CS6316-Fall2017

Exam-ML-CS6316 for dra2zp

Denny Anderson

Your original score on this exam was 99 of 110 points (90.00%).

Your breakdown of points per page is below.

Page	Score	Max
1	0	0
2	10	10
3	9	10
4	20	20
5	7	10
6	20	22
7	16	18
8	17	20
9	0	0
10	0	0

A graded scan of each page follows.

CS6316 Exam – Fall 2017

Name: Denny Anderson

Pledge:

On my honor as a student, I have neither given nor received aid on this exam.

Denny R. Anderson III

Your signature here

You MUST do the following three things:

1. Fill in your name and sign the pledge above. If you do not sign the pledge we will not grade your exam.
2. On the top of **every page**, including this one, write your email ID (e.g., mst3k).
3. On the bottom of **this page**, bubble-in your email ID (one character per row).

Grading is done one page at a time. Keep each answer on the same page as the question it answers.

Policies

- This exam is open-book, open-notes, but closed-internet. No laptops or phones allowed. You are allowed to bring a calculator. You may use a pen or pencil.
- Be sure to **state any assumptions** you make next to the question(s) to which they apply.
- Write clearly. If your answer is illegible it cannot be graded.
- When writing your solutions, pay attention to both content and presentation/writing style.
- You have the entire class period to complete this exam.

Good Luck!

ⓐ ⓑ ⓒ ● ⓔ ⓕ ⓖ ⓗ ⓘ ⓙ ⓚ ⓛ ⓜ ⓝ ⓞ ⓟ ⓠ ⓡ ⓢ ⓣ ⓤ ⓥ ⓦ ⓧ ⓨ ⓩ
ⓐ ⓑ ⓒ ⓓ ⓔ ⓕ ⓖ ⓗ ⓘ ⓙ ⓚ ⓛ ⓜ ⓝ ⓞ ⓟ ⓠ ● ⓢ ⓣ ⓤ ⓥ ⓦ ⓧ ⓨ ⓩ
● ⓑ ⓒ ⓓ ⓔ ⓕ ⓖ ⓗ ⓘ ⓙ ⓚ ⓛ ⓜ ⓝ ⓞ ⓟ ⓠ ⓡ ⓢ ⓣ ⓤ ⓥ ⓦ ⓧ ⓨ ⓩ
⓪ ① ● ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨
ⓐ ⓑ ⓒ ⓓ ⓔ ⓕ ⓖ ⓗ ⓘ ⓙ ⓚ ⓛ ⓜ ⓝ ⓞ ⓟ ⓠ ⓡ ⓢ ⓣ ⓤ ⓥ ⓦ ⓧ ⓨ ●
ⓐ ⓑ ⓒ ⓓ ⓔ ⓕ ⓖ ⓗ ⓘ ⓙ ⓚ ⓛ ⓜ ⓝ ⓞ ● ⓠ ⓡ ⓢ ⓣ ⓤ ⓥ ⓦ ⓧ ⓨ ⓩ

Do not write in this area

Department	Course	Exam	Page	Version
● ○	○ ○ ○	●	●	●
○ ○	● ○ ○	○	○	○
○ ○	○ ○ ○	○	○	○
● ○	○ ○ ○	○	○	○
○ ○	○ ○ ○	○	○	○
● ○	○ ○ ○	○	○	○

Question 1 [**10 points**]: Sometimes a dataset can be very **unbalanced.** Let's assume there are two classes represented, 'A' and 'B', and the proportions are: A: 90%, and B: 10%.

(a) For performance evaluation of a classifier (algorithm), accuracy is a metric that is used very often. However, by itself accuracy can be misleading. Describe a scenario in which a misleading high percentage accuracy can be obtained.

4/4

The classifier could classify everything into class A, making the accuracy 90%, which is misleading because it's not making any meaningful classifications.

(b) Name a performance evaluation metric that we have talked about in class that is useful when the dataset has unbalanced classes.

cost matrix

2/2

(c) When performing k-fold cross validation on a dataset that is unbalanced, name and briefly describe one (1) strategy that can be employed to improve the effectiveness.

double resampling – within each of our k folds, we perform k-fold cross validation again; since we perform the algorithm again on each of the k folds, it significantly improves the effectiveness rather than simply doing k-fold cross validation because we're maintaining the proportion of the classes in each fold (stratified k-fold cross validation)

4/4

The score on this page is 10/10

Question 2 [**10 points**]: You are stranded on a deserted island. Mushrooms of various types grow wildly all over the island, but no other food is anywhere to be found. Some of the mushrooms have been determined as poisonous (1) and others as not (0). Consider the data:

Example	IsHeavy	IsSmelly	IsSpotted	IsSmooth	IsPoisonous
A	0	0	0	0	0
B	0	0	1	0	0
C	1	1	0	1	0
D	1	0	0	1	1
E	0	1	1	0	1
F	0	0	1	1	1
G	0	0	0	1	1
H	1	1	0	0	1

(a) What is the **entropy** of *IsPoisonous* (the class)? (Show all of your work)

$E(\text{IsPoisonous}) = -\left(\frac{5}{8}\log_2\frac{5}{8}\right) - \left(\frac{3}{8}\log_2\frac{3}{8}\right) = 0.954$

3/3

(b) Luckily you recently took the machine learning class, so you decide to create a decision tree to help you classify the mushrooms on the island. Answer parts (i) and (ii) below.

