# 19F-COMSCIM146-1 Midterm

YUAN CHENG

TOTAL POINTS

**95 / 100**

QUESTION 1

## 1 Name 2 / 2

✓ **- 0 pts** Correct

QUESTION 2

## Short Questions 27 pts

### 2.1 True/False 10 / 10

✓ **- 0 pts** Correct

- **- 2 pts** a) incorect
- **- 2 pts** b) incorrect
- **- 2 pts** c) incorrect
- **- 2 pts** d)incorrect
- **- 2 pts** e) incorrect

### 2.2 Multiple Choice 9 / 9

✓ **- 0 pts** Correct

- **- 3 pts** a) incorrect
- **- 3 pts** b) NE[X}
- **- 3 pts** c) 0.6

### 2.3 Regression 8 / 8

Write down the J(w)

✓ **+ 2 pts correct write down the square error**

**+ 1 pts** partly wrong of the error formulation

correct gradient w.r.t. w1

✓ **+ 2 pts all three terms correct**

**+ 1 pts** partly correct (pos/neg sign)

**+ 0 pts** wrong gradient

correct gradient w.r.t. w2

✓ **+ 2 pts all three terms are correct**

**+ 1 pts** partly correct (+/-)

**+ 0 pts** wrong gradient

correct simplified result for w1

✓ **+ 1 pts correct result for w1**

**+ 0.5 pts** correct result based on the wrong gradient calculation or partly correct simplification; or no final substituation at all (get equation w1=sth. but has w2 there)

**+ 1 pts** correct closed form result

**+ 0.5 pts** trying to use closed form but wrong calculation

**+ 0 pts** no solution for w1 or completely wrong answer

correct simplified result for w2

✓ **+ 1 pts correct final result**

**+ 0.5 pts** partly wrong because the wrong gradient calculation or wrong simplification procedure; or no final simplification at all (get w2=some equation but still have w1/w2 there)

**+ 1 pts** correct matrix formulation

**+ 0.5 pts** using closed form but wrong formulation

**+ 0 pts** no solution for w2 or completley wrong answer

Directly use the closed form

**+ 8 pts** correct w result

**+ 2 pts** correct closed form equation

**+ 3 pts** correctly write down the X^TX

**+ 3 pts** correctly write down the (X^Ty)

**+ 1 pts** wrong shape of the matrix

**+ 1 pts** partly mistake in the last final step

**+ 1 pts** partly correct closed format

**+ 1 pts** partly correct substituation of the matrix

**+ 0 pts** nothing

QUESTION 3

## 3 Decision Tree 15 / 15

Q1

**- 2 pts** 100% wrong formula

**- 0 pts** Wrong answer, we didn't ask to compute exact number, so as long as entropy formulation is correct, we don't remove points.

**- 1 pts** Half wrong formula

**- 3 pts** Anything else not above

Q2

**- 2 pts** 100% wrong formula

**- 0 pts** Wrong answer, we didn't ask to compute exact number, so as long as entropy formulation is correct, we don't remove points.

**- 1 pts** Half wrong formula

**- 3 pts** Anything else not above

Q3

**- 2 pts** 100% wrong formula

**- 0 pts** Wrong answer, we didn't ask to compute exact number, so as long as entropy formulation is correct, we don't remove points.

**- 1 pts** Half wrong formula

**- 3 pts** Anything else not above

Q4

**- 1 pts** Wrong tree but correct use of entropy and reasoning in Info Gain

**- 2 pts** Wrong tree without any reasoning. Remove points when entropy formula and numbers are correct. Conditioned on Q1-3, students should know how to use Info Gain. This is important concept.

