CSCE 633: Machine Learning

EXAM # 1

Fall 2019

Total Time: 50 minutes

Name: _____

UID: _____

Question	*Point*	*Grade*
1	*25*	
2	*25*	
3	*25*	
4	*25*	
Total	*100*	

Person Sitting to Your Left:

Person Sitting to Your Right:

1. (25 points) Concepts.
For each of the following questions, please provide your answer and an explanation/justification. Please answer the following questions

a) (5 points) Please explain the difference between Bootstrapping and k-fold cross-validation. What are the pros and cons of each?

Bootstrapping takes a random test/train split and repeats it a lot of time. The random splitting + repetition provides for good robustness to variability

K-fold cross-validation splits data into k equal parts where each then takes turn as the test set once to provide some robustness in variability.

Bootstrapping needs a lot of repetitions and may take time
k-fold allows for every element to be the test subject just once but isn't as robust

3 points – difference
2 points – pros and cons

b) (5 points) What are the differences between Bagging and Boosting? What are some pros and cons of each?

Bagging is just creating a bunch of random decision trees – while boosting sequentially builds trees up.
Bagging Pros: Robust to overfitting – allows for better variability Cons; might have all trees favor dominant variables
Boosting Pros: Very good at building strong classifiers – accounts for re-weighing samples
Cons: Can overfit – learns very slowly

3 points – differences
2 points – pros and cons

c) (5 points) Why does Lasso regularization select features while ridge regression does not? Please provide any formulations necessary to support your answer. (Be precise!)

Because of L1 it drives features to 0 whereas L2 shrinks them but usually does not hit 0.
Formulation: |x1| + |x2| < s vs. b1^2 + b2^2 < s

+3 for reason
+2 for formulation (< s)

d) (10 points) Please provide the formulation for the vanilla Elastic Net. Please provide some pros and cons to this formulation. How can this be solved using just a regular lasso solver?

L(beta, lamba1, lambda2) = RSS + lambda2 ||beta||_2^2 + lambda1 \\beta_1

B_tilde = argmin RSS + \lambda1 / (\sqrt (1 + lambda 2)) \\beta_tilde\\ 1

X_tilde = 1 / (sqrt (1 + lambda 2) (x, sqrt(lambda 2) Ip) <- identify matrix – use ISTA
+5 for formula
+5 for writing L2 in terms of L1 and then solving with ISTA

2. (25 points) Supervised Machine Learning Models

a) (5 points) Which of the follow are possible hyperparameters of the corresponding models? Select all that apply.

1) The number K neighbors in K-NN classification.
+1

2) The number of trees, depth of each tree, and weight of each weak learner in Gradient Descent Boosting.
-.5 if they pick it (weight)

3) The step size in gradient descent.
+1

4) The λ weight of regularization.
+1

5) The weights β in logistic regression.
Nope

b) (5 points) Assume a non-linear regression model that predicts the stock price price $\in$ R of a company based on the average number of sales $\in$ R, ranging between 1 and 900, such that price = g_w(sales,$\sqrt{sales}$). Further assume that the regression equation is written as $g_w = w_0 + w_1x_1 + w_2x_2$. The weights w of the model are found through gradient descent. What would be good choices of x_1 and x_2 so that the gradient descent method converges in a reasonable time? Justify your answer for full credit.

A. $x_1 = \text{sales}, x_2 = \frac{\sqrt{\text{sales}}}{30}$

B. $x_1 = \text{sales}, x_2 = 1000 * \sqrt{\text{sales}}$

C. $x_1 = \frac{\text{sales}}{30}, x_2 = \sqrt{\text{sales}}$

D. $x_1 = \text{sales}, x_2 = 30 * \sqrt{\text{sales}}$

D -> sqrt (up to 900) -> 30

+3 for choice
+2 for justification

c) (5 points) Recall that Precision and Recall are defined at a specific probability threshold as:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Please calculate the precision, recall and F1 score with thresholds p=0.45 and p=0.55. (You can approximate the fraction values to 2 decimal points).

