MIS 382N

Mid Term Exam, Fall 2019

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: ________________________________

Signature: ________________________________

Q. 1: (10 pts total)

(a) (2+2=4 pts) (i) Explain what you understand by the "model bias" of a parametric regression model y=f(**x;w**) (where the parameters are denoted by **w**), learnt using a training dataset of size |D|, either in English or by writing a mathematical expression.

Bias represents the extent to which the average prediction over all datasets differs from the desired regression function. It is the error caused due to assumptions built into the model.

$$\int (\mathbb{E}_{\mathcal{D}}[f(\mathbf{x};\mathcal{D})] - h(\mathbf{x}))^2 p(x)dx$$

(Expectation is over all datasets of same size.)

(ii) What will happen to model bias for a Multi-layered perceptron, when
The size of a separate test set is increased (all other factors are kept the same)?**Circle one: increase decrease [✓ no change]**
The number of hidden units is increased (all other factors are kept the same)? **Circle one: increase [✓ decrease] no change**

(b) (6 pts) Suppose data is being obtained from the following generative process:
(i) Draw x from a uniform distribution between -5 and 5.
(ii) Given x, the corresponding t is obtained as t = q(x) + ε ;
where q(x) is an unknown fifth order polynomial in x, and ε is drawn from a distribution with zero mean, finite variance.

You suspect that a polynomial will be a good regression function to fit to the data, but do not know what degree is suitable. So you fit polynomials of different degrees (from 1 through 9) to the data using OLS. *Qualitatively* sketch (model_bias)^2 (B), model variance (V) and true MSE (E), as functions of model complexity, given training data sets of size 20 and size 100 respectively. (you should show 6 curves on the space given below, labelled B20, B100, V20, V100, E20 and E100, respectively.

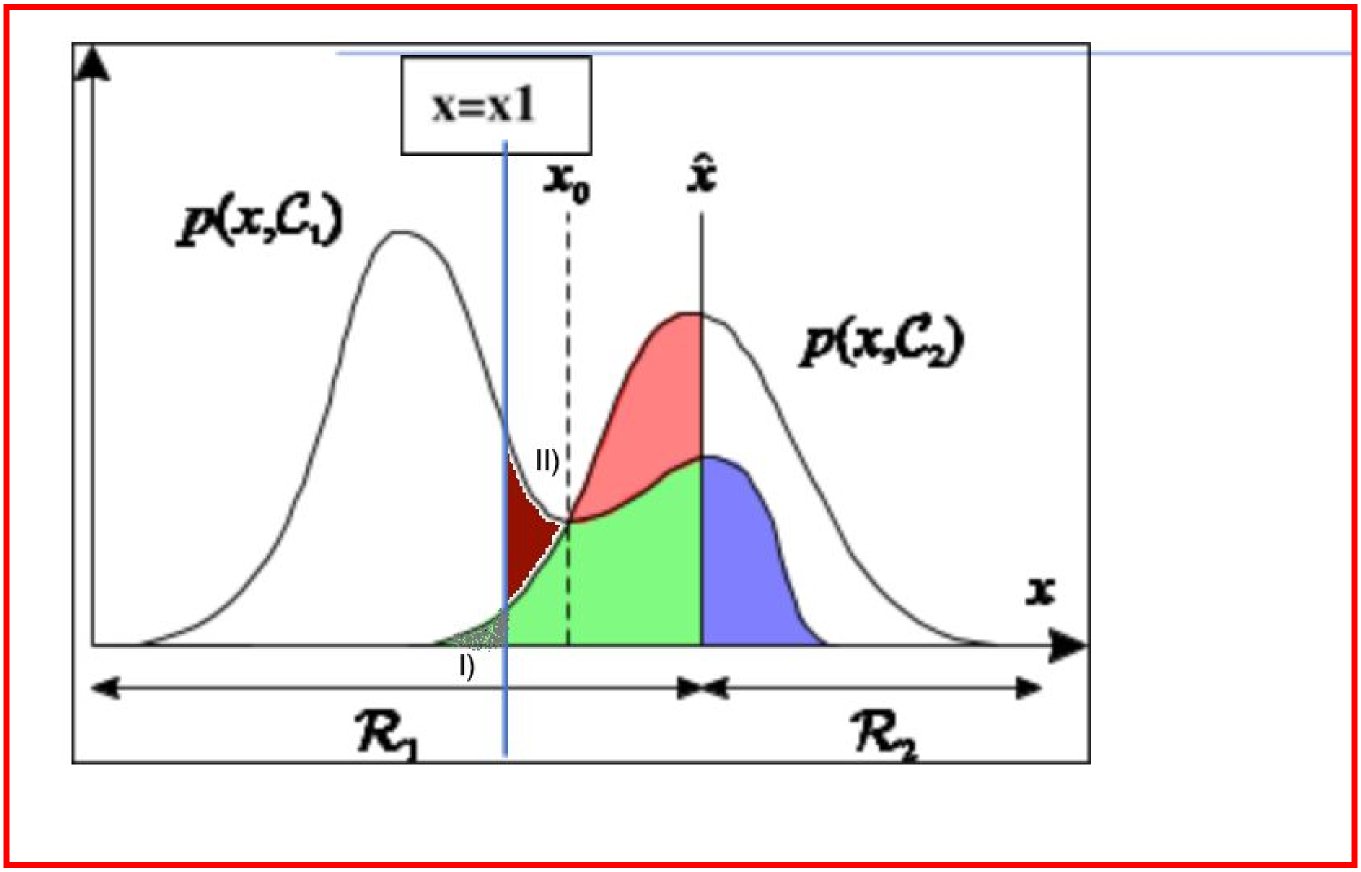
- B20=B100: goes down, zeros at 5 and flattens
- V20 > V100, both rise with complexity.
- E20 > E100, minima of E100 lower and to the right of (or same place as) E20 min, both min should be at or to the left of 5.

