Bus 674

Machine Learning I

EXAM
Objective Questions: T/F and Multiple Choice

Fall 2020

Instructions:

This exam is open book, open notes, and open computer and open internet. That is, you may access the internet as a resource for documentation, to look at responses to questions on technical websites, and so on. You may not have any synchronous communication with another person during the exam or post questions to forums in hopes of getting a fast response, or any other behavior contrary to the idea that the exam is your own individual work.

The exam is intended to be a 3-hour exam. However, I am allowing an extra half hour for downloading and printing the exam and for filling in the Google Forms answer sheet and uploading your files. Thus, the exam must be submitted by 4:30pm.

There are two parts to this exam. This document contains the objective part (T/F and Multiple Choice). There are a total of 52 points on this part of the exam. The long answer part (Part 2) is worth 48 points for a total of 100 points on the exam.

For the multiple choice questions, choose the best answer. For numerical answers, if your answer does not match any of the choices exactly, check your work, and then choose the closest answer.

The questions are not intended to be ambiguous or deliberately tricky, so keep that in mind when you are answering them. If you find that you are splitting hairs, you have probably missed the point of the question or you are over-thinking things.

I reserve the right to adjust the point values in response to analysis of how students have responded to the questions as it is difficult or impossible to anticipate all of the issues that might occasionally come up.

When you submit you answers to this part of the exam, sign the electronic pledge at the end of the answer sheet. The pledge applies to both parts of the exam.

Part 1: True/False (10 points, 1 points each)

(1) T F The cube-root transformation ($x^{1/3}$) is a stronger transformation than the square-root transformation ($\sqrt{x}$) in that it will correct more severe right skewness in the distribution of a variable.

(2) T F "Slowing the learning rate" in fitting a machine learning model frequently means reducing the step length in the numerical optimization method that is used to train the model.

(3) T F When building a complex machine learning model, if you want an estimate of your final model's performance that is not biased by fitting or model selection, then you need to have a hold-out test data set.

(4) T F Figure 1 below shows some output produced by a step while performing a step-wise logistic regression using the step() function in R. At the beginning of this step, the variable "weight" was in the model.

Figure 1

	Df	Sum of Sq	RSS	AIC
+ origin	1	220.8	4348.1	951.24
<none>			4569.0	968.66
+ acceleration	1	10.5	4558.5	969.77
+ cylinders	1	5.0	4564.0	970.24
+ horsepower	1	3.3	4565.7	970.38
+ displacement	1	0.0	4568.9	970.66
- year	1	2752.3	7321.2	1151.49
- weight	1	11222.4	15791.3	1452.81

(5) T F You should rescale the x-variables in exactly the same way if you are training a kNN (k nearest neighbor) model or if you are training a neural net.

(6) T F A bootstrap sample has the same number of observations (i.e., the same sample size) as the sample from which is it drawn but only contains about 2/3rds of them.

(7) T F In an ROC plot, the specificity is plotted on the y-axis.

(8) T F Bagging and random forests are examples of ensemble methods.

(9) T F Suppose you are fitting a regression where both you Y-variable and X-variables are continuous variables. If the normal probability plots (Q-Q plots) of your Y-variable and all of your X-variables are approximately straight lines, then the regression of Y on the X-variables is likely to be close to linear.

(10) T F The key difference between supervised and unsupervised learning is that, in supervised learning you have, for each observation, either the correct label (for classification problems) or actual resulting value (for prediction problems) while in unsupervised learning you do not.

Part 2: Multiple Choice (42 points total, about 3 points each)

(1) Suppose that in a machine learning application that you have been working on, you have randomly divided your data into training, validation, and test samples. Among all of the approaches you have considered, you have selected the one what you believe is the best one. Now that the best procedure has been selected, you have gone ahead and evaluated its performance on the test sample.

Consider the performance of your approach on the three samples (training, validation, and test). The performance is most likely the best on which is the three samples? For which is it most likely to be the worst?

(a) Best on the test sample. Worst on the training sample.

(b) Best on the test sample. Worst on the validation sample

(c) Best on the validation sample. Worst on the training sample.

(d) Best on the training sample. Worst on the test sample.

(e) Best on the validation sample. Worst on the test sample.

(f) None of the above.

(2) For which of the following does a larger value indicate better performance?

(a) AIC (b) AUC (c) k in kNN Classification

(d) the bias (e) The root MSE (f) None of the above.

(3) Which of the following has a meaning that is substantially different from the others.

(a) Training the model.

(b) Transforming the model.

(c) Fitting the model.

(d) Estimating the model.

(e) All of the terms have the same meaning.

(f) All of the terms have substantially different meanings.

(4) Which of the following occurs at every node split in constructing the trees that make up a random forest.

(a) Drawing a bootstrap sample.

(b) Calculation of the AIC.

(c) K-fold cross-validation.

(d) Regularization.

