Advanced Machine Learning

Mid Term Exam, Fall 2023
Total Marks = 40.
Time: 1 hr 15 mins.

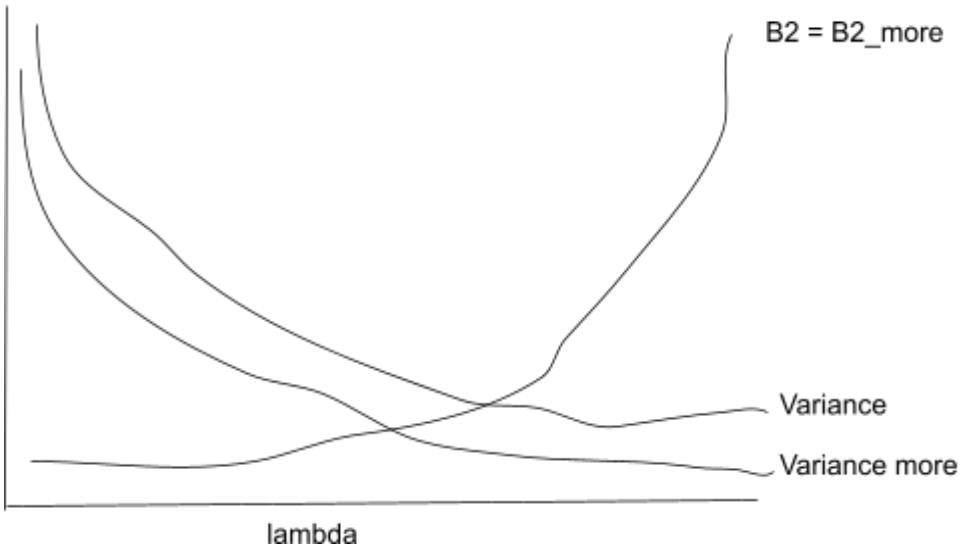
Answer each problem in the space provided; use the back of the preceding sheet in extreme conditions only.
Declaration: By submitting this examination for grading, I affirm that I have neither given nor received help from another examinee.

Name: _____ EID: _____

Signature: _____

Q1. [10 pts total]
(a) (1 +1 pts) Mention two situations where you may prefer to use stochastic gradient descent (SGD) to determine the parameters of a multiple linear regression model instead of solving the (batch) least squares problem?
Any two of the following will suffice.
(i) Large dataset. Full-batch is too expensive.
(ii) Non-stationary problem, relationship between x and y changes with time, so need to adapt online.
(iii) Data is not available all at once but is streaming.

(b) (2+2+1 = 5 pts) (a) Qualitatively plot how (square of) model bias (B2) and model variance (V) should behave as a function of the amount of regularization (lambda) for a ridge-regularized linear regression model.
(b) For the same problem, also plot B2_more and V_more which represent how these two terms will qualitatively change for the same problem if you use more training data (validation set size remains the same).
You should show both curves on the same plot of Bias/variance versus lambda so one can compare them.



(c) What will the MSE be of a regularized regression model if lambda = infinity?

Variance(y)

(c) (2+1 pts).
(i) When training an MLP using SGD, how do you decide how many epochs should the network be trained for?

Look at the validation curve

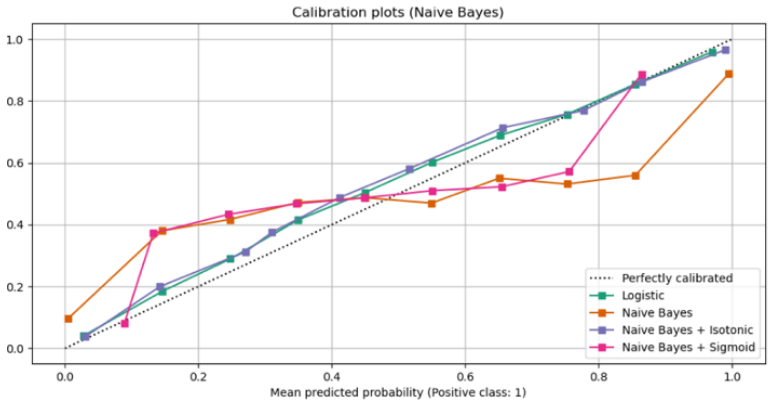
(ii) For the same problem, how do you expect the number of training epochs determined in part (i) to change if the size of the training dataset is substantially increased? Circle one: **Increase. Decrease. Remain about the same.**

Q2. [10 pts total]

(a) (1+2 pts). What is Huber loss? Explain how linear regression using the Huber loss function can be viewed as solving for a weighted version of standard linear regression that is based on minimizing the "(weighted) sum of squared errors" loss function. (with the appropriate choice of weights, the solutions will be identical).

Linear regression uses squared loss (r2, where residual r = y – y') for all data points, whereas huber loss uses squared error function for smaller residuals, and linear error function (|r|) for larger residuals.
So we can view linear regression with huber loss as a weighted version of standard linear regression, where weights for the smaller residuals are 1, and weights for the larger residuals are proportional to 1/|r| so that the error function becomes linear.

(b) (2+2 pts). A calibration plot is shown below, based on evaluating 4 classifiers on a validation dataset.
(i) What is the purpose of such a plot? Name one application where having a well-calibrated classifier is desired.
(ii) What is the Y-axis for such a plot?



i) Calibration curve evaluates how well a classification is calibrated, how the predicted probability given by the model differs from the observed probability. Any critical application like healthcare, finance, etc.
ii) fraction of positives in that bin.

