CM146: Introduction to Machine Learning | Fall 2019

Midterm Solution

Nov. 4^{th}, 2019

- Please do not open the exam unless you are instructed to do so.
- This is a closed book and closed notes exam.
- Everything you need in order to solve the problems is supplied in the body of this exam.
- Mark your answers ON THE EXAM ITSELF. If you make a mess, clearly indicate your final answer (box it).
- If you think something about a question is open to interpretation, please make a note on the exam.
- If you run out of room for your answer in the space provided, you can write it down in the last page and indicate clearly that you've done so.
- You may ask TA for scratch paper or scratch in the last page of the exam.
- You have 1 hour 30 minutes (90 minutes).
- **Besides having the correct answer, being concise and clear is very important. For full credit, you must show your work and explain your answers.**

Good Luck!

Legibly write your name and UID in the space provided below to earn **2 points**.

Name: Varun Raju

UID: 204979452

Short Questions (27pts)

1. (10 pts) True OR False (check the box).

   (a) Decision tree with a larger depth is more likely to generalize better to new data points.
   ☐ True ☑ False

   (b) 5-NN (KNN with K=5) is more robust to outliers than 1-NN.
   ☑ True ☐ False

   (c) Comparing to stochastic gradient descent, gradient descent can always find the global minimum.
   ☑ True ☐ False

   (d) A classifier that attains 100% accuracy on the training set is always better than a classifier that attains 70% accuracy on the same training set.
   ☐ True ☑ False

   (e) If data is not linearly separable, K-NN algorithm cannot reach training error zero.
   ☐ True ☑ False

2. (9 pts) Multiple Choice (check the box).

   (a) Suppose we want to compute 10-Fold Cross-Validation error on 1000 training examples. We need to compute error N_1 times, and the Cross-Validation error is the average of the errors. To compute each error, we need to build a model with data of size N_2, and test the model on the data of size N_3. What are the appropriate numbers for N_1, N_2, N_3?

   ☑ A. $N_1 = 10$, $N_2 = 900, N_3 = 100$ 10 900 100
   ☐ B. $N_1 = 1, N_2 = 800, N_3 = 200$
   ☐ C. $N_1 = 10, N_2 = 1000, N_3 = 100$
   ☐ D. $N_1 = 1, N_2 = 1000, N_3 = 1000$

   (b) Let $X_1, \ldots, X_N$ are i.i.d. random variables with the same distribution of a random variable X. Let $E[X]$ to be the expectation of X. What is the expectation of $X_1 + X_2 + \ldots + X_N$?
   ☐ A. $E[X]$ ☑ B. $NE[X]$ ☐ C. $N^2E[X]$ ☐ D. 0

   (c) A coin is tossed 100 times and lands heads 60 times. What is the maximum likelihood estimate for the probability of heads.
   ☐ A. 1 ☐ B. 0.2 ☑ C. 0.6 ☐ D. 0

3. (8 pts) We are given two-dimensional inputs x_i and their corresponding output y_i. We denote $x_{i,1}$ and $x_{i,2}$ to be the first and second dimension of x_i. We use the following linear regression model to predict y:

$$y_i = w_1 x_{i,1} + w_2 x_{i,2}.$$

Given a data set $\{(x_i, y_i)\}, i = 1, \ldots, N$, derive the best w_1 and w_2 that minimize the square error. To simplify the answer, you can use the following notations:

$$\alpha_1 = \sum_{i=1}^{N} x_{i,1}^2, \quad \alpha_2 = \sum_{i=1}^{N} x_{i,2}^2, \quad \alpha_{12} = \sum_{i=1}^{N} x_{i,1}x_{i,2}, \quad \beta_1 = \sum_{i=1}^{N} x_{i,1}y_i, \quad \beta_2 = \sum_{i=1}^{N} x_{i,2}y_i.$$

Square error:

$$J = \boxed{\sum_{i=1}^{N} \left(y - (w_1 x_{i,1} + w_2 x_{i,2})\right)^2} \rightarrow \sum_{i=1}^{N} (y - y_i)^2$$

$$\frac{\partial J}{\partial w_1} = \sum_{i=1}^{N} \frac{\partial J}{\partial w_1} \left(y - (w_1 x_{i,1} + w_2 x_{i,2})\right)^2$$

we set $\frac{\partial J}{\partial w_1}$ to 0 to minimize.