(e) Random selection of a subset of the x-variables (the features) to consider.

(f) None of the above.

(5) Which of the following functions is not typically used as an activation function in a neural net?

(a) the sine function (b) the ReLU function (c) the logistic function

(d) the hyperbolic tangent function (e) they are all frequently used as activation functions

(f) None are typically used as activation functions.

(6) In logistic regression, the log odds is modeled as a linear function of the x-variables. For a particular dataset, the fitted logistic regression equation is

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -2.3 + 2x_1 - 0.5x_2 .$$

What is the predicted value of $\hat{p}$ when $x_1 = 2.2$ and $x_2 = 3.6$?

(a) 0.43 (b) 0.77 (c) 0.99

(d) 0.57 (e) 0.98

(f) None of the above.

(7) The $\sqrt{MSE}$ of a particular prediction is 30. The standard deviation of the prediction is 24. What is the bias?

(a) 6 (b) 324 (c) 14.5

(d) 18 (e) 2.5

(f) None of the above.

(8) After transforming the variables and fitting regular least squares regression to a data set, we end up with the following fitted model:

$$\log(\hat{Y}) = 1.34 + 2.21\sqrt{x_1} - 1.32\log(x_2) + 1.64x_3^{1/3}$$

What is the prediction $\hat{Y}$ when $x_1 = 4$, $x_2 = 6$, and $x_3 = 0.5$.

(a) 4.7 (b) 21.8 (c) 109.6

(d) 28.7 (e) 67.7 (f) None of the above.

(9) Figure 2 shows the output of a fairly simple tree fit to the Auto data. Based on this tree, what is the predicted value for a car with the following characteristics:

Year = 80
Cylinders = 6
Horsepower = 90
Displacement = 180
Weight = 2200
Acceleration = 18.6
Origin = 2

(a) 19.44 (b) 33.12 (c) 36.22

(d) 26.71 (e) 33.67 (f) None of the above.

Figure 2

```
node), split, n, deviance, yval
      * denotes terminal node

 1) root 392 23820.0 23.45
   2) displacement < 190.5 222  7786.0 28.64
     4) horsepower < 70.5 71  1804.0 33.67
       8) year < 77.5 28   280.2 29.75 *
       9) year > 77.5 43   814.5 36.22 *
     5) horsepower > 70.5 151  3348.0 26.28
      10) year < 78.5 94  1222.0 24.12
        20) weight < 2305 39   362.2 26.71 *
        21) weight > 2305 55   413.7 22.29 *
      11) year > 78.5 57   963.7 29.84
        22) weight < 2580 24   294.2 33.12 *
        23) weight > 2580 33   225.0 27.46 *
   3) displacement > 190.5 170  2210.0 16.66
     6) horsepower < 127 74   742.0 19.44 *
     7) horsepower > 127 96   457.1 14.52 *
```

(10) Figure 3 shows some calculations for the e-mail spam dataset we used in class for the Naïve Bayes example. Specifically, it shows the marginal probabilities of an e-mail being spam or ham together with two tables giving the conditional probabilities of the frequency categories (Zero, Low, Med, High) for four words (features) computed conditionally on the e-mail being spam or ham. These conditional probabilities are computed for each word separately, consistent with the naïve Bayes approach (nothing tricky or unexpected here).

A new e-mail arrives with the follow values for the four features:

Order = Zero
Free = High
Credit = Low
Money = Zero

What is the predicted probability that the e-mail is spam using the naïve Bayes approach.

