CM146 Midterm

Eric Lawrence Kong

TOTAL POINTS

51 / 60

QUESTION 1

1 Machine Learning Basics 6 / 6

✓ - 0 pts Correct

QUESTION 2

2 Logistic Regression 2 / 4

✓ - 2 pts a) Incorrect, logistic regression would best suit the problem description

QUESTION 3

3 True/False 8 / 12

✓ - 2 pts Q6 Incorrect

✓ - 2 pts Q7 Incorrect

QUESTION 4

Multiple Choice 7 pts

4.1 2 / 2

✓ - 0 pts Correct

4.2 2 / 2

✓ - 0 pts Correct

4.3 3 / 3

✓ - 0 pts Correct

QUESTION 5

5 Maximum Likelihood 5 / 5

✓ - 0 pts Correct

QUESTION 6

Decision Trees 10 pts

6.1 Entropy Y 1 / 1

✓ - 0 pts Correct

6.2 Information Gain 4 / 4

✓ - 0 pts Correct

6.3 Root split 1 / 1

✓ - 0 pts Correct

6.4 Zero training error tree 2 / 2

✓ - 0 pts Correct

6.5 Change instance 2 / 2

✓ - 0 pts Correct

QUESTION 7

Weighted Linear Regression 16 pts

7.1 objective function 3 / 3

✓ - 0 pts Correct

7.2 optimal value 2 / 5

✓ - 3 pts wrong intermediate steps

7.3 OLS log likelihood 3 / 3

✓ - 0 pts Correct

7.4 MLE; variance and weight relation 5 / 5

✓ - 0 pts Correct

CM 146 — Machine Learning: Midterm

Fall 2017

Name: Eric Kong

UID: [redacted]

Yup, redacted.

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	6	
2	Application	4	
3	True/False	12	
4	Multiple choice	7	
5	Maximum likelihood	5	
6	Decision Trees	10	
7	Regression	16	
Total		60	

1. (6 pts) **Machine Learning Basics**

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

In supervised learning, the inputs are labelled, and the objective is to build a function which can predict the labels of unseen inputs based on their feature values.
In unsupervised learning, the inputs are unlabelled. Rather than explicitly labelling them, the goal is to find some sort of intrinsic structure in the data.

(b) (2 pts) What is the main difference between classification and regression?

In classification, the labels are discrete;
in regression, they are continuous real values.

(c) (2 pts) What is the motivation to separate the available data into training and test data?

It is instructive to be able to compare the performance of different classifiers, in terms of accuracy. Separating the available data into training and test sets allows us to evaluate the training and test errors for our classifiers, and to plot their learning curves.

2. (4 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 would prefer a decision tree, because the features involved in predicting the presence of cancer are generally discrete, and usually binary. The other models generally involve real-valued features (and selection of a norm in these spaces), and are therefore not as suited to our decidedly medical task.

   (b) (2 pts) A model that attains 100% accuracy on the training set and 70% accuracy on the test set is better than a model that attains 80% accuracy on the training set and 75% accuracy on the test set.

   True (False)

   A model which attains zero training error and high test error epitomises the unwanted phenomenon of *overfitting*, wherein the model's predictions become too dependent on the data and cannot generalise. Because generalisation is at the heart of learning (machine or otherwise), overfitting indicates a failure in this regard.

True/False

3. (2 pts) 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 (so your new training set will have $n+1$ instances), 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.

True (False)

It is possible that the addition of an instance changes the information gains of the attribute in such a way that a different attribute is selected as the root.

4. (2 pts) Stochastic Gradient Descent is faster per iteration than Batch Gradient Descent.

(True) False

Stochastic Gradient Descent runs in $O(D)$ time per iteration whereas Batch Gradient Descent runs in $O(nD)$ time per iteration. where D is the dimensionality of the space and n is the number of training instances.

5. (2 pts) 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.

True (False)

The training data could be linearly separable, but take longer than 100 iterations to converge.

6. (2 pts) We want to learn a non-linear regression function to predict y from $\boldsymbol{x}$ where $y \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^D$ given training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. To do so, we transform $\boldsymbol{x}$ by a function $\boldsymbol{\phi}(\boldsymbol{x})$ and minimize the residual sum of squares objective function on the transformed features: $\sum_{i=1}^n \left(y_i - \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_i)\right)^2$. This optimization problem is convex.