$$0 \Rightarrow \sum_{i=1}^{N} 2\left(y - w_1 x_{i,1} - w_2 x_{i,2}\right) \cdot -x_{i,1}$$

$$\sum_{i=1}^{N} w_1 x_{i,1}^2 \Rightarrow \sum_{i=1}^{N} \left(y - w_2 x_{i,2}\right) x_{i,1} \quad \text{and} \quad \boxed{w_1 = \sum_{i=1}^{N} \frac{y - w_2 x_{i,2}}{\sum_{i=1}^{N} x_{i,1}}}$$

$$\frac{\partial J}{\partial w_2} = \sum_{i=1}^{N} \frac{\partial J}{\partial w_2} \left(y - (w_1 x_{i,1} + w_2 x_{i,2})\right)^2$$

$$= \sum_{i=1}^{N} 2\left(y - w_1 x_{i,1} - w_2 x_{i,2}\right) \cdot -x_{i,2}$$

$$\boxed{w_2 = \sum_{i=1}^{N} \frac{y - w_1 x_{i,1}}{\sum_{i=1}^{N} x_{i,2}}}$$

Decision Tree (15 pts)

Consider the following training dataset with 2 features (Age and Weight), and the outcome is Diabetes. You don't have to simplify your answer and note $\log_2 3 \approx 1.6, \log_2 5 \approx 2.3$.

Patient	*Age*	*Weight*	*Diabetes*
1	Young	Heavy	No
2	Young	Heavy	No
3	Young	Heavy	No
4	Young	Heavy	No
5	Young	Light	No
6	Young	Light	No

Patient	*Age*	*Weight*	*Diabetes*
7	Young	Light	No
8	Young	Light	No
9	Old	Heavy	Yes
10	Old	Heavy	Yes
11	Old	Light	No
12	Old	Light	No

1. (3 pts) What is the entropy H(Diabetes)?
   Hint: $H(S) = -\sum_{v=1}^{K} P(S = a_v) \log_2 P(S = a_v)$

$H(\text{Diabetes}) = -\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6}$

$- \left(-0.431 + -0.219\right) \rightarrow \boxed{0.65}$

2. (3 pts) What is the information gain if we partition the data on the attribute *Age*?
   Hint: $\text{Gain}(S, A) = H(S) - \sum_{v \in \text{Value}(A)} \frac{|S_v|}{|S|} H(S_v)$.

Partition on Age: Young = 8, Old = 4

Entropy of (D | Young) = 0, Entropy

$\left(-0\log 0 - 1\log 1 \times \frac{8}{12}\right) + \left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right)\cdot\frac{4}{12}$

$\therefore \text{Gain}(D, A) = \left(-\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6}\right) + \frac{4}{12}\left(\log\frac{1}{2}\right) \rightarrow = \cancel{0.33}$

3. (3 pts) What is the information gain if we partition the data on the attribute *Weight*?

Partition on weight: Heavy = 1/2, Light = 1/2

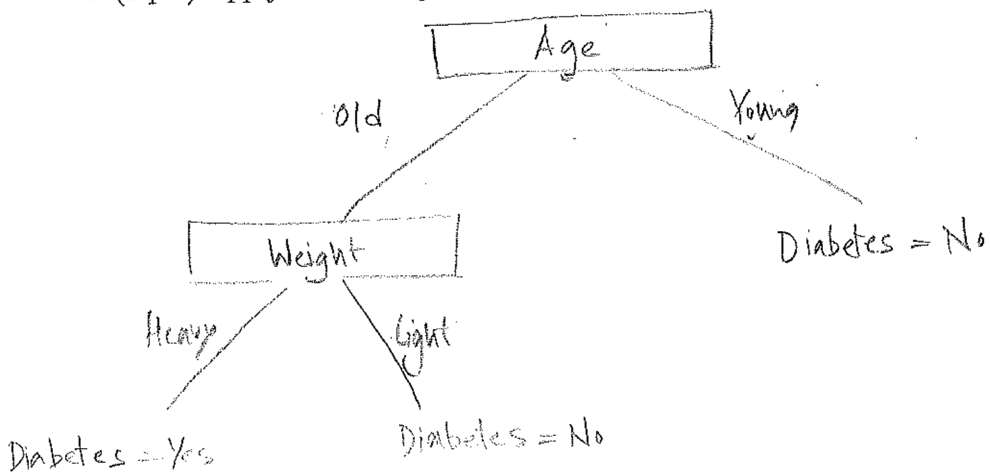
$H(D \mid \text{Light}) = 0$, $H(D \mid \text{Heavy}) =$

$\Rightarrow -\frac{2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6} \times \frac{1}{2} \rightarrow \cancel{0.52}$

