CS188 Midterm

Koyoshi Shindo

TOTAL POINTS

50 / 50

QUESTION 1

1 ML basics 4 / 4

- **- 0 Correct**
- - 1 Supervised learning in part (a) missing/incorrect
- - 1 Unsupervised learning in part (a) missing/incorrect
- - 1 part (b) incorrect
- - 1 part (c) incorrect
- - 0.5 Supervised learning in part (a) incomplete/incorrect
- - 0.5 Unsupervised learning in part (a) incomplete/incorrect
- - 0.5 part of part (c) incorrect
- - 0.5 part (b) incomplete/incorrect

QUESTION 2

2 Applications 6 / 6

- **- 0 Correct**
- - 2 part a) incorrect
- - 2 part b) incorrect
- - 2 part c) incorrect
- - 1 part a) incomplete/missing
- - 1 part b) incomplete/missing
- - 1 part c) incomplete/missing

QUESTION 3

3 True/False 12 / 12

- **- 0 Correct**
- - 2 Q3 incorrect
- - 2 Q4 incorrect
- - 2 Q5 incorrect
- - 2 Q6 incorrect
- - 2 Q7 incorrect
- - 2 Q8 incorrect
- - 1 Q3 incomplete/missing
- - 1 Q4 incomplete/missing
- - 1 Q5 incomplete/missing
- - 1 Q6 incomplete/missing
- - 1 Q7 incomplete/missing
- - 1 Q8 incomplete/missing

QUESTION 4

4 Multiple Choice 7 / 7

- **- 0 Correct**
- - 2 problem 9 incorrect
- - 2 problem 10 incorrect
- - 1 11a incorrect
- - 1 11b incorrect
- - 1 11c incorrect

QUESTION 5

5 Maximum likelihood 5 / 5

- **- 0 Correct**
- - 1 Answer for specific dataset (a)
- - 2 Incorrect (a)
- - 2 Incorrect (b)
- - 1 Partially incorrect (a)
- - 1 Math Error (b)
- - 1 Incorrect (c)
- - 1 Partially incorrect (b)

QUESTION 6

6 Decision Trees 10 / 10

- **- 0 Correct**
- - 2 Incorrect decision tree
- - 1 Incorrect answer (a)
- - 4 Incorrect answer (b)
- - 1 Incorrect answer (c)
- - 2 Partially incorrect (b)
- - 1 Math Error (b)
- - 2 Incorrect answer (e)

QUESTION 7

7 Linear Regression 6 / 6

- **- 0 Correct**
- - 1 Missing solution theta1 = 0
- - 2 Incorrect 2nd possible solution
- - 1 Math Error (b)
- - 1 Incorrect derivative
- - 1 Wrong cost function
- - 1 Needs more simplification
- - 6 No answer
- - 1 Set derivative equal to 0

CS 188 — Machine Learning: Midterm

Winter 2017

Name: Koyoshi Shindo

UID: 004632397

Instructions:

1. This exam is CLOSED BOOK and CLOSED NOTES.
2. You may use scratch paper if needed.
3. The time limit for the exam is 1hour, 45 minutes.
4. Mark your answers ON THE EXAM ITSELF. If you make a mess, clearly indicate your final answer (box it).
5. For true/false questions, CIRCLE True OR False and provide a brief justification for full credit.
6. Unless otherwise instructed, for multiple-choice questions, CIRCLE ALL CORRECT CHOICES (in some cases, there may be more than one) and provide a brief justification if the question asks for one.
7. If you think something about a question is open to interpretation, feel free to ask the instructor or make a note on the exam.

Q	Problem	Points	Score
1	ML basics	4	
2	Application	6	
3	True/False	12	
4	Multiple choice	7	
5	Maximum likelihood	5	
6	Decision Trees	10	
7	Regression	6	
Total		50	

1. (4 pts) **Machine Learning Basics**

   (a) (2 pts) Consider supervised and unsupervised learning. What is the main difference in the inputs and the goals?

   Supervised learning's inputs have labels, and goal is to predict an unknown ~~st~~ instance's label. Unsupervised learning's inputs do not have labels, and the goal is typically grouping, clustering etc.

   (b) (1 pts) What is the main difference between classification and regression?

   Classification's label y is discrete/categorical. Regression's label y is real number.

   (c) (1 pts) Learning is about generalizing from training data to (unseen) test data. What does this assume about the training and test set?

   It assumes that they are related and rules learned from traing set can be applied on test set. Training set and test set should be produced under same condition, from same underlying distribution.

2. (6 pts) **Application**

   Suppose you are given a dataset of cellular images from patients with and without cancer.

   (a) (2 pts) Consider the models that we have discussed in lecture: decision trees, k-NN, logistic regression, perceptrons. If you are required to train a model that predicts the probability that the patient has cancer, which of these would you prefer, and why?