Q2 (11 points total)

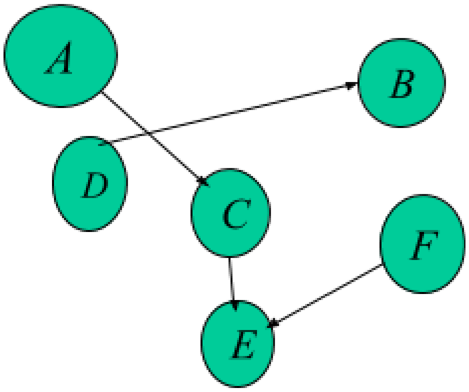
(a) **(2+2 pts).** A figure from the notes related to a 2-class problem given a single independent variable, x, is reproduced below. Suppose your classifier labels all objects with "x" < x1 as belonging to Class 1, and labels all others as Class 2.

I have drawn a solid vertical line to the left of the dotted vertical line to denote x=x1)

Indicate by shading the appropriate regions (i) the probability that a point belonging to Class 2 is misclassified. (ii) the Error in excess of the Bayes error rate that your classifier incurs

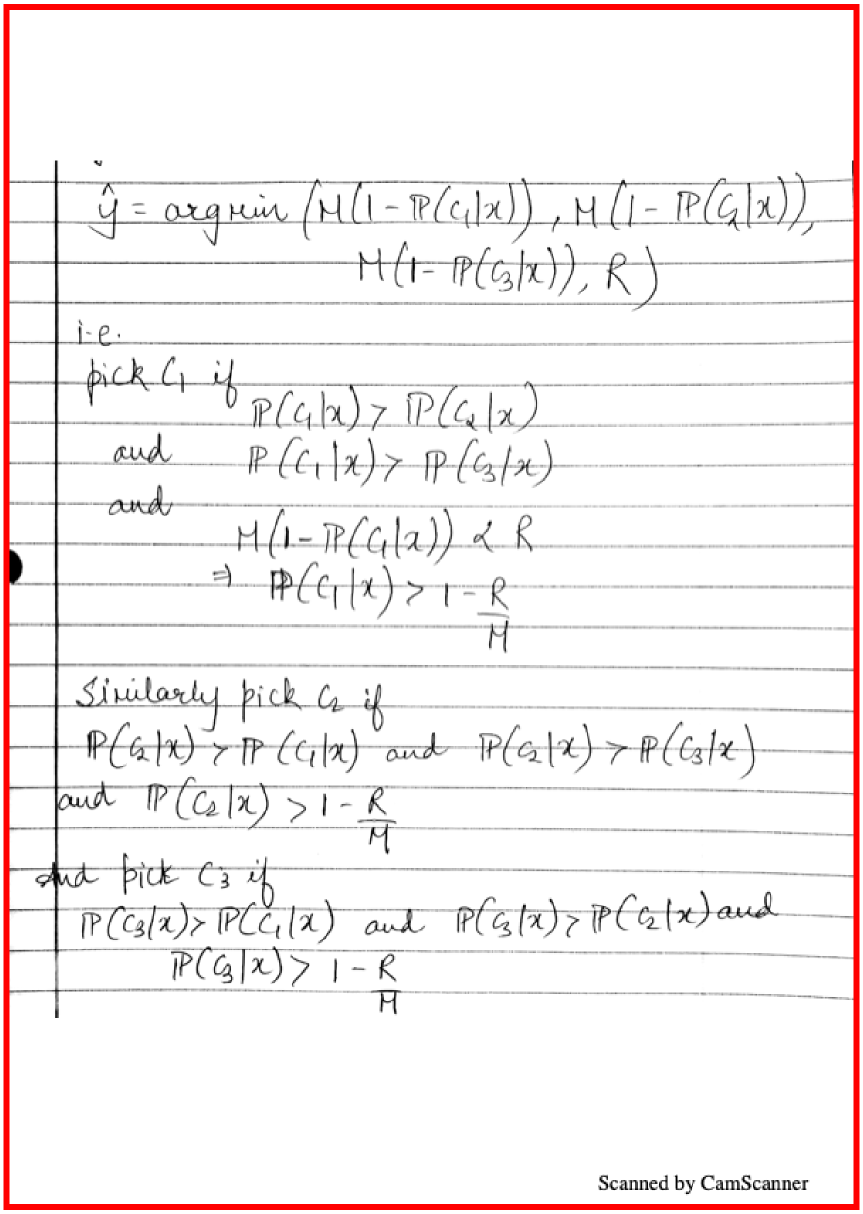


(b) **(3 pts)** Write down an expression for the joint distribution of the six random variables, (A through F) using the Bayesian network shown below to simplify the expression wherever possible.



$P(A)\,P(D)\,P(F)\,P(B|D)\,P(C|A)\,P(E|C)\,P(E|F)$