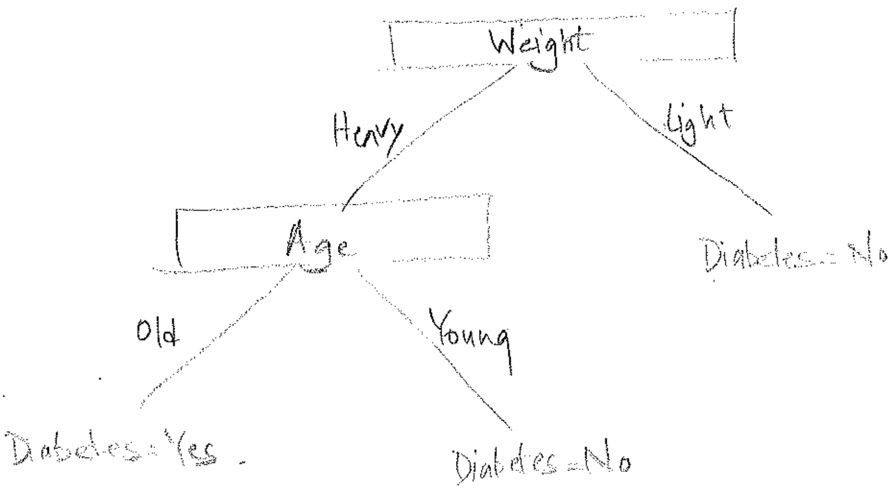
$\cancel{0.154}$

$\boxed{\therefore \text{Gain}(D, W) = \left(-\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6}\right) + \frac{1}{2}\left(\frac{2}{6}\log\frac{2}{6} + \frac{4}{6}\log\frac{4}{6}\right)}$

4. (3 pts) Apply the ID3 algorithm to build the tree using both features Age and Weight.



5. (3 pts) Find another tree that yields the same training error as the tree built by ID3.



Perceptron (20 pts)

In this problem we consider various variants of Perceptron and explore their properties.

1. (2 pts) First, complete the Line 3 of the Perceptron algorithm by choosing from the following options. (Hint: Given a data point (x_i, y_i), if the current model parametrized by w makes a wrong decision, the model will update.)
   (a) $w^T x_i \geq 0$ (b) $y_i = 1$ (c) $y_i w^T x_i \geq 0$ (d) $y_i w^T x_i \leq 0$
   (e) $w^T x_i < 0$ (f) $y_i = -1$

Algorithm 1 Perceptron with learning rate α

1: Initialize $w = \vec{0}$
2: **for** each $(x_i, y_i) \in \mathbb{D}_{train}$, where $y_i = 1$ or -1 **do**
3: **if** (d) $y_i w^T x_i \leq 0$ **then**
4: $w \leftarrow w + \alpha y_i x_i$
5: **end if**
6: **end for**
7: **return** w

2. (3 pts) Can we set $\alpha = 0$? First answer yes/no then explain your answer.
   No. If α is zero, the model is useless as it will never "learn" (update its weights) from its mistakes.

3. (5 pts) Choose a learning rate such that when the algorithm sees two consecutive occurrences of the same example, it will never make a mistake on the second occurrence. Prove your answer is correct. (Hint: before update, statement in Line 3 is True. After the update, it has to be False)
   We have to prove that if $\boxed{y_i w_t^T x_i \leq 0, \; y_i w_{t+1}^T x_i \geq 0}$

   So, we know that $w_{t+1} = w_t + \alpha y x$
   Multiply both sides with y and x.
   We get: $y \cdot w_{t+1} \cdot x = y \cdot w_t \cdot x + \alpha y^2 x^2$
   We want LHS ≥ 0 so RHS must also be ≥ 0

   $$y w_t \cdot x + \alpha y^2 x^2 \geq 0$$
   $$\alpha y^2 x^2 \geq y w_t \cdot x$$

   $y^2 = 1$

   6

   $$\therefore \boxed{\alpha \geq \frac{y w_t \cdot x}{x^2}}$$

   If α obeys this condition, it will never make a mistake on the second occurrence.

4. We prove the convergence of Perceptron algorithm in class. In the following, we will derive the mistake bound of the Perceptron algorithm with learning rate α. We assume (1) there exist a vector u and $\gamma > 0$ such that $||u|| = 1$ and for all data $(x_i, y_i) \in \mathbb{D}_{train}$, $y_i(u^T x_i) \geq \gamma$; (2) there exist $R > 0$ such that $||x_i|| \leq R$. Complete the following proof.

(a) (3 pts) Let w^t represent the weight vector w after t updates. We further assume $w^0 = \vec{0}$ (i.e., w is initialized with a vector with all zeros). Prove $w^t \cdot u \geq t\alpha\gamma$.

$w^t \cdot u \geq t \cdot \alpha \cdot \gamma$ | at w^0, $t\alpha\gamma = 0$ so true
at w^1, $t\alpha\gamma \leq w^1 u$

At each update, we add an αx_i to w where $x_i \geq \gamma$ because $y_i(u^T x_i) \geq \gamma$.
So over t iterations $w^t \cdot u = \sum_{i=1}^{t} x_i \cdot \alpha$ which is $\geq t \cdot \alpha\gamma$ ← as x_i "stays ahead" of γ.