I will chose logistic repression, because only logistic repression gives the probability as hypothesis, and logistic repression is a binary classification algorithm

(b) (2 pts) (True/False) Suppose this dataset had 900 cancer-free images and 100 images from cancer patients. If you train a model which achieves 85% accuracy, it is a good model (Hint: think about a baseline).

False. It isn't a good model because a baseline can do 90% accuracy by guessing all is cancer-free (~~say~~ such as majority vote)

(c) (2 pts) (True/False) Suppose that you split your dataset into a training set and test set. A model that attains 100% accuracy on the training set and 70% accuracy on the test set is better than a model that attains 70% accuracy on the training set and 75% accuracy on the test set.

False. Training accuracy is a poor indication of how well a model does because memorization is possible. In fact, having too-high accuracy often means possible overfitting which does poorly on test set. So second model can be better.

True/False

3. (2 pts) (True/False) You are given a training dataset with attributes $A_1, \ldots, A_m$ and instances $x^{(1)}, \ldots, x^{(n)}$ and you use the ID3 algorithm to build a decision tree D_1. You then take one of the instances, add a copy of it to the training set, and rerun the decision tree learning algorithm (with the same random seed) to create D_2. D_1 and D_2 are necessarily identical decision trees.

False. The ~~copy~~ copied instance may alter ~~the~~ entropy which causes greedy choice to be different in some steps. For example it might make a tie into not a tie any more. of entropy

4. (2 pts) (True/False) Stochastic Gradient Descent is faster per iteration than Gradient Descent.

True. Stochastic is $O(D)$ Batch Gradient Descent is $O(ND)$

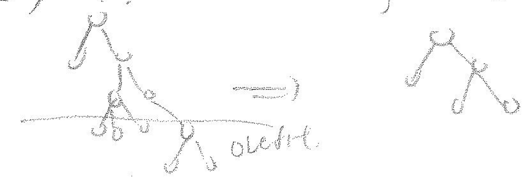
5. (2 pts) (True/False) You run the PerceptronTrain algorithm with $maxIter = 100$. The algorithm terminates at the end of 100 iterations with a classifier that attains a training error of 1%. This means that the training data is not linearly separable.

False. Not necessary. More iterations may reduce training error to 0.

6. (2 pts) (True/False) You learn a decision tree with the $MaxDepth$ parameter set to

infinity and then prune the resulting decision tree. Pruning the decision tree tends to reduce overfitting.

True. Pruning restricts the final depth of the tree, which tends to reduce overfitting.



7. (2 pts) (True/False) We want to use 1-NN to classify data into one of two classes. It is possible for 1-NN to always classify all new instances as positive even though there are negative instances in the training data. (If true, show an instance. If false, briefly explain.)

False. 1-NN looks at closest neighbor.
Say that the negative instance in training data is $\vec{X}_{neg}$
Then the new instance with feature vector $= \vec{X}_{neg}$
will be classified as negative

8. (2 pts) (True/False) You run gradient descent to minimize the function $f(x) = (2x-3)^2$. Assume the step size has been chosen appropriately and you run gradient descent till convergence. Then gradient descent will return the global minimum of f.

True. Because f is convex:
$f'' = 8 > 0$ for all x.

Multiple choice

9. (2 pts) In k-nearest neighbor classification, which of the following statements are true? (circle all that are correct)

   (a) The decision boundary is smoother with smaller values of k.
   (b) k-NN does not require any parameters to be learned in the training step (for a fixed value of k and a fixed distance function).
   (c) If we set k equal to the number of instances in the training data, k-NN will predict the same class for any input.
   (d) For larger values of k, it is more likely that the classifier will overfit than underfit.

   overfit = memorization

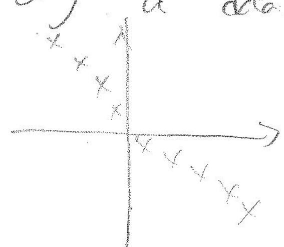
10. (2 pts) Assume we are given a set of one-dimensional inputs and their corresponding output (that is, a set of $\{x, y\}$ pairs). Further assume that we have an unlimited amount of data. We would like to compare the following two models on our input dataset:

$$A : y = \theta^2 x \qquad \Rightarrow \theta \text{ restricted to be } \geq 0$$
$$B : y = \theta x$$

For each one, we split into training and testing set to evaluate the learned model. Which of the following is correct? Choose the answer that best describes the outcome, and provide justification.

   (a) There are datasets for which A would perform *better* than B.
   (b) There are datasets for which B would perform *better* than A.
   (c) Both (i) and (ii) are correct.
   (d) They would perform equally well on all datasets.