**- 1 pts** Draw 2 trees but did not say which one is chosen

**- 1.5 pts** Correct reasons, but incorrectly use entropy formula

**- 1 pts** Wrong application of entropy formula, and wrong tree without explanation.

**- 1 pts** Not consistent with entropy computation.

**- 1 pts** write algorithm but not say what is the tree, or say using info gain

Q5

**- 2 pts** Anything absolutely wrong. We do not take points off Q5 if its answer depends on Q4.

**- 3 pts** Does not attempt at all.

✓ **- 0 pts 100% correct everything**

QUESTION 4

## 4 Perceptron 18 / 20

✓ **+ 2 pts Q1 [2 pts]: answer is "d".**

Q2 [3 pts]

✓ **+ 1 pts Answer is No.**

✓ **+ 2 pts The weight will not change and stay the same as the initial value.**

**+ 0 pts** Note: weight can be initialized as any value and not necessarily to be zero.

**- 0.5 pts** Partially Correct for the second checkpoint: not only the weight not converged, the weights actually will not be updated

**+ 0 pts** Note: I think you understand this question, although we shouldn't allow the parameter not updated.

Q3 [5 pts]

✓ **+ 2 pts Before Update: y_i w^T x_i <= 0;
After Update: we want y_i (w + \alpha y_i x_i)^T x_i > 0**

✓ **+ 2 pts \alpha > -y_i w^T x_i / ||x_i||^2 y_i^2**

✓ **+ 0.5 pts Since y_i = +1 or -1, y_i^2 = 1**

✓ **+ 0.5 pts Therefore: \alpha > -y_i w^T x_i / ||x_i||^2**

**- 1 pts** Sign is reversed for checkpoint 1.

**- 0.5 pts** Lack of explanation for the second, third and fourth checkpoints.

**+ 0 pts** We can actually choose any learning rate larger than -yi w xi / ||xi||^2, instead of only the marginal value.

**+ 0 pts** Note: I assume you know the third checkpoint: y_i^2 = 1, but you should actually clarify it.

**- 0.5 pts** Note: you cannot delete x or y from nominator and denominator, as it's matrix multiplication

**- 0.5 pts** Note: x not necessarily only have two dimension.

**- 0.5 pts** Note: It's ||x||^2, not ||x^2||. Please see the definition of norm.

**- 0.5 pts** Note: miss a x for norminator

**- 0.5 pts** Sign is reversed for checkpoint 2

Q4(a) [3 pts]

✓ + **1 pts** **u^T w^t = u^T (w^{t-1} + \alpha y_{I_t} x_{I_t}).**

✓ + **1 pts** **\forall i, y_i(u^Tx_i) >= \gamma.**

✓ + **1 pts** **induction and initial state (w^0 = 0).**

- **1 pts** Missing alpha for the first point, and more alpha in the second checkpoint.

✓ - **0.5 pts** **Lack of explanation for the third checkpoint on initial state (w^0=0)**

- **1 pts** Didn't combine the first two checkpoints and get correct bound.

+ **0 pts** Note: in 4(b) you mention w^0 = 0, here I assume you know it, but you should actually clarify it.

- **0.5 pts** Didn't finish the proof

Q4(b) [3 pts]

✓ + **0.5 pts** **||w^t||^2 = ||w^{t-1} + \alpha y_{I_t} x_{I_t}||^2 = ||w^{t-1}||^2 + 2\alpha y_{I_t} (w^{t-1})^T x_{I_t} + (\alpha y_{I_t})^2 ||x_{I_t}||^2.**

+ **0.5 pts** Since the perceptron with parameter (w^t) makes mistake on data point (x_{I_t}, y_{I_t}): y_{I_t} (w^{t-1})^T x_{I_t} < 0.

✓ + **0.5 pts** **Since y_i = +1 or -1, y_i^2 = 1.**

✓ + **0.5 pts** **By definition, ||x_t||^2 <= R^2**

✓ + **1 pts** **induction and initial state (w^0 = 0).**

- **0.5 pts** Lack of explanation for the second, third and fourth checkpoints (how to derive ||w^t||^2 - ||w^{t-1}||^2 <= alpha^2 R^2).

✓ - **0.5 pts** **Lack of explanation for the fifth checkpoint on initial state (w^0=0)**

+ **2.5 pts** Another proof. I think it's correct, except that you should clarify w^0=0. Also, better to clarify why ||sum(yi*xi)||^2 <= max(||yixi||^2)

+ **2.5 pts** Another proof. I think it's correct, except that you should clarify ||yi||=1.

