Indian Institute of Technology Kharagpur
Mid-Spring Semester 2018-19

Subject No.: CS60050 Subject: Machine Learning
Department: Computer Science and Engineering
Duration: 2 hours Full Marks: 50
Specific charts, graph paper, log book etc., required: NA
Instructions:
This question paper contains three (03) printed pages.
Attempt all questions. All parts of the same question must be answered together.
No clarifications can be provided during the exam. Make reasonable assumptions if necessary, and state any assumptions made.
All workings must be shown. You can use calculators.

1. State whether the following statements are True or False, with brief reasons. **No credit will be given if the explanation is incorrect even if the answer is correct.** $[2 \times 6 = 12]$

   (i) Growing a decision tree to full depth makes it more likely to fit the noise in the data.

   (ii) In a least-squares linear regression problem, adding regularization is likely to decrease the error of the solution on the training data.

   (iii) If you train a linear regression estimator with only half the data, its bias is smaller.

   (iv) A classifier that attains 100% accuracy on the training set is definitely better than a classifier that attains 80% accuracy on the training set.

   (v) For logistic regression, with parameters optimized using a stochastic gradient method, setting parameters to 0 is an acceptable initialization.

   (vi) Suppose you are given m data points, and you use half the points for training and half for testing. The difference between training error and test error is likely to increase as m increases.

2. Answer the following Multiple Choice Questions. Each question may have any number of correct answers, including zero. List all choices that you believe to be correct. **There is 1 mark for each correct answer, -0.5 for each incorrect answer.**

   (i) A given dataset can contain noise objects and outliers. Which of the following statements are true?
   - (a) A noise object can be an outlier
   - (b) An outlier can never be a noise object
   - (c) An outlier can be a noise object
   - (d) Noise objects and outliers should be removed before training a classifier

   (ii) As model complexity increases, which of the following statements are true?
   - (a) The variance is likely to increase
   - (b) The variance is likely to decrease
   - (c) The bias is likely to increase
   - (d) The bias is likely to decrease

(iii) Suppose you suspect that your model is overfitting. Which of the following is a valid way to try and reduce the overfitting?
(a) Increase the amount of test data
(b) Improve the optimization algorithm being used for error minimization
(c) Decrease the model complexity
(d) Using regularization on the model paremeters

(iv) You have trained a decision tree for spam mail classification, and it is getting abnormally bad performance on both your training and test sets. You know that your implementation has no bugs, so what could be causing the problem?
(a) Your decision tree is too shallow
(b) Your classifier is overfitting
(c) You need to change the optimization algorithm
(d) All of the above

(v) Suppose you have a regularized linear regression model. What is the effect of increasing the regularization parameter λ on bias and variance?
(a) Increases bias, increases variance
(b) Increases bias, decreases variance
(c) Decreases bias, increases variance
(d) Decreases bias, decreases variance

3. Consider the dataset in Table 1 showing various attributes of a set of herbs, that can be used to classify a herb as poisonous or not. It is known whether herbs A through H are poisonous (training set), but nothing is known about U through W. **[2 + 4 + 6 + 3 = 15]**
(a) What is the entropy of Poisonous?
(b) Which attribute should be chosen as the root of a decision tree, considering *information gain* as the measure of impurity?
(c) Build a decision tree to classify herbs as poisonous or not.
(d) Classify herbs U, V and W using this decision tree as poisonous or non-poisonous.

Training

Example	Heavy	Smelly	Spotted	Smooth	Poisonous
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
U	1	1	1	1	?
V	0	1	0	1	?
W	1	1	0	0	?

Table 1: Dataset for Question 3

4. Consider Table 2 that reports, for eight instances, the actual class of each instance and the score given by a classifier to that instance. For a given threshold τ, if the score is equal to or greater than τ, the classifier assigns + class, else it assigns − class. Considering the thresholds (i) $\tau = 5$, (ii) $\tau = 1$, (iii) $\tau = -3$, compute Precision, Recall and False Positive Rate (FPR) of the classifier for the + class. [7]