(b) Say a dataset looks like this:



which is generated by $y = -x$
B can perform it, $\theta = -1$
but A ~~cannot~~ does poorly because $\theta^2 \geq 0$

11. (3 pts) If your model is overfitting, increasing the training set size (by drawing more instances from the underlying distribution) will tend to result in which of the following? (circle the best answer for each)

   (a) training error will ... increase / decrease / unknown
   (b) test error will ... increase / decrease / unknown
   (c) overfitting will ... increase / decrease / unknown

For these problems, you must show your work to receive credit! Blank pages have been provided for this purpose, or you may attach extra pages as needed.
(If you use additional pages, please indicate clearly the problem being solved and write your name and UID on each page.)

Maximum likelihood

12. We observe the following data consisting of four independent random variables $X_n, n \in \{1, \ldots, 4\}$ drawn from the same Bernoulli distribution with parameter θ (*i.e.*, $P(X_n = 1) = \theta$): $(1, 1, 0, 1)$.

   (a) Give an expression for the log likelihood $l(\theta)$ as a function of θ given this specific dataset. [2 pts]

$$L(\theta) = \prod P(X_n = x_n)$$
$$= \theta^3 (1-\theta)$$
$$l(\theta) = \log(L(\theta)) = \log(\theta^3) + \log(1-\theta)$$
$$= 3\log(\theta) + \log(1-\theta)$$

   (b) Give an expression for the derivative of the log likelihood. [2 pts]

$$\frac{d\,l(\theta)}{d\theta} = 3 \cdot \frac{1}{\theta} + -\frac{1}{1-\theta}$$
$$= \frac{3}{\theta} - \frac{1}{1-\theta}$$

(c) What is the maximum likelihood estimate of θ? [1 pts]

$$\frac{3}{\theta} - \frac{1}{1-\theta} = 0$$

$$3(1-\theta) - \theta = 0$$

$$3 - 3\theta - \theta = 0$$

$$4\theta = 3$$

$$\hat{\theta} = \frac{3}{4}$$

estimate is $\frac{3}{4}$

Decision Trees

13. We would like to learn a decision tree given the following pairs of training instances with attributes (a_1, a_2) and target variables.

Instance number	a_1	a_2	Target
1	T	T	T
2	T	T	T
3	T	F	F
4	F	F	T
5	F	T	F
6	F	T	F

For reference, for a random variable X that takes on two values with probability p and $1-p$, here are some values of the entropy function (we use **log to the base 2** in this question):

$$p = \{\tfrac{1}{2}\} : H(X) = 1 \qquad\qquad p \in \{\tfrac{1}{3}, \tfrac{2}{3}\} : H(X) \approx .92$$

(a) What is the entropy of the Target variable? [1 pts]

$$H[\text{Target}] = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)$$

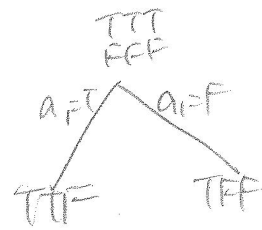
$$= -\left(\frac{1}{2}(-1) + \frac{1}{2}(-1)\right)$$

$$= -\left(-\frac{1}{2} - \frac{1}{2}\right)$$

$$= 1$$

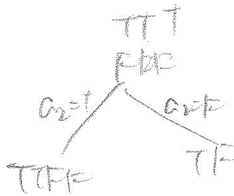
$\boxed{\text{alwys use base-2!}}$

(b) What is the information gain of each of the attributes a_1 and a_2 relative to the Target variable? [4 pts]



$H[\text{Target} \mid a_1] = \frac{1}{2} \times 0.92 + \frac{1}{2} \times 0.92 = 0.92$

Info Gain $= H[\text{Target}] - H[\text{Target} \mid a_1] = 0.08$



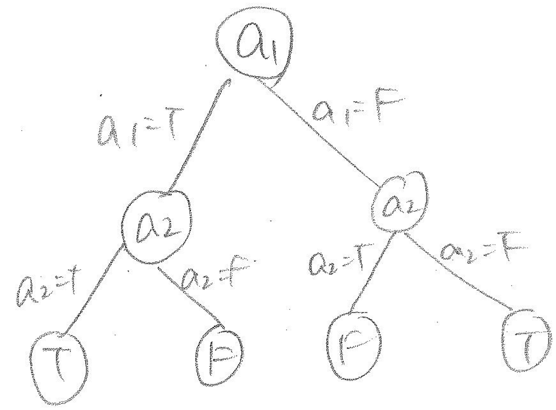
$H[\text{Target} \mid a_2] = \frac{2}{3} \times 1 + \frac{1}{3} \times 1 = 1$

Info Gain $= H[\text{Target}] - H[\text{Target} \mid a_2] = 0$

(c) Using information gain, which attribute will the ID3 decision tree learning algorithm choose as the root? [1 pts]

a_1.