(i) Which attribute should you choose as the root of the decision tree? [Hint: you might be able to figure this out by looking at the data without explicitly computing the information gain of all four attributes]

(ii) What is the **information gain** of the chosen attribute?

3/3

I would choose the IsSmooth attribute since that would cause the largest split in the data.

3/4

IsSmooth	IsPoisonous = 1	IsPoisonous = 0	Total
0	2	2	4
1	3	1	4
Total			8

$E(\text{IsSmooth}=0) = -\left(\frac{2}{4}\log_2\frac{2}{4}\right) - \left(\frac{2}{4}\log_2\frac{2}{4}\right) = 0$

$E(\text{IsSmooth}=1) = -\left(\frac{3}{4}\log_2\frac{3}{4}\right) - \left(\frac{1}{4}\log_2\frac{1}{4}\right) = 0.811$

$G(\text{IsPoisonous}, \text{IsSmooth}) = 0.954 - 0.811 = 0.143$

0.9544- 1/2*E(isSmooth=0) -1/2*E(isSmooth=1)

=0.9543-1/2*1-1/2*0.8112 = 0.0487

The score on this page is 9/10

Question 3 [10 points]: A randomly selected cast iron skillet produced by "Ashley's Kitchen Company" was found to be defective (D). There are three factories (A, B, and C) where such skillets are produced. Following is the information known about the company's skillet production and the probability of defective skillets from each of the factories:

Factory	% of total production	Probability of defective skillet
A	0.50 = *P(A)*	0.01 = *P(D | A)*
B	0.30 = *P(B)*	0.02 = *P(D | B)*
C	0.20 = *P(C)*	0.03 = *P(D | C)*

What is the probability that the skillet was manufactured in factory **B**? (Show <u>all</u> of your work.)

$P(B|D) = P(B \cap D) \, / \, P(D)$

$= P(D|B) * P(B) \, / \, P(D)$

$= 0.02 * 0.30 \, / \, (P(D \cap A) + P(D \cap B) + P(D \cap C))$

$= 0.006 \, / \, (P(D|A) * P(A) + P(D|B) * P(B) + P(D|C) * P(C))$

$= 0.006 \, / \, (0.01 * 0.50 + 0.02 * 0.30 + 0.03 * 0.20)$

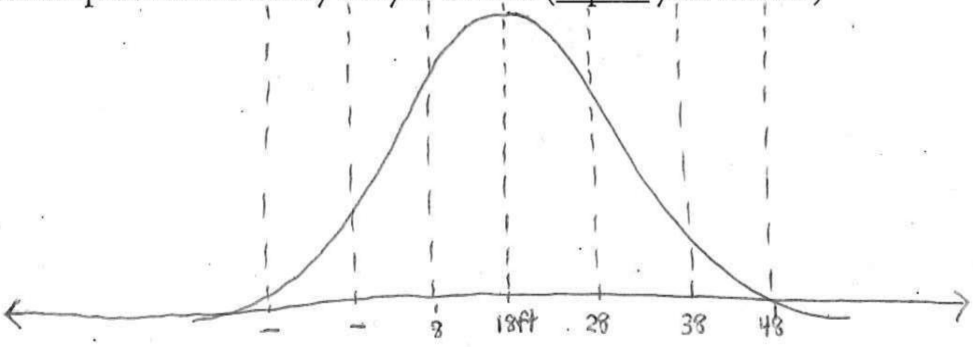
$= 0.006 \, / \, 0.017$

$= 0.353$

6/6

4/4

Question 4 [10 points]: A large river floods once annually in June. The flooding data of the last 100 years shows a normal distribution with a mean of 18 feet, and a standard deviation of 10 feet. Kim predicts the flood will be 28 feet or less. Ali predicts the flood will be between 8 and 28 feet. Whose prediction is mostly likely to be true? (<u>Explain</u> your answer.)



4/4

4/4

2/2

Kim's prediction is most likely to be correct because the range of their guesses are:

- Kim: 0-28ft
- Ali: 8-28ft

Kim's guess <u>includes</u> the range for Ali's guess, so it is more likely that her prediction will be correct.