$(7, \tau)$ +

Instance	Actual class	Score
1	+	7
2	+	4
3	−	2
4	−	1
5	−	−1
6	+	−4
7	−	−5
8	−	−6

Table 2: Dataset for Question 4

5. In a supervised learning problem, let the hypothesis be $H(x) = w_0 + \sum_{j=1}^{d} w_j x_j$, where $x_j, j = 1, 2, \ldots, d$ are the features. Let the cost function for a data point be $C(H(x), y) = \frac{1}{1+\exp(yH(x))}$. Suppose you use gradient descent to obtain the optimal parameters w_0 and $w_j, j \to 1, 2, \ldots, d$. Derive the update rules for the parameters. [8]

$$\frac{d}{dx} \text{of} \frac{1}{1+e^{ax}}$$

$$\frac{-1}{(1+e^{ax})^2} \frac{d}{dx}(1+e^{ax})$$

$$\frac{-e^{ax}}{(1+e^{ax})^2} \frac{d}{dx}(ax)$$

$$\left(\frac{1}{1+e^{ax}}\right)\left(\frac{1}{1+e^{ax}} - 1\right)$$

(1 i). A decision tree is a classifier that develops decision boundaries by building of a tree like Model whereby each path from Root → Leaf is a Conjunction of Attributes.

2

If a decision tree is build to its full depth → all the leaves will be homogenous with Zero Entropy and the Training Data will be perfectly classified.

However this is very akin to the Overfitting Problem in general. Each Data Points classification will lead to NOISE getting fit too.

Hence TRUE.

(1).

(ii) FALSE

2 Regularisation is done to tackle the ~~Information~~ Overfitting problem which arises due to the learned model fitting the idiosynchrosies of the training data.
Regularisation tones down the co-efficients of the parameter vector to reduce overfitting by turning the hypothesis into a simpler one from a complex one.
If at all, it might lead to a slight increase in the Training Error since the overfit model is being changed.

(iii) 2 The bias is a property of the algorithm or model we aim to learn on the Hypothesis Space. The Restrictions we put on the Hypothesis Function and the preference of the family of functions is what comprises of bias.
Training Linear Regressor with half data would not lead to a change in the bias as such, will definitely not make it smaller. FALSE

(iv) FALSE

2 There is nothing that can be said with 100%. Certainty and conviction in Machine Learning. The "better" of a classifier is inferred by its performance on the test-data, not the training data. This is because, measuring error on the data-set which was fed to it, which it was modelled to fit sounds absurd.

The true generalisation error of the 80% classifier might actually be better than the 100% classifier. Nothing can be said about that, but one thing can be said for sure that the Performance on the Training Data is not a Metric of Evaluation.

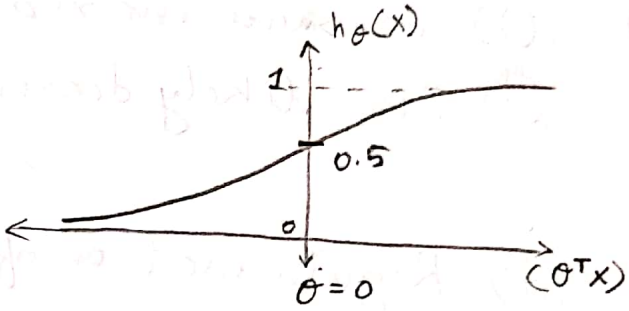
(vi). FALSE

2 $\frac{m}{2}$ points for Training and Testing.

as the value of 'm' → the number of data points increases, the properties and features of the target function will be met/realised better as opposed to a smaller dataset. Hence the Training & Testing Errors will both converge to the

generalisation Error; and the difference b/w them would actually decrease.

(v) $h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$



if initially $\theta = 0$

$$h_\theta(x) = \frac{1}{1+1} = \frac{1}{2} = 0.5.$$

Stochastic descent:

```
while( NOT CONVERGENT)
{   for j = 1 to m
    {
        θ_i := θ_i - α( h_θ(x^j) - y^j ) . x_i^j
        ∀ i ∈ [0 .n] Simultaneously.
    }
}.
```

TRUE; there seems to be no problem whatsoever; the cost function is a CONVEX function and any point will converge to the Global Minima using Gradient Descent with suitable α.

2

(ii) (a) Variance increases
(d) Bias likely decrease

(iii). (d) Regularisation of Model Parameters
(c) Decrease Model Complexity
~~(b) Increase/Add Data~~

(iv) (a) Decision Tree is too shallow.

(v) (b) Increases bias, decreases variance

(i) (a) Noise object can be outlier
(c) outlier can be a noise object.