(c) (4 pts). Consider a 3 class problem. The cost of any correct classification is 0, and cost of any single misclassification is M. Suppose that you also have a *reject option*, i.e., if you are not sure of the correct class, you have the option of making a “not known” call, which will incur a cost of R, where 0 < R < M. If you have a reasonable way of estimating the posterior probabilities of all the classes for any input **x**, what should your optimum decision be so as to minimize the expected cost of your decision?



$$\hat{y} = \text{argmin}\left(M(1-\mathbb{P}(C_1|x)),\ M(1-\mathbb{P}(C_2|x)),\ M(1-\mathbb{P}(C_3|x)),\ R\right)$$

i.e.

pick C_1 if $\mathbb{P}(C_1|x) > \mathbb{P}(C_2|x)$

and $\mathbb{P}(C_1|x) > \mathbb{P}(C_3|x)$

and $M(1-\mathbb{P}(C_1|x)) < R$

$\Rightarrow \mathbb{P}(C_1|x) > 1 - \frac{R}{M}$

Similarly pick C_2 if

$\mathbb{P}(C_2|x) > \mathbb{P}(C_1|x)$ and $\mathbb{P}(C_2|x) > \mathbb{P}(C_3|x)$

and $\mathbb{P}(C_2|x) > 1 - \frac{R}{M}$

and pick C_3 if

$\mathbb{P}(C_3|x) > \mathbb{P}(C_1|x)$ and $\mathbb{P}(C_3|x) > \mathbb{P}(C_2|x)$ and

$\mathbb{P}(C_3|x) > 1 - \frac{R}{M}$

Q 3 [2 pts each: Total 10 pts] PICK ANY 5.