Q4(c) [3 pts]

✓ + **1 pts** **Cauchy Schwarz Inequality: w^t \cdot u <= ||u||||w^t||.**

✓ + **1 pts** **||u|| = 1.**

✓ + **1 pts** **mistake bound is R^2/gamma^2.**

✓ - **0.5 pts** **Lack of explanation for the first and second checkpoints.**

+ **0 pts** Note: □ means vertical, and // means parallel. The in-equation w \cdot u = ||w|| only holds when u // w, instead of u □ w. And also this inequality is cauchy schwarz inequality, better to clarify it.

✓ + **1 pts** **Q4(d) [1 pts]: No, as \alpha doesn't appear in the mistake bound.**

QUESTION 5

## 5 Logistic Regression 21 / 24

- **3 pts** Logistic Regression Q1 : correct ans is 1. Your ans is incorrect.

- **3 pts** Logistic Regression Q2. correct ans is 0. Your ans is incorrect.

- **3 pts** Logistic Regression Q3: correct ans is 1/2. Your ans is incorrect.

Logistic Regression Q4. Given gamma = 0

- **1 pts** sigma = 1 when w^Tx>0 (not mentioned)

✓ - **1 pts** **sigma = 1/2 when w^Tx=0 (not mentioned)**

✓ - **1 pts** **sigma = 0 when w^Tx<0 (not mentioned)**

- **0 pts** You have slight mistake

Logistic Regression Q5.

- **1 pts** Did not mention Probability condition unchanged or P(y=1|x) >=0.5

- **2 pts** Did not derive From P(y=1|x) >=0.5, to w^Tx >= 0

- **1 pts** w^Tx >= 0

+ **2 pts** Simply mention that From P(y=1|x) >=0.5, it can be derived that w^Tx >= 0 but does not derive it.

- **0 pts** You have slight mistake in calculation.

- **4 pts** Logistic Regression Q6. Correct ans is 1-P(y=1|x) = 1/(1+exp(w^Tx/gamma)). Your answer is incorrect.

Logistic Regression Q7.

✓ - **1 pts** **Need to consider both cases when y=1 and y=-1.**

- **2 pts** incorrect/no mention of product of P(y|x) or sum of log(p(y|x))

- **1 pts** incorrect/no mention of (i) max of -log function or (ii) min of +log or max of P or (iii) product/sum but mention P or log P

**- 0 pts** You have slight mistake. or Your answer is unclear.

**- 1 pts** either in the equation (1) You used sigma in place of y or (2) did not use any y or (3) you forgot to use log

**- 0 pts** Logistic Regression: All correct

QUESTION 6

## 6 Maximum Likelihood **12 / 12**

1)

**+ 1.5 pts** Product decomposition

**+ 1.5 pts** Correct indices/expression

2)

**+ 1.5 pts** Log of joint distribution

**+ 1.5 pts** Correct log transformation

3)

**+ 2 pts** Partial with respect to lambda

**+ 1 pts** mistake on partial with respect to lambda

**+ 1 pts** set to zero + wrong math

**+ 2 pts** Setting to zero

**+ 2 pts** Correct final expression

**+ 1 pts** Final expression minor mistake (sign, N missing)

**+ 0 pts** No attempt/wrong

✓ **+ 12 pts All correct**

| 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:** Yuan Cheng

**UID:** 404931874

## 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. $N_1 = 10,\ N_2 = 900, N_3 = 100$
   ☐ 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:

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

$$J(w) = \frac{1}{2}\sum_{i=1}^{N}(y_i - w^T x)^2 = \frac{1}{2}\sum_{i=1}^{N}(y_i - w_1 x_{i,1} - w_2 x_{i,2})^2$$

$$\frac{\partial J(w)}{\partial w_1} = \frac{1}{2}\sum_{i=1}^{N} 2(y_i - w_1 x_{i,1} - w_2 x_{i,2})(-x_{i,1})$$

$$= -\sum_{i=1}^{N} x_{i,1}y_i + w_1 \sum_{i=1}^{N} x_{i,1}^2 + w_2 \sum_{i=1}^{N} x_{i,2}x_{i,1}$$