(b) (3 pts) Show that after t updates, $||w^t||^2 \leq t\alpha^2 R^2$.

We also know that $||x_i|| \leq R$ for all i.
Similar to the proof above, at each update we add αx_i to w where $||x_i|| \leq R$. So $||w|| \leq \alpha R$ always. As nothing is negative here, we can square to get over t iterations: $||w^t||^2 \leq t\alpha^2 R^2$

(c) (3 pts) What is the mistake bound of the Perceptron algorithm with learning rate α? Prove your answer.

It is $\left(\frac{R}{\gamma}\right)^2$.

$t\alpha\gamma \leq w^t \cdot u$

square to get $t\alpha^2\gamma^2 \leq ||w^t||^2 \cdot 1 \leq t\alpha^2 R^2$

bounds to $\frac{R^2}{\gamma^2}$

(d) (1 pts) Does the choice of α affect the mistake bound? (Yes/No)

No.

Logistic Regression (24 pts)

Remember we mentioned in the lecture, for a binary classification problem $y = 1, -1$ logistic regression model $P(y = 1|x)$ by $P(y = 1|x) = \sigma(w^T x) = 1/(1 + \exp(-w^T x))$. In the following, we consider a variant of logistic regression and model $P(y = 1|x)$ with

$$\sigma_\gamma(w^T x) = \frac{1}{1 + \exp(-w^T x/\gamma)},$$

where $\gamma > 0$ is a hyper-parameter that can be tuned. Answer the following questions.

1. (3 pts) When $w^T x \to \infty$, what is the value of $\sigma_\gamma(w^T x)$?

   $\sigma_y(w^T x) = 1$

2. (3 pts) When $w^T x \to -\infty$, what is the value of $\sigma_\gamma(w^T x)$?

   $\sigma_y(w^T x) = 0$

3. (3 pts) What happen when $\gamma \to \infty$?

   $\sigma_\gamma(w^T x) \to 1/2$ $1/e^{(w^T x/y)}$

4. (3 pts) What happen when $\gamma \to 0$?

   $\sigma_\gamma(w^T x) \to 1$

5. (4 pts) Show that for any γ the decision boundary is a linear function.

   $\frac{w^T x}{\gamma}$ = decision boundry

   We can prove that $\sigma_\gamma(w^T x) \geq 1/2$ reduces to $\frac{w^T x}{\gamma} \geq 0$

   $\frac{1}{1+e^{-w^T x/\gamma}} \geq 1/2$, $2 \geq 1 + e^{-w^T x/\gamma}$, $e^{w^T x/\gamma} \geq 1$ $\therefore \boxed{w^T x/\gamma \geq 0}$

   linear function.

6. (4 pts) Write down $P(y = -1|x)$.

   $P(y=-1|x) = 1 - P(y=1|x)$

   $\Rightarrow \boxed{1 - \frac{1}{1+\exp(-w^T x)}}$

7. (4 pts) Given a dataset $(x_i, y_i), i = 1, \ldots, N$. Write down the optimization problem maximizing the log-likelihood of the above model.

   $$\max \sum_{i=1}^{N} P(y=1|x_i)^{x_i} \cdot P(y=-1|x_i)^{1-x_i}$$

   $$\boxed{\Rightarrow \max \sum_{i=1}^{N} x_i \log(\sigma_\gamma(w^T x_i)) - (1-x_i)\log(1-\sigma_\gamma(w^T x_i))}$$

Maximum Likelihood (12 pts)

Let $x_1, \ldots, x_N$ are i.i.d. random samples from the exponential distribution with the probability density function (pdf):

$$P(x) = \lambda \exp(-\lambda x).$$

Answer the following questions.

1. (3 pts) Write down the joint probability of $P(x_1, x_2, \ldots, x_N)$.

Joint probability

$P(x) = \lambda \cdot e^{-\lambda x}$

$P(x_1, x_2, \ldots, x_n) = \lambda \cdot e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \cdots \lambda e^{-\lambda x_n}$

2. (3 pts) What is the log likelihood of λ given the dataset $\{x_1, x_2, \ldots, x_N\}$?

$F(\lambda) = \log \lambda \left(-\lambda x_1 - \lambda x_2 - \cdots - \lambda x_n\right)$

$\sum_{i=1}^{N} P(x_i) \cdot P($

3. (6 pts) Derive the maximum likelihood estimator of λ (i.e., find the λ that maximizes the likelihood).

$\frac{\partial F}{\partial \lambda} = \frac{-1}{\lambda} \frac{\partial}{\partial \lambda}\left(\lambda x_1 + \lambda x_2 + \cdots + \lambda x_n\right)$

$0 = -\frac{1}{\lambda} \cdot \sum_{i=1}^{N} (x_i)$

If you run out of room in answering questions, you can continued your answer here. Please indicate clearly that the answer is in the last page.