(c) (3 pts).
Suppose that due to asymmetric costs, you have decided that you will call an input as belonging to Class 1 (and not the reference class 0) if your (estimated) P(C1|x) >= 0.7. Show that the decision boundary obtained will be a hyperplane (and specify the algebraic equation that defines this hyperplane) if you are using logistic regression to solve this problem.

P(C1|x) = log(mu/(1-mu)) = w*x = log(0.7/0.3)
which is linear, i.e., w*x > log(0.7/0.3) predict C1 else predict C0

Advanced Machine Learning

Q3. [10 pts total]
(a) ((1+2)+2 pts)
Consider the loss matrix below specified for a certain 2-class problem:

		Decision		
		C1	C2	Reject (for part (ii))
Truth	C1	-1	4	1
	C2	3	0	1

(i) Suppose there is no "Reject" option, i.e. you have to label each datapoint as belonging to either C1 or C2. For what range of values of P(C1 | x) will you classify the input as belonging to Class C1 if:
(a) Your goal is to minimize the expected error rate?

To classify as C1, $P(C1|x) > P(C2|x)$
$\Rightarrow P(C1|x) > 1 - P(C1|x)$
$\Rightarrow P(C1|x) > 0.5$.

(b) Your goal is to minimize the expected loss?
loss of C1: $(-1)*P(C1|x) + 3*(1 - P(C1|x))$
loss of C2: $4*P(C1|x) + 0*(1-P(C1|x))$

To go for the C1, we let loss of C1 < loss of C2 $\Rightarrow P(C1|x) > 3/8$

(ii) (a) Suppose the Reject option is also available with a fixed cost of 1 per data point rejected, as shown in the table above. For what range of values of P(C1 | x) will you go for the **reject** option if your goal is to minimize the expected loss?
Loss for Reject: $1*P(C1|x) + 1*(1-P(C1|x)) = 1$
Loss for C1: $(-1)*P(C1|x) + 3*(1 - P(C1|x))$
Loss for C2: $4*P(C1|x)$

To go for Reject option, we let loss of Reject<loss of C1, AND loss of Reject<loss of C2:
$1 < (-1)*P(C1|x) + 3*(1 - P(C1|x))$ and $1 < 4*P(C1|x)$
$\Rightarrow ¼ < P(C1|x) < ½$

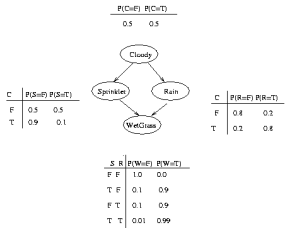
(b) If the fixed cost of the Reject option was "c", then what will be the minimum value of c for which you will never exercise the Reject option?

Ans.
To exercise the Reject option,
$c < (-1)*P(C1|x) + 3*(1 - P(C1|x))$ AND $c < 4*P(C1|x)$
$\Rightarrow P(C1|x) < (3-c)/4$ AND $P(C1|x) > c/4$

To not exercise the Reject option,
$(3-c)/4 <= c/4$
$\Rightarrow c >= 1.5$ (cost at $P(C1|x) = ⅜$)

(c) (2+3 pts). Consider the Bayesian belief network below with the associated conditional probability tables (CPT)s.
(i) Why is the Cloudy variable not present in the CPT for "WetGrass"?
A node is independent of its ancestors given its parents. There is no direct arc from Cloudy variable to WetGrass, so its not a parent. W is independent of C given its parents S and R.

(ii) What is the probability that a day was cloudy, sprinkler was ON and it did not rain, given that the grass was not wet? (Hint: first write the algebraic equation before trying to fill in the probability values. You should be able to at least get an arithmetic expression, if not solve for it, without a calculator).
P(C=1, S=1, R=0 | W=0) = P(C=1, S=1, R=0, W=0)/ P(W=0)
= P(C=1)*P(S=1|C=1)*P(R=0|C=1)*P(W=)|S=1, R=0) / P(W=0)
= 0.5*0.1*0.2*0.1/P(W=0)



Q4. [2x5 = 10 pts total] Answer any 5 of the 6 questions below.
(a) What is the key difference between transfer learning and multi-task learning?

Transfer learning involves training on the target/task domains sequentially, while in multi-task learning, we train multiple tasks simultaneously.

(b) How is the "scree plot" helpful in indicating whether PCA may be a suitable choice for dimensionality reduction for a given dataset?
The scree plot indicates the variances along principal components. If there is a significant difference in variance between top principal components and the rest, then PCA is suitable.

(c) Briefly describe what you understand by the "softmax" operation. What is its purpose?
It does a monotonic transformation of a set of numbers x->t(x). All t(x) are non-negative, sum to one, potential for interpretation as discrete probabilities. Softmax operation is often used to normalize the output of a network to a probability distribution over predicted output classes.

(d) Briefly, what is an "AI foundation model"?

A type of machine learning model pretrained on a large scale dataset which can then be finetuned for a variety of downstream tasks.

(e) How do you reconstruct (a noisy version of) the original data from the eigenvectors and the scores obtained through PCA?

The original data is reconstructed by multiplying the selected eigenvectors by their corresponding scores and summing the products.

(f) While solving a 3 class problem using logistic regression, you obtained the parameter vectors $\boldsymbol{\beta}_1$ (that distinguishes class 1 from class 0) and $\boldsymbol{\beta}_2$ (that distinguishes class 2 from class 0). Suppose instead you chose class 2 to be the reference class. What will be the obtained parameter vector when solving class 0 vs. class 2 binary classification problem using the same dataset?

$-\boldsymbol{\beta}_2$

.

Space for doodle, jokes, etc.