True (False)

If the function $\phi(x) = -x$, then the optimisation problem becomes concave in x. Therefore, in general, the transformed optimisation problem does not generally remain convex in x. (However, the problem remains convex in $\phi(x)$.)

7. (2 pts) We want to use 1-Nearest Neighbors (1-NN) to classify houses into one of two classes (cheap vs expensive) given a single feature that measures the area of the house. The predictions made by the 1-NN classifier data can change if the area of the house is measured in square metres instead of square feet. (You can neglect the effect of ties *i.e.*, two training instances that are both nearest neighbors to a test instance.)

(True) False

The distance metric used in k-NN is a hyperparameter which can change the predictions in general. Changing the units used to measure the area is tantamount to changing the distance metric.

8. (2 pts) 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) False

f is convex, for $\forall x \in \mathbb{R}$, $f''(x) \geq 0$.

$f'(x) = 2(2x - 3)$

$f''(x) = 4$.

Because f is convex, gradient descent returns its global minimum necessarily.

Multiple choice

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

X (a) The decision boundary is smoother with smaller values of k. Larger 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). Just store entire training set. Nonparametric method.

(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. It will predict the class with a majority.

X (d) For larger values of k, it is more likely that the classifier will overfit than underfit. See above. For k too large, classifier will miss patterns.

10. (2 pts) Assume we are given a set of one-dimensional inputs and their corresponding output (that is, a set of $\{(x_i, y_i)\}, x_i \in \mathbb{R}, y_i \in \mathbb{R}$). We would like to compare the following two models on our input dataset where $\theta \in \mathbb{R}$:

$$\text{A} : y = \theta^2 x$$
$$\text{B} : y = \theta x$$

For each model, 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 be more *accurate* than B.

(b) There are datasets for which B would be more *accurate* than A.

(c) Both (a) and (b) are correct.

(d) They would perform equally well on all datasets.

Consider the data set $\{(1, 2), (-2, -4)\}$.
Then B would perfectly predict y given x using $\theta = -2$, but A cannot do so, for $\theta^2 \geq 0$ because $\theta \in \mathbb{R}$, and the data have a perfect negative correlation.
However, B can predict any data set to as much accuracy as A does, for if θ works for A, then $\sqrt{\theta}$ works for B.
(Aside: I have a feeling this question is confusingly worded.)

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 — increase complexity of training set allows for more errors

(b) test error will ... increase / (decrease) / unknown —

(c) overfitting will ... increase / (decrease) / unknown — more training data allows the model to free itself from the data that it has overfit itself to, thus decreasing test error

For these problems, you must show your work to receive credit!

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.*, $P(X_n = 1) = \theta$): $(X_1, X_2, X_3, X_4) = (1, 1, 0, 1)$.

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

$$L(\theta) = \mathbb{P}(x_1, x_2, x_3, x_4; \theta)$$
$$= \mathbb{P}(x_1 = 1; \theta)\mathbb{P}(x_2 = 1; \theta)\mathbb{P}(x_3 = 0; \theta)\mathbb{P}(x_4 = 1; \theta)$$
$$= \theta^3(1 - \theta)$$
$$l(\theta) = \log L(\theta) = 3\log\theta + \log(1 - \theta)$$

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

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{3}{\theta} - \frac{1}{1 - \theta}$$
$$= \frac{3(1 - \theta) - 1(\theta)}{\theta(1 - \theta)}$$
$$= \frac{3 - 4\theta}{\theta(1 - \theta)}$$

(c) What is the maximum likelihood estimate $\hat{\theta}$ of θ? [1 pts]

$$\hat{\theta} = \underset{\theta = \Theta}{\arg\max}\, l(\theta)$$

We find θ which maximises $l(\theta)$.

$$0 = \frac{dl(\theta)}{d\theta} = \frac{3 - 4\theta}{\theta(1-\theta)}$$

$$0 = 3 - 4\theta$$

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

Observe that

$$\frac{d^2 l}{d\theta^2} = -\frac{3}{\theta^2} - \frac{1}{(1-\theta)^2}$$

so

$$\left.\frac{d^2 l}{d\theta^2}\right|_{\theta = \frac{3}{4}} = -\frac{3}{\left(\frac{3}{4}\right)^2} - \frac{1}{\left(1 - \frac{3}{4}\right)^2}$$

$$= -\frac{16}{3} - 16 = -\frac{64}{3} < 0,$$

and by the Second Derivative Test,

$\theta = \frac{3}{4}$ gives a maximum.

