability vector $\mathbf{p} = (p_1, p_2, \ldots, p_K)$ where $\sum_i p_i = 1$ and all $p_i \geq 0$. The PMF of an observation is then given by

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) = \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_K = x_K) = \frac{n!}{x_1! x_2! \ldots x_K!} p_1^{x_1} p_2^{x_2} \ldots p_K^{x_K}$$

where $(x_1, x_2, \ldots, x_K)$ are the number of observations in each category.

For example, if a 3-class multinomial has probability parameters $(0.5, 0.25, 0.25)$ and we observe $(3, 1, 1)$, the PMF of that observation is:

$$\frac{5!}{3!1!1!} 0.5^3 0.25^1 0.25^1 = \frac{120}{6 * 1 * 1} (0.125)(0.25)(0.25) = 0.15625$$

You want to use a multinomial generative classifier to distinguish emails based on certain words. After discussion with your IT department, you have collected a series of valid and spam emails and found that they contain the following word counts:

Word	Valid	Spam	Total
Deal	20	80	100
Double	10	100	110
Money	80	100	180
Free	20	100	120
Spreadsheet	20	5	25
Revenue	40	12	52
Classifier	10	3	13
Total	200	400	600

(Emails may contain other words, but you do not include them in your model.) In this context, we are using a *bag of words* approach, where the only thing that matters is the counts of various words, not the order in which they appear or any other words not on our list.

You also know that your company's domain receives nine times as many spam messages as valid ones.

(1) Given the above information, what should your *prior* probabilities before for $\mathbb{P}(\text{Valid})$ and $\mathbb{P}(\text{Spam}) = 1 - \mathbb{P}(\text{Valid})$? (3 points)

- $\mathbb{P}(\text{Valid}) =$ _____
- $\mathbb{P}(\text{Spam}) =$ _____

---

(2) Using the above text, what are the $\mathbf{p}$ probabilities for both classes? Write your answer as two probability vectors. (5 points)

- $\mathbf{p}_{\text{Valid}} =$ _____
- $\mathbf{p}_{\text{Spam}} =$ _____

---

(3) In order to calibrate the decision boundary for your classifiers, you perform a user experience study that reveals it takes 2 minutes on average to discard a spam email, while it takes 10 minutes to find an improperly labeled valid email and move it to the inbox.

What *posterior* probability threshold should you select in order to minimize the expected amount of wasted time? (Find p_{thresh} such that if the posterior probability of being spam is greater than p_{thresh}, the optimal choice is to treat the email as spam.)

Hint: When the posterior probability is equal to the threshold, the expected time loss of both decisions is equal. (5 points)

A message should be labeled as spam if its posterior probability of being spam is greater than _____%

---

(4) Your classifier receives a new message with the text:

> Hello Friend!
>
> I have a great deal for you - if you send me $100 today, I will double your money and send you $200 dollars next week. That is $100 absolutely free!!
>
> How am I able to offer such an amazing deal? I have a system for investing in the markets. I can study price patterns and identify stocks that are going up and ride them to the moon! No spreadsheet needed - just pure skill.
>
> I'm offering you this opportunity to double your money because I believe that this path to properity should be free to all. We don't need any fancy banks with their lies - power to the people!

i. What is the data vector ($\boldsymbol{x}$) associated with the above text? (5 points)

$\boldsymbol{x} =$ _____

---

ii. What is the PMF of each class associated with this data vector? (5 points)

- $\mathbb{P}(\boldsymbol{x}|\text{Valid}) =$ _____
- $\mathbb{P}(\boldsymbol{x}|\text{Spam}) =$ _____

---

iii. What are the *posterior* probabilities of each class? (5 points)

- $\mathbb{P}(\text{Valid}|\boldsymbol{x}) =$ _____
- $\mathbb{P}(\text{Spam}|\boldsymbol{x}) =$ _____

---

iv. Should this email be marked as spam or not? (2 points)

(Blank page for scratch work - not graded)

(Blank page for scratch work - not graded)