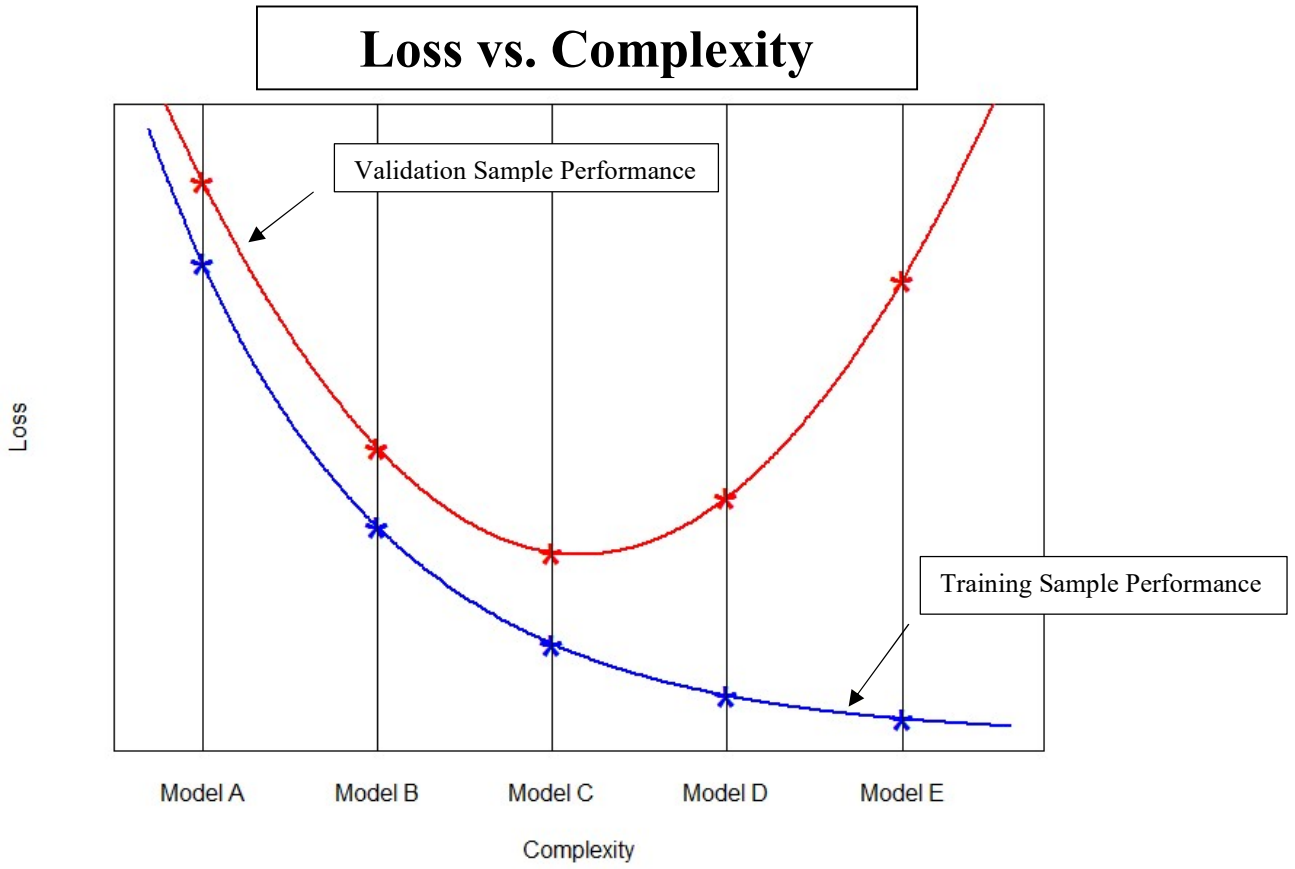
(a) 0.99 (b) 0.16 (c) 0.71
(d) 0.78 (e) 0.57 (f) None of the above.

Figure 3

Note: For your ease of calculation, I have also provided an Excel file containing this information.

P(Spam) =	0.4033			
P(Ham) =	0.5967			
P(Feature|Spam)				
	Order	Free	Credit	Money
Zero	0.5697	0.6685	0.6891	0.4546
Low	0.1114	0.0782	0.0647	0.1186
Med	0.2147	0.1688	0.1599	0.2875
High	0.1042	0.0845	0.0863	0.1393
P(Feature|Ham)				
	Order	Free	Credit	Money
Zero	0.9848	0.9247	0.9314	0.9071
Low	0.0018	0.0225	0.0248	0.0316
Med	0.0116	0.0352	0.0298	0.0455
High	0.0018	0.0176	0.014	0.0158

Figure 4



Questions 11 to 24 refer to Figure 4 above showing Loss vs. Model Complexity.

In the generic supervised machine learning problem,

$$\min_{\beta,\eta,\theta}\{\text{loss}(Y,F(X,\beta),\eta)+\lambda\,\text{penalty}(\beta,\theta)\},$$

the second term in the minimization, $\text{penalty}(\beta,\theta)$, is a penalty function that penalizes model complexity. That is $\text{penalty}(\beta,\theta)$ is a non-negative function that is large when the model is complex and small when it is simple. The parameter λ ($\lambda \geq 0$) is a tuning parameter that controls how much of a penalty there is for model complexity. When $\lambda = 0$ there is no penalty for complexity. On the other hand, for large positive values of λ there is a heavy penalty for increasing model complexity.

Consider a family of five models that differ only in the value of the tuning parameter λ that is used. Referring to Figure 4, answer the following questions.

(11) Which model is most likely to correspond to the one with the smallest value of the regularization tuning parameter λ?

(a) Model A (b) Model B (c) Model C

(d) Model D (e) Model E

(12) Which model is most likely to correspond to the one with the largest value of the regularization tuning parameter λ?

(a) Model A (b) Model B (c) Model C

(d) Model D (e) Model E

(13) Which model is most likely to have the smallest value of the penalty function penalty(β,θ)?

(a) Model A (b) Model B (c) Model C

(d) Model D (e) Model E

(14) Which model is most likely to be the best model among those considered?

(a) Model A (b) Model B (c) Model C

(d) Model D (e) Model E