Thus

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

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 variable Y.

Instance number	a_1	a_2	Y
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 Y? [1 pts]

$$H(Y) = -\sum_y P(Y=y) \lg P(Y=y)$$

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

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

$$= -(-1) = 1.$$

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

Information gain for a_1:

$H[Y] - H[Y \mid a_1]$

$H[Y \mid a_1 = T] = -\sum_y \mathbb{P}(Y = y \mid a_1 = T) \lg \mathbb{P}(Y = y \mid a_1 = T)$

$= -\left(\frac{2}{3} \lg \frac{2}{3} + \frac{1}{3} \lg \frac{1}{3}\right) = 0.92.$

$H[Y \mid a_1 = F] = -\sum_y \mathbb{P}(Y = y \mid a_1 = F) \lg \mathbb{P}(Y = y \mid a_1 = F)$

$= -\left(\frac{1}{3} \lg \frac{2}{3} + \frac{1}{3} \lg \frac{1}{3}\right) = 0.92.$

for a_1

$H[Y \mid a_1] = \frac{1}{2}(0.92) + \frac{1}{2}(0.92) = 0.92.$ Information gain: 0.08.

Information gain for a_2:

$H[Y \mid a_2 = T] = 1$ $H[Y \mid a_2 = F] = 1.$

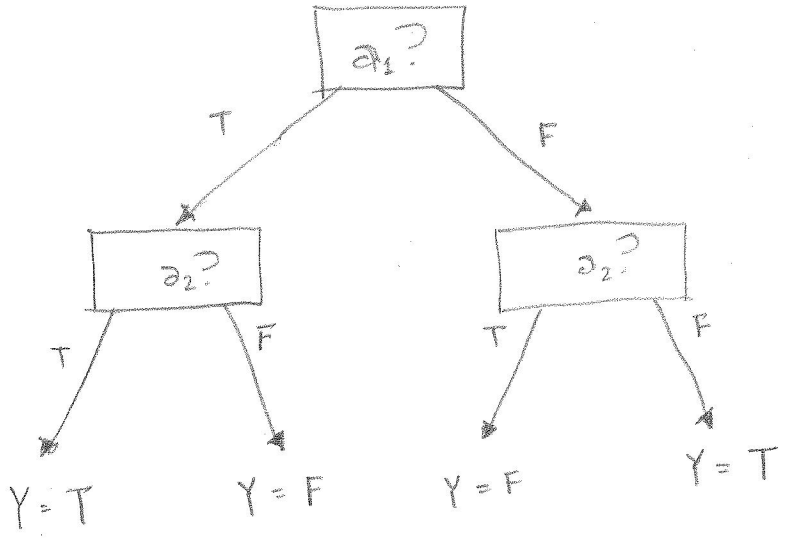
for a_2

$H[Y \mid a_2] = \frac{2}{3}(1) + \frac{1}{3}(1) = 1.$ Information gain: 0.

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

The ID3 algorithm will choose a_1 as the attribute for the root, for it provides maximum information gain.

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



(e) Change exactly one of the instances (by changing either the attributes or labels but not both) so that **no decision tree can attain zero training error** on this dataset (indicate the instance number and the change). [2 pts]

If we change instance 2 to read

a_1	a_2	Y
T	F	T

then it will have identical feature values as instance 3, but a different label. Thus no decision tree can classify both instances correctly and attain zero training error.

Weighted linear regression

14. In the problem set, we considered weighted linear regression where the input features are 1-dimensional. We now extend this to D-dimensional features. Thus, we want to find $\boldsymbol{\theta}$ that minimizes the cost function

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{N} w_n (y_n - \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}_n)^2$$

Here $w_n > 0$, $\boldsymbol{x}_n \in \mathbb{R}^{D+1}, \boldsymbol{\theta} \in \mathbb{R}^{D+1}$. $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_N^{\mathrm{T}} \end{pmatrix} \in \mathbb{R}^{N \times (D+1)}$, $\boldsymbol{y} \in \mathbb{R}^N$. For this problem, assume that the intercept term is included in the $\boldsymbol{\theta}$ and that the linear regression solution exists in this setting.