$$= -\beta_1 + w_1\alpha_1 + w_2\alpha_{12} = 0 \quad (1)$$

$$\frac{\partial J(w)}{\partial w_2} = \frac{1}{2}\sum_{i=1}^{N} 2(y_i - w_1 x_{i,1} - w_2 x_{i,2})(-x_{i,2})$$

$$= -\sum_{i=1}^{N} x_{i,2}y_i + w_1 \sum_{i=1}^{N} x_{i,1}x_{i,2} + w_2 \sum_{i=1}^{N} x_{i,2}^2$$

$$= -\beta_2 + w_1\alpha_{12} + w_2\alpha_2 \quad (2)$$

from (1): $$w_1 = \frac{\beta_1 - w_2\alpha_{12}}{\alpha_1} \quad (*)$$

(*) into (2) $$-\beta_2 + \frac{\alpha_{12}\beta_1 - w_2\alpha_{12}^2}{\alpha_1} + w_2\alpha_2 = 0$$

$$w_2\left(\alpha_2 - \frac{\alpha_{12}^2}{\alpha_1}\right) = \beta_2 - \frac{\alpha_{12}\beta_1}{\alpha_1}$$

$$w_2 \frac{\alpha_1\alpha_2 - \alpha_{12}^2}{\alpha_1} = \frac{\beta_2\alpha_1 - \alpha_{12}\beta_1}{\alpha_1}$$

$$\boxed{w_2 = \frac{\beta_2\alpha_1 - \alpha_{12}\beta_1}{\alpha_1\alpha_2 - \alpha_{12}^2}}$$

$$w_1 = \frac{\beta_1}{\alpha_1} - \frac{\alpha_{12}\alpha_1\beta_2 - \alpha_{12}^2\beta_1}{\alpha_1\alpha_2 - \alpha_{12}^2} \cdot \frac{\alpha_{12}}{\alpha_1} = \boxed{\frac{\alpha_2\beta_1 - \alpha_{12}\beta_2}{\alpha_1\alpha_2 - \alpha_{12}^2}}$$

# 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 |

0 - 1.6 - 2.3 - (1+1.6) - 0.3

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

   $H(S) = -\frac{5}{6}\log_2\frac{5}{6} - \frac{1}{6}\log_2\frac{1}{6} = \frac{5}{6}\times 0.3 + \frac{1}{6}\times 2.6$

   $= 0.683$

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

   $\text{Gain}(S, age) = H(S) - \frac{2}{3}(-\log 1 - \log 0 \times 0) - \frac{1}{3}(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2})$

   $= 0.683 - 0 - \frac{1}{3} \times 1 = 0.35$

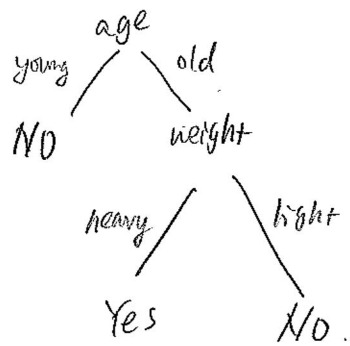
3. (3 pts) What is the information gain if we partition the data on the attribute *Weight*?

   $\text{Gain}(S, weight) = H(S) - \frac{1}{2}(-\log 0 - \log 1) - \frac{1}{2}(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3})$

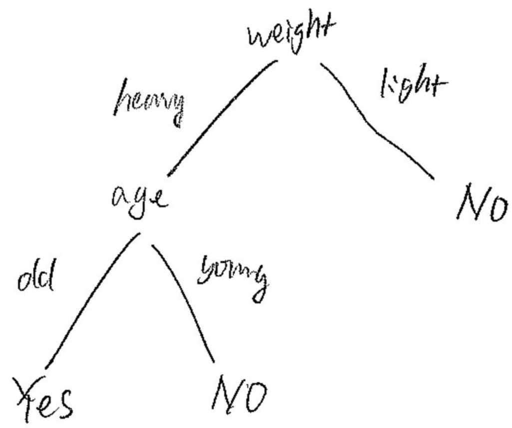
   $= H(S) - 0 - \frac{1}{2}(\frac{1}{3}\times 1.6 + \frac{2}{3}\times 0.6)$

   $= 0.683 - 0.46667 = 0.217$

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 $\alpha$

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___ **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, because that way weight is never updated.