The score on this page is 20/20

Question 5 [10 points]: **Pre-scaling** of input variables (to the same range) is often performed for supervised learning problems prior to training (or model estimation). Consider several regression methods presented in this course: k-nearest neighbor regression, linear regression (in classical statistics), SVM regression, and Multilayer Perceptron (MLP). For each of these methods (a through d below), discuss briefly whether pre-scaling of inputs is necessary (**Yes or No**), and explain why. [*Hint*: What is being measured (if anything)?] Finally answer question (e).

(a) k-NN regression:.

No – since we're checking the k closest things, it's important to not scale it so it doesn't affect the data.

1/2

(b) Linear regression:

Yes – we may want to pre-scale the input variables so outliers don't have too much of an effect on the regression algorithm.

1/2

(c) SVM regression:

No – we want to find the boundary line with the highest margin, so scaling it might affect the boundary it chooses

1/2

(d) MLP regression:

Yes – we want the learning signal to be the difference between the desired and actual neuron's response, so we want to pre-scale the input variables so that the learning signal isn't affected by outliers.

2/2

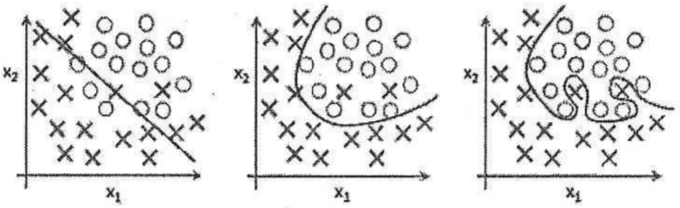
(e) Scaling input variables is one thing you can do when performing preprocessing. Name one (1) other preprocessing task and briefly describe why it is useful.

Removing outliers is also useful because it can increase the accuracy for regressions on the other data. (Ex. in class, the chart showing brain size compared to body weight – humans + dinosaurs were outliers, and removing them increased the accuracy for other organisms because it improved the regression algorithm).

2/2

The score on this page is 7/10

Question 6 [10 points]: Three students compete with one another to come up with a classifier that will attain the best prediction accuracy on future data samples. The three students came up with the models shown below (represented by the diagrams):



Student 1 Student 2 Student 3

6/6

4/4

Using **Occam's Razor**, and any other arguments from predictive learning, discuss which student will be the most successful. Explain why.

Student 2 will be the most successful because Occam's Razor states that the simpler the model, the better. If we make it more complex like Student 3, we risk overfitting the data. However, if we make it any simpler like Student 1's linear regression, we risk underfitting the data and losing accuracy. Simpler is typically better until we begin to make it too simple and lose accuracy.

Question 7 [12 points]: Apply **Popper's** philosophical arguments to determine if the following assertions can be qualified as scientific hypotheses or not. Briefly explain your answers.

(a) We cannot see our noses, because they all live on the Moon

This statement can be qualified because Popper's falsifiability means that we can physically go to the moon to check if our noses are there.

4/4

(b) The weather at any given moment reflects all the physical factors such as temperature, humidity, pressure, clouds etc. that are responsible for weather conditions

This statement cannot be qualified because Popper would say that it's impossible to determine all the possible factors that determine the weather conditions.

4/4

(c) Any two objects dropped from the same height (in the vacuum, to remove the effect of air resistance), will hit the ground at the same time

~~This statement can be qualified because Popper's falsifiability says that we can test this by drop~~

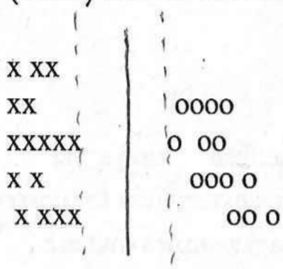
2/4

This statement cannot be qualified because Popper would want us to check the truth of this for all objects, which cannot physically be done.

The score on this page is 20/22

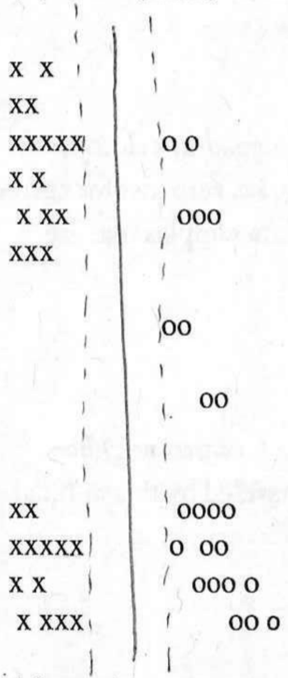
Question 8 [**18 points**]: The following question requires you to draw on the following diagrams based on your understanding of **Support Vector Machines** (SVM):

(a) Show (<u>draw</u>) the linear SVM decision boundary for the following bivariate training data: (binary class labels of data samples are indicated by symbols 'x' and 'o')



6/6

(b) Show (<u>draw</u>) the linear SVM decision boundary for another linearly separable dataset:



6/6

(c) Do these SVM models in part (a) and (b) have approximately the same margin size?