(d) Construct a decision tree with zero training error on this training data. [2 pts]



(e) Change exactly one of the instances (either the attributes or labels) so that **no decision tree can attain zero training error** on this dataset (indicate the instance number and the change). [2 pts]

Instance number 2.
Change Target to F

	a_1	a_2	Target
2	T	T	F

lin reg. $\|X\theta - Y\|^2$

Linear Regression

14. We are given a set of N two-dimensional inputs and their corresponding output: $\{\boldsymbol{x}_n, y_n\}, \boldsymbol{x}_n = \begin{pmatrix} x_{n,1} \\ x_{n,2} \end{pmatrix} \in \mathbb{R}^2, y_n \in \mathbb{R}, n \in \{1, \ldots, N\}$. Given $\boldsymbol{x}_n$, we would like to use the following regression model to predict y_n:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}_n) = \theta_1^2\, x_{n,1} + \theta_2^2\, x_{n,2}.$$

We learn this model by finding values of the parameters (θ_1 and θ_2) that minimize the cost function defined as the sum of squared errors between the predicted and true labels (also called the residual sum of squares).

(a) Write out the cost function that is minimized (your answer should be expressed in terms of y_n, $x_{n,1}$, $x_{n,2}$, θ_1 and θ_2). [1 pts]

$$J = RSS = \sum_n (h_\theta(x_n) - y_n)^2 = \sum_n (\theta_1^2 x_{n,1} + \theta_2^2 x_{n,2} - y_n)^2$$

(b) Derive the optimal value(s) for θ_1. (You should find a closed-form solution. Note that θ_2 may appear in your resulting equation and that there may be more than one possible value for θ_1.) [5 pts]

$$\frac{\partial J}{\partial \theta_1} = \sum_n \frac{\partial}{\partial \theta_1}\left[(\theta_1^2 x_{n,1} + \theta_2^2 x_{n,2} - y_n)^2\right]$$

$$= \sum_n 2(\theta_1^2 x_{n,1} + \theta_2^2 x_{n,2} - y_n) \times 2x_{n,1} \times \theta_1 = 0$$

$$\sum_n (4x_{n,1}\theta_1)(\theta_1^2 x_{n,1} + \theta_2^2 x_{n,2} - y_n) = 0$$

$$\sum_n 4x_{n,1}^2 \theta_1^3 + \sum_n 4x_{n,1}x_{n,2}\theta_2^2\theta_1 - \sum_n 4x_{n,1} y_n \theta_1 = 0$$

see next page

$$\left(\sum_n 4x_{n,1}^2\right)\theta_1^3 + \left(\sum_n 4x_{n,1}x_{n,2}\right)\theta_2^2\theta_1 - \left(\sum_n 4x_{n,1}y_{n,1}\right)\theta_1 = 0$$

$\boxed{\text{Solution 1: } \theta_1 = 0}$

If $\theta_1 \neq 0$:

$$\left(\sum_n 4x_{n,1}^2\right)\theta_1^2 + \left(\sum_n 4x_{n,1}x_{n,2}\right)\theta_2^2 - \left(\sum_n 4x_{n,1}y_{n,1}\right) = 0,$$

$$\left(\sum_n x_{n,1}^2\right)\theta_1^2 + \sum_n x_{n,1}x_{n,2}\theta_2^2 - \sum_n x_{n,1}y_{n,1} = 0$$

$$\left(\sum_n x_{n,1}^2\right)\theta_1^2 + \sum_n x_{n,1}x_{n,2}\theta_2^2 - x_{n,1}y_{n,1} = 0$$

$$\left(\sum_n x_{n,1}^2\right)\theta_1^2 + \sum_n\left(x_{n,1}x_{n,2}\theta_2^2 - x_{n,1}y_{n,1}\right)$$

$$\theta_1^2 = \frac{\sum_n\left(x_{n,1}y_{n,1} - x_{n,1}x_{n,2}\theta_2^2\right)}{\sum_n\left(x_{n,1}^2\right)}$$

$$\boxed{\text{Solution 2, } \quad \theta_1 = \pm\sqrt{\frac{\sum_n\left(x_{n,1}y_{n,1} - x_{n,1}x_{n,2}\theta_2^2\right)}{\sum_n\left(x_{n,1}^2\right)}}}$$

Final solution

$$\boxed{\begin{aligned} &\theta_1 = 0 \\ &\quad\text{or} \\ &\theta_1 = \pm\sqrt{\frac{\sum_n\left(x_{n,1}y_{n,1} - x_{n,1}x_{n,2}\theta_2^2\right)}{\sum_n\left(x_{n,1}\right)^2}} \end{aligned}}$$

(Blank page provided for your work)