3 $m = 8$

(a) Poisonous ≡ PO = { 0, 1 }

2 $P(PO=0) = \frac{3}{8}$

$P(PO=1) = \frac{5}{8}$

$$\text{Entropy}(PO) = -\left(\frac{3}{8}\log_2\left(\frac{3}{8}\right) + \frac{5}{8}\log_2\left(\frac{5}{8}\right)\right)$$

$$= -\left(-0.530639 + (-0.423794)\right)$$

$$\text{Entropy}(PO) = 0.954434$$

(b). 4 Split on Heavy

Heavy	n(PO=0)	n(PO=1)	I(PO|H) Entropy
0	2	3	0.970950
1	1	2	0.918296

$-\left(\frac{2}{5}\log\left(\frac{2}{5}\right) + \frac{3}{5}\log\left(\frac{3}{5}\right)\right) = 0.970950$

$-\left(\frac{1}{3}\log\left(\frac{1}{3}\right) + \frac{2}{3}\log\left(\frac{2}{3}\right)\right) = 0.918296.$

$$\text{Total Entropy} = \frac{5}{8}(0.970950) + \frac{3}{8}(0.918296)$$

$$= 0.951208$$

Split on Smelly

Smelly	n(PO=0)	n(PO=1)	I(PO|Sm)
0	2	3	0.970950
1	1	2	0.918296

Same as Heavy → 0.951208

Split on Spotted

Spotted	n(PO=0)	n(PO=1)	I(PO|Sp)
0	2	3	0.970950
1	1	2	0.918296

Same as Heavy → 0.951208

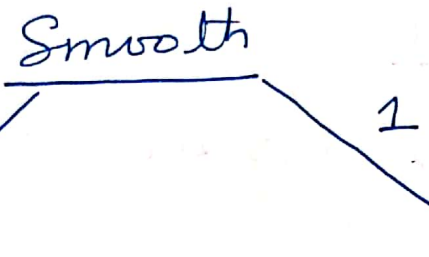
Split on Smooth

Smooth	n(PO=0)	n(PO=1)	I(PO|Smt)
0	2	2	1
1	1	3	0.811278

$$-\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right)$$

Total Entropy = $\frac{1}{2}(1 + 0.811278) = \underline{0.905639}$

Since Entropy is Least for Split on Smooth, Information Gain is highest (E(P0) - E(A))
Hence, we split on Smooth ✓



	H	Smelly	Sp	P0
A	0	0	0	0
B	0	0	1	0
E	0	1	1	1
H	1	1	0	1

	H	Smelly	Sp	P0
C	1	1	0	0
D	1	0	0	1
F	0	0	1	1
G	0	0	0	1

lets Split on Smelly

Smelly	n(P0=0)	n(P0=1)
0	2	0
1	0	2

Perfect classification

No further check Reqd.

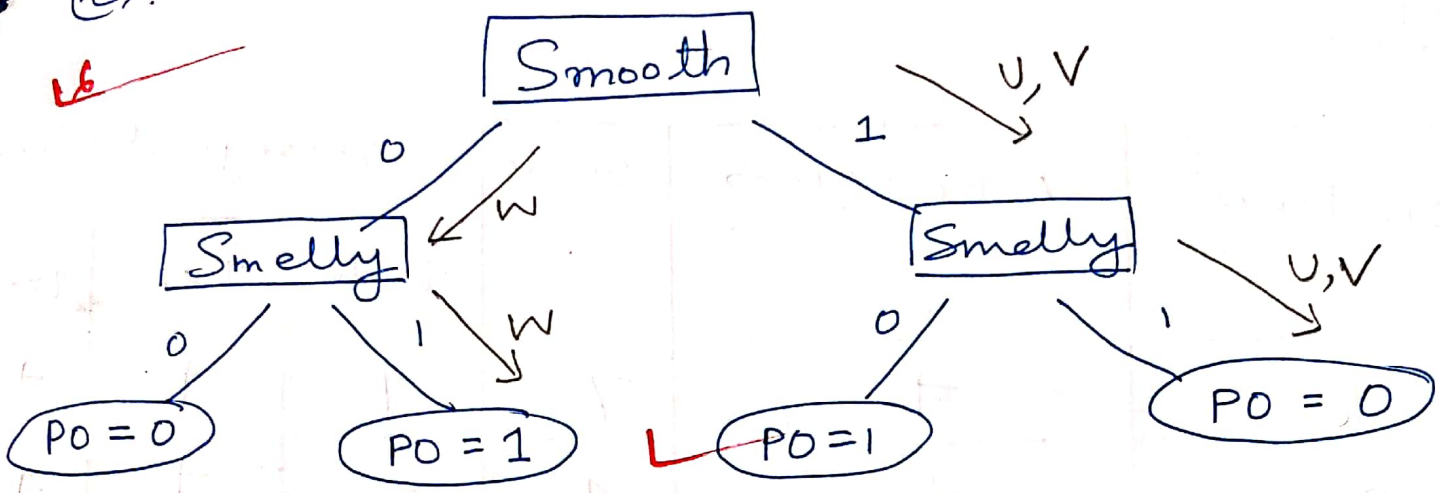
lets Split on Smelly

Smelly	n(P0=0)	n(P0=1)
0	0	3
1	1	0

Perfect Classification

No Check Required.