Predicted Probability	0	5	10	14	24	50	53	57	75	100
Ground Truth	No	No	No	Yes	No	Yes	Yes	No	Yes	Yes

+1 P, +1 Rec, .5 F
+1 P, +1 Rec, .5 F

P = 0.45 +3

Prediction No No No No No Yes Yes Yes Yes Yes
TP = 4
TN = 4
FP = 1
FN = 1

Precision = 4/5
Recall = 4/5 +
F1 = 2 * (P R) / (P + R) = 2 * (.8 *.8) /(.8 + .8) = .8 + 1

P = 0.55 + 2

Prediction No No No No No No No Yes Yes Yes
TP =2
TN = 4
FP = 1
FN = 3

Precision = 2/2 +1 (+ 2

Recall = 2/2 + 3 (+ 2
F1 = 2 * (P R) / (P + R) = 2 * (.8 *.8) /(.8 + .8) = .8

d) (10 points) Please show two iterations of gradient descent with the following problem. Assume we want to find the right regression coefficients for the following dataset – to estimate final grade based upon hours studied:

Hours Studied	9	8	1	6	2	10
Grade	98	73	24	57	18	83

Start with $\beta_0 = 0$, $\beta_0 = 10$, $\alpha = 0.01$, and $\varepsilon = 0.01$

Start with $\beta_0^{(0)} = 0$, $\beta_1^{(0)} = 10$, $\alpha = 0.01$, and $\varepsilon = 0.01$. α is the step size and ε is the stopping criterion. You can round your intermediate results to simplify the computation. It's enough to get approximated results.

+5 formulation

+5 computation

```
Iteration: 1 New Beta 0: -0.0116666666666667 New Beta1: 9.73 magnitude: 27
.0251940069098
```

```
Iteration: 2 New Beta 0: -0.0070166666666667 New Beta1: 9.5894 magnitude:
14.0676872655032
```

3. (25 Points) Logistic Regression

Assume you have collected data for a group of students with variables X_1 = hours studied, X_2 = undergrad GPA, and outcome Y = whether they received an A. You fit a logistic regression model and produce estimated coefficients $\widehat{\boldsymbol{\beta_0}} = -\mathbf{6}$, $\widehat{\boldsymbol{\beta_1}} = \mathbf{0.05}$, and $\widehat{\boldsymbol{\beta_2}} = \mathbf{1}$

a) (5 points) Please write the logistic regression formulation for probability of an A given some new text subject X. (You may leave the answer in terms of log or exp). Then, estimate the probability that a student who studies 40 hours and had an undergrad GPA of 3.5 gets an A.

log (p/ 1-p) = b0 + b1x_1 + b2_x2

Or

P(A) = exp(b0 + b1x_1 + b2_x2) / (1 + exp (b0 + b1x_1 + b2_x2))

+2 for the log odds

+2 for the linear formula

+1 to show probability of A

b) (5 points) How many hours would the student in (a) need to study to have a 50% chance of getting an A?

$50 = \exp(b0 + b1\_x1 + b2\ (3.5) / (1 + \exp\ (b0 + b1x\_1 + b2\_3.5)$

+3 fixing 3.5 and intercept and just varying x1

+2 for correct x1

c) (10 points) Assume you add a new categorical variable to the model, X3, that represents the major: EECS, Mathematics, Statistics. You create three one-hot binary variables for each major and fit a new logistic regression model. You measure the z-statistic for each variable and are given the following table:

	Coefficient	Z-statistic	p-value
Intercept	2	22.08	< 0.001
Hours Studied	0.05	24.74	< 0.001
GPA	1	0.37	< 0.001
Major: EECS	1.5	2.74	0.0023
Major: Math	0.25	2.12	0.0014
Major: Stats	0.085	0.38	0.7115

Explain how the model works and the relative (ordered) importance of the variables

+5 order of importance (EECS, GPA Math, Studied, Stats… intercept?) (allow variance if they say GPA because of scaling)

+5 understanding that Stats is not included but is included in the intercept

d) (5 points) You now want to compare your model's performance from part (a) with your model from part (c). What metrics can you use to compare the models? Give one sentence to explain how each metric works.

+3 for any measures of comparisons
+2 if they correctly state they need ones that allow model comparisons with different number of predictors

4) (25 Points)

a) (5 points) The formulation for the Maximal Marginal Classifier is the following:

$$x_1, \cdots, x_n \in \mathbb{R}^p$$
$$y_1, \cdots, y_n \in \{-1, +1\}$$

Then we want to:

$$maximize_{\beta_0, \beta_1, \cdots, \beta_p, M} M$$

Subject to constraints:

$$\sum_{j=1}^{p} \beta_j^2 = 1$$

and

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M \forall i = 1, \cdots, n$$

How do you change this formulation to the soft-margin classifier (Support Vector Classifer)? Please give the formulation:

+3 for slack variable in constraint
+2 for slack variable constraint with C

b) (10 points) Please explain what each parameter in the Support Vector Classifier does and how changing the values impacts classification performance:

slack variable – error (+3)

M – defines width of margin (+3)

C – how much error is allowed (+3)

B – defines hyperplane (+1)

c) (15 points) Support Vector Machines are defined as:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i)$$

Please provide formulation for 3 popular kernels and explain how they work/are different from each other.

+3 for kernel + 2 for how it works