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)

define $W^{t+1} \stackrel{\Delta}{=} W^t + \alpha y_i X_i$, we want $y_i (W^{t+1})^T X_i \geq 0$.

$$y_i (W^t + \alpha y_i X_i)^T X_i \geq 0.$$

$$(y_i W^t)^T X_i + \alpha y_i^2 X_i^T X_i \geq 0.$$

$$\alpha \|X_i\|^2 \geq -y_i (W^t)^T X_i$$

$$\alpha \geq -\frac{y_i (W^t)^T X_i}{\|X_i\|^2}$$



we can choose any learning rate that satisfies the above condition

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 $\alpha$. 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 $w^t \cdot u \geq t\alpha\gamma$.

$w^{j+1} \cdot u = w^j \cdot u + \alpha y_i x_i \cdot u \geqslant w^j \cdot u + \alpha\gamma$

this is valid because we know $y_i(u^T x_i) \geqslant \gamma$

by induction, we get $\boxed{w^t \cdot u \geqslant t\alpha\gamma}$

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

$||w^{j+1}||^2 = ||w^j + \alpha y_i x_i||^2 = ||w^j||^2 + 2\alpha y_i x_i + \alpha^2 ||x_i||^2$ (*)

when it makes mistake. $y_i x_i \leqslant 0$,

and by definition $||x_i||^2 \leqslant R^2$

$\Rightarrow ||w^{j+1}||^2 \leqslant ||w^j||^2 + \alpha^2 R^2$

by induction, $\boxed{||w^t||^2 \leqslant t\alpha^2 R^2}$ D

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

from (*)

$t\alpha^2 R^2 \geqslant ||w^t||^2 \geqslant (w^t \cdot u)^2 \geqslant (t\alpha\gamma)^2$

$\Rightarrow t\alpha^2 R^2 \geqslant t^2 \alpha^2 \gamma^2$

$\boxed{t \leqslant \frac{R^2}{\gamma^2}}$

$(w^t \cdot u)^2$

$= ||w^t||^2 \cdot ||u||^2 \cdot \cos^2\theta$

$= ||w^t||^2 \cdot \cos^2\theta \leq ||w^t||^2$ (*)

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

No, because $\alpha$ is not a parameter in my final expression.

# 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 $\sigma_\gamma(w^T x)$? 1

2. (3 pts) When $w^T x \to -\infty$, what is the value of $\sigma_\gamma(w^T x)$? 0

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

   $\sigma_\gamma(w^T x) = \frac{1}{2}$

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

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

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

   $P(y=1|x) = \frac{1}{2}$

   $\frac{1}{1+\exp(-w^T x/\gamma)} = \frac{1}{2}$

   $\exp(-w^T x/\gamma) = 1$

   $-w^T x/\gamma = 0$

   $\frac{w^T x}{\gamma} = 0$

   $\boxed{w^T x = 0}$

   the decision boundary is still a linear function for any $\gamma$

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

   $P(y=-1|x) = \sigma_\gamma(-w^T x)$

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.

   $\max_w \prod_{i=1}^{N} \sigma_\gamma(w^T x) = \max_w \prod_{i=1}^{N} \frac{1}{1+\exp(-w^T x_i/\gamma)}$

   take log: $\max_w \sum_{i=1}^{N} -\log(1+\exp(-w^T x_i/\gamma))$

   $= \min_w \sum_{i=1}^{N} \log(1+\exp(-w^T x_i/\gamma))$

# 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 $P(x_1, x_2, \ldots, x_N)$.

$$P(x_1, \cdots, x_N) = \lambda^N exp(-\lambda x_1 - \lambda x \cdots - \lambda x)$$
$$= \lambda^N exp(-\lambda \sum_{i=1}^{N} x_i)$$

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

$$N \log \lambda - \lambda \sum_{i=1}^{N} x_i$$

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

take derivative of $\lambda$: $\frac{N}{\lambda} - \sum_{i=1}^{N} x_i = 0$

$$\boxed{\lambda = \frac{N}{\sum_{i=1}^{N} x_i}}$$

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.