Questions:

(a) Show that $J(\boldsymbol{\theta})$ can also be written as:

$$J(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})$$

Here $\boldsymbol{W}$ is a diagonal matrix where the entry on the diagonal on row n, column n is w_n. [3 pts]

Let $z = y - X\theta$. Then $z_n = y_n - \theta^T x_n$.

By definition of matrix multiplication,

$$z^T W z = \sum_{j,k} z_j z_k W_{jk}$$

Because W is diagonal with $W_{nn} = w_n$, $W_{jk} = 0$ $(j \neq k)$,

$$z^T W z = \sum_n z_n^2 W_{nn} = \sum_n w_n z_n^2$$

$$= \sum_n w_n (y_n - \theta^T x_n)^2$$

But $z^T W z$ is just $J(\theta)$!

So $J(\theta) = (y - X\theta)^T W (y - X\theta) = \sum_n w_n (y_n - \theta^T x_n)^2$

(b) Show that the optimal value for $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{y}$. For reference, here are some useful gradient identities (where $\boldsymbol{x}, \boldsymbol{b}$ are vectors and $\boldsymbol{A}$ is a symmetric matrix).

(I) $\quad f(\boldsymbol{x}) = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} \qquad \nabla f(\boldsymbol{x}) = \boldsymbol{b}$

(II) $\quad f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} \qquad \nabla f(\boldsymbol{x}) = 2\boldsymbol{A}\boldsymbol{x}$

[5 pts]

We minimise $J(\theta)$ by solving $\nabla J(\theta) = 0$. ← zero vector in $\mathbb{R}^n$.

$$\nabla\left((y - X\theta)^T W (y - X\theta)\right) = 0.$$

$$2W(y - X\theta) = 0 \qquad \text{by identity (II)}.$$

$$W(y - X\theta) = 0.$$

$$Wy - WX\theta = 0$$

$$WX\theta = Wy$$

$$X^T WX\theta = X^T Wy$$

$$\theta = (X^T WX)^{-1} X^T Wy \quad \blacksquare$$

(c) In class, we provided a probabilistic interpretation of ordinary least squares (OLS). We now try to provide a probabilistic interpretation of weighted linear regression. Consider a model where each of the N samples is independently drawn according to a normal distribution

$$P(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(y_n - \boldsymbol{\theta}^T\boldsymbol{x}_n)^2}{2{\sigma_n}^2}\right)$$

In this model, each y_n is drawn from a normal distribution with mean $\boldsymbol{\theta}^T\boldsymbol{x}_n$ and variance ${\sigma_n}^2$. The ${\sigma_n}^2$ are **known**. Write the log likelihood of this model as a function of $\boldsymbol{\theta}$. [3 points]

$$L(\theta) = \prod_{n=1}^{N} P(y_n|x_n, \theta)$$

$$l(\theta) = \log \prod_{n=1}^{N} P(y_n|x_n, \theta)$$

$$= \log \prod_{n=1}^{N} \left(\frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(y_n - \theta^T x_n)^2}{2\sigma_n^2}\right)\right)$$

$$= \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma_n^2}}\right)\left(-\frac{(y_n - \theta^T x_n)^2}{2\sigma_n^2}\right)$$

$$= \sum_{n=1}^{N} \frac{\log\left(\sqrt{2\pi\sigma_n^2}\right)}{2\sigma_n^2}(y_n - \theta^T x_n)^2.$$

(d) Show that finding the maximum likelihood estimate of $\boldsymbol{\theta}$ leads to the same answer as solving a weighted linear regression. How do ${\sigma_n}^2$ relate to w_n? [5 points]

Observe that $l(\theta)$ is in the form

$$l(\theta) = \sum_{n=1}^{N} w_n(y_n - \theta^T x_n)^2$$

where $w_n = \frac{\log\left(\sqrt{2\pi\sigma_n^2}\right)}{2\sigma_n^2}$.

The above expression should look familiar, for it is $J(\theta)$ from two pages ago. Therefore solving the weighted linear regression, where the weights $w_n = \frac{\log\left(\sqrt{2\pi\sigma_n^2}\right)}{2\sigma_n^2}$, is equivalent to finding the maximum likelihood estimate of θ. ■

(Blank page provided for your work)