(i) What is the Huber loss function and in what situation(s) may it serve as a more useful loss function for regression models?

See notes: The Huber loss function is more robust to outliers since it penalizes outliers with less error thus influencing our model less during training due to errors on outliers.

(ii) 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?

Ans:

Any two of the following will suffice.

(i)Large dataset. Batch 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

(iii) When SGD is applied to a regression problem, suppose your model (with parameters or weights **w**(t)) at some time t is realizing a certain function $h_t(x, w)$. Of what function is the gradient taken, and what variable(s) is the gradient "with respect to"?

Function: Loss function of h_t(x, w) versus the true value at time t.
The gradient will be taken with respect to the weights.

(iv) Briefly, what do you understand by a classification model being well calibrated?

Well calibrated classifiers are probabilistic classifiers for which the predicted probabilities can be directly interpreted as a confidence level. They well estimate the true posterior probabilities.

(v) How will you formulate a 3-class classification problem so that you can solve it using logistic regression?

This can be done by running 3-1 = 2 independent binary logistic regression models, in which one class is chosen as a "pivot" and then the other 2 classes are separately regressed against the pivot class.

(vi) While using Naïve Bayes, it may happen that there is no record in the training data that has a specific value of a categorical variable for a certain class. Why is this a problem and how can it be mitigated?

This is a problem since the probability will drop to 0. This can be mitigated by Laplace smoothing.

$$P\left(A_i = v_j | c_k\right) = \frac{n_{ijk} + 1}{n_k + s_i}$$

here,

n_{ijk} is the number of examples in c_k where $A_i = v_j$. n_k is the number of examples in c_k. s_i is the number of possible values for A_i.

Q 4 [1.5x6=9 pts total] Choose the best alternative. **If multiple answers are correct, select all the correct ones.** 1.5 for fully correct answer(s), -0.5 for each incorrect answer; 0 for no answer.

(a) The decision boundary obtained by a logistic regression model (for a 2-class problem) is

(i) linear in the original feature space (i.e., the space of the independent variables corresponding to the coefficients of the model)
(ii) linear in the space of log-odds transformed features
(iii) non-linear in general to cater to curved decision boundaries
(iv) linear in an implicit projection to infinite dimensions.

(b) When linear regression is applied using OLS, we are assuming that:

(i) the noise (in the dependent variable) is Normally distributed with constant variance
(ii) the noise is zero mean
(iii) both dependent and independent variables have zero-mean Gaussian noise added
(iv) All the independent variables are uncorrelated with one another.

(c) In PCA, the first principal component indicates:

(i) the direction of maximum separation between classes
(ii) the direction in which the projected data has maximum variance
(iii) the direction that maximizes the R-squared value if the extracted feature is used in linear regression

(d) The (two-class) LDA (Linear Discriminant Analysis) classifier

(i) Is a generalization of the QDA classifier
(ii) Assumes that both classes are normally distributed with the same covariance.
(iii) Provides Bayes optimal decisions if adequate training data is available
(iv) Assumes that the decision boundary is linear, and so is ideal if the classes are actually linearly separable

(e) The Bayes Decision Rule

(i) proves the optimality of the Naïve Bayes Classifier
(ii) Maximizes the Area Under the ROC curve for a two-class problem
(iii) Minimizes the expected probability of misclassification
(iv) Maximizes the accuracy on a given test set

(f) Stochastic Gradient Descent (SGD):

(i) can give different results for same training data if the order in which the data-points are stored is changed
(ii) can be used to update the weights of a multi-layered perceptron with two hidden layers
(iii) can also be applied to cost functions other than "squared loss"
(iv) Is called stochastic because it does gradient descent on the mean squared error

Space for a doodle or a joke..