Circle: [✓ Yes] No

2/2

(d) Which SVM model (a) or (b) is expected to have better prediction (smaller test error) for future samples (from the same distribution as training data)? <u>Briefly explain</u> your answer.

SVM model (b) is expected to have higher prediction for future samples because it was trained on more data.

2/4

The score on this page is 16/18

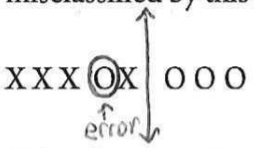
Question 9 [20 points]: Consider the following classification problem: a single input (temperature x) is used to decide whether a patient is healthy (class X) or sick (class O). The training data for this problem is shown below, where single-input (x) feature values are represented as coordinates in a horizontal direction:

input x-axis →

XXX OX OOO

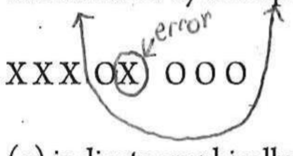
For this data set:

(a) indicate graphically (<u>draw</u>) the decision boundary obtained using a *linear classifier* via empirical risk minimization (using standard classification loss function, i.e. zero loss for correct / unit loss for incorrect classification decisions). Also, circle errors, i.e. data samples that are misclassified by this optimal decision boundary.



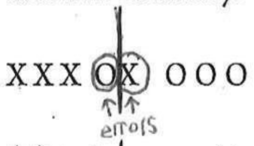
4/4

(b) indicate graphically (<u>draw</u>) the decision boundary obtained using a *quadratic classifier* via empirical risk minimization (using standard classification loss function, i.e. zero loss for correct / unit loss for incorrect classification decisions). Also, circle errors, i.e. data samples that are misclassified by this optimal decision boundary.



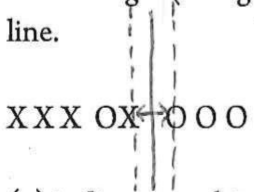
3/4

(c) indicate graphically (<u>draw</u>) the decision boundary obtained using a *k-nearest neighbor classifier* (with k=3). Also, circle errors, i.e. data samples that are misclassified by this optimal decision boundary.



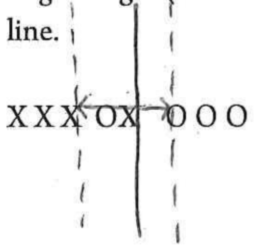
2/4

(d) indicate graphically (<u>draw</u>) the decision boundary estimated using *linear SVM classifier* with *small margin* (~ large value of tuning parameter C). Also, show the margin borders in dashed line.



4/4

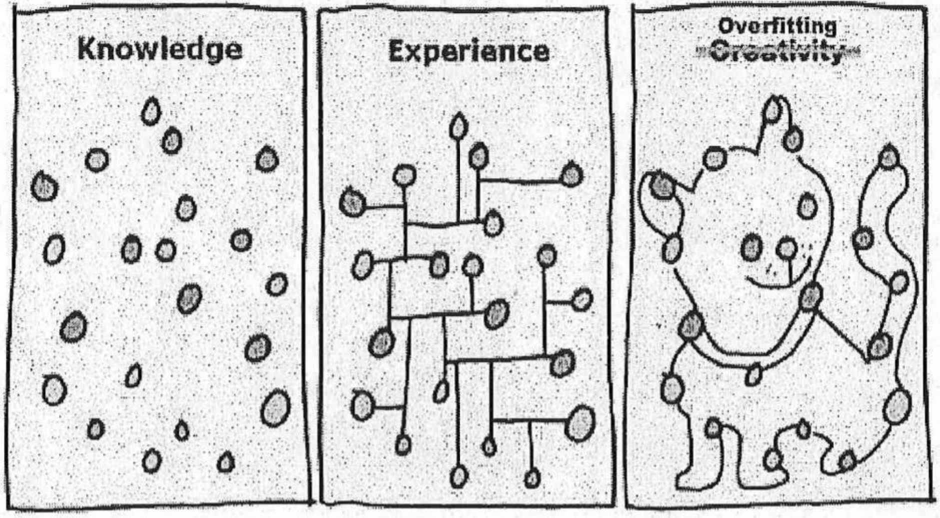
(e) indicate graphically (<u>draw</u>) the decision boundary estimated using *linear SVM classifier* with *large margin* (~ small value of tuning parameter C). Also, show the margin borders in dashed line.



4/4

The score on this page is 17/20

Scratch paper – you can use this page for notes, nothing on this page will be graded. Do not tear this page from your exam.



Smile! You've completed the exam! ☺

Scratch paper – you can use this page for notes, nothing on this page will be graded.
Do not tear this page from your exam.