(c).



This is our decision Tree.

(d).

	Smooth	Smelly	Poisonous
U	1	1	0
V	1	1	0
W	0	1	1

Poisonous = Smooth ⊕ Smelly.

(4).

Instance	Actual Class	Score	Assigned Class		
			$\tau = 5$	$\tau = 1$	$\tau = -3$
1	+	7	+	+	+
2	+	4	−	+	+
3	−	2	−	+	+
4	−	1	−	+	+
5	−	−1	−	−	+
6	+	−4	−	−	−
7	−	−5	−	−	−
8	−	−6	−	−	−

$\tau = 5$.

h \ y	1	0
1	1 (TP)	0 (FP)
0	2 (FN)	5 (TN)

$$\text{Precision} = \left(\frac{TP}{TP + FP}\right) = \frac{1}{1+0} = 1$$ ✓

$$FPR = \frac{0}{8} = 0$$ ✓

$\left(\frac{FP}{Total}\right)$

$$\text{Recall} = \left(\frac{TP}{TP + FN}\right) = \frac{1}{1+2} = \frac{1}{3}$$ ✓

$T = 1$

h \ y	1	0
1	2 (TP)	2 (FP)
0	1 (FN)	3 (TN)

$$\text{Precision} = \frac{2}{2+2} = \frac{1}{2}$$ ✓

$$\text{Recall} = \frac{2}{2+1} = \frac{2}{3}$$ ✓

$$\text{FPR} = \frac{2}{8} = \frac{1}{4}$$ ✗

$T = -3$

h \ y	1	0
1	2 (TP)	3 (FP)
0	1 (FN)	2 (TN)

$$\text{Precision} = \frac{2}{2+3} = \frac{2}{5}$$ ✓

$$\text{Recall} = \frac{2}{2+1} = \frac{2}{3}$$ ✓

$$\text{FPR} = \frac{3}{8} = \frac{3}{8}$$ ✗

	T	Precision	Recall	FPR
(i)	5	1	$\frac{1}{3}$	0
(ii)	1	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{4}$
(iii)	-3	$\frac{2}{5}$	$\frac{2}{3}$	$\frac{3}{8}$

(5) d features. $\rightarrow x_j \in \{1, 2 \cdots d\}$

$$H(x) = w_0 + \sum_{j=1}^{d} w_j x_j$$

defining $x_0 = 1$

$$\boxed{H(x) = \sum_{j=0}^{d} w_j x_j}$$

Cost Function

$$c(H(x), y) = \frac{1}{1 + e^{yH(x)}}$$

Let $w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}$ and $X = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}$ and $y = \begin{bmatrix} y_0 \\ \vdots \\ y_d \end{bmatrix}$

Then

$$H(x) = H_w(X) = w^T X$$

$$c(H_w(X), y) = \frac{1}{1 + \exp(y w^T X)}$$

$$\nabla_w C = ?$$

$$\frac{\partial C}{\partial w_i} = \frac{\partial}{\partial w_i}\left(\frac{1}{1+\exp(yw^Tx)}\right)$$

$$= \left(\frac{1}{1+\exp(yw^Tx)}\right) \times \left(\frac{1}{1+\exp(yw^Tx)} - 1\right) \times \frac{\partial}{\partial w_i}(yw^Tx)$$

$$\frac{\partial C}{\partial w_i} = (C) \times (C-1) \times y x_i$$

So;

$$w_i := w_i - \alpha \frac{\partial C}{\partial w_i}$$

or

$$w_i := w_i - \alpha \times \left(\frac{1}{1+\exp(yw^Tx)}\right)\left(\frac{1}{1+\exp(yw^Tx)} - 1\right) y x_i$$

is the update rule for w_i where $i \in \{0, 1, \ldots, d\}$

To be applied simultaneous to all the w_i

with x_i as $x_0 = 1, x_1, x_2 \ldots x_d$; $y \rightarrow$ given

and $w^Tx = H(x)$. and $\alpha \rightarrow$ Learning Rate

$$w_i := w_i - \alpha \times (Cost) \times (Cost - 1) \times y \cdot x_i$$

PTO

While (NOT converge)

{

$w_i := w_i - \alpha \times Cost \times (Cost - 1) \times y \times x_i$

⋮ all simultaneously for the $i \in \{0 \dots d\}$

}