- 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: Hongyi Xie

UID: 604998533

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. $N^2 E[X]$ ☐ 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:

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

$$L(w) = \frac{1}{2} \sum_{i=1}^{N} \left(y_i - (w_1 x_{i,1} + w_2 x_{i,2}) \right)^2$$

$$\frac{\partial L}{\partial w_1} = -\sum_{i=1}^{N} \left(y_i - (w_1 x_{i,1} + w_2 x_{i,2}) \right) x_{i,1} = 0$$

$$\frac{\partial L}{\partial w_2} = -\sum_{i=1}^{N} \left(y_i - (w_1 x_{i,1} + w_2 x_{i,2}) \right) x_{i,2} = 0$$

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

$$W_1 = \frac{\beta_1 - W_2 \alpha_{12}}{\alpha_1}$$

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

$$W_1 = \frac{\beta_2 - W_2 \alpha_2}{\alpha_{12}}$$

$$\Rightarrow \quad \frac{\beta_1 - W_2 \alpha_{12}}{\alpha_1} = \frac{\beta_2 - W_2 \alpha_2}{\alpha_{12}}$$

$$\alpha_{12} \beta_1 - \alpha_{12}^2 W_2 = \alpha_1 \beta_1 - \alpha_1 \alpha_2 W_2$$

$$W_2 = \frac{\alpha_{12} \beta_1 - \alpha_1 \beta_1}{\alpha_{12}^2 - \alpha_1 \alpha_2}$$

$$W_1 = \frac{\beta_1 - \alpha_{12} \frac{\alpha_{12} \beta_1 - \alpha_1 \beta_1}{\alpha_{12}^2 - \alpha_1 \alpha_2}}{\alpha_1}$$

3

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: $H(S) = -\sum_{v=1}^{K} P(S = a_v) \log_2 P(S = a_v)$

   $$H(Diabetes) = -\frac{10}{12} \log_2 \frac{10}{12} - \frac{2}{12} \log_2 \frac{2}{12}$$
   $$= -\frac{5}{6} \log_2 \frac{5}{6} - \frac{1}{6} \log_2 \frac{1}{6}$$

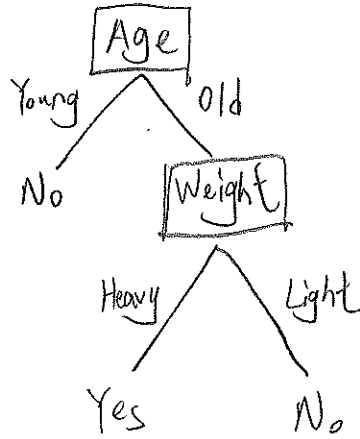2. (3 pts) What is the information gain if we partition the data on the attribute *Age*?
   Hint: $Gain(S, A) = H(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} H(S_v)$.

   $$G(Diabetes, Age) = H(Diabetes) - \frac{8}{12} \times 0 - \frac{4}{12}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right)$$
   $$= -\frac{5}{6}\log_2\frac{5}{6} - \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{6}\log_2\frac{1}{2} + \frac{1}{6}\log_2\frac{1}{2}$$
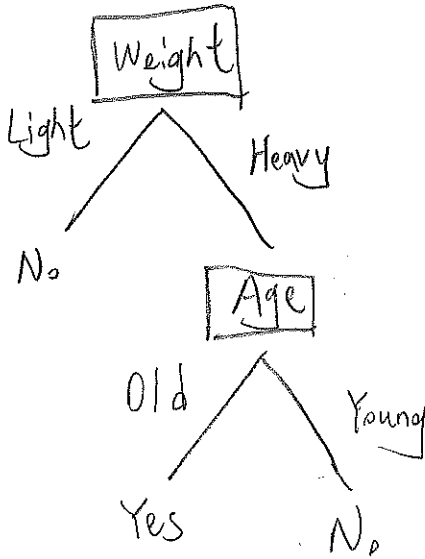
3. (3 pts) What is the information gain if we partition the data on the attribute *Weight*?

   $$G(Diabetes, Weight) = H(Diabetes) - \frac{1}{2}\left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right)$$
   $$- \frac{1}{2} \times 0$$
   $$= -\frac{5}{6}\log_2\frac{5}{6} - \frac{1}{6}\log_2\frac{1}{6} + \frac{1}{3}\log_2\frac{2}{3} + \frac{1}{6}\log_2\frac{1}{3}$$

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

```
                    ┌─────┐
                    │ Age │
                    └─────┘
            Young  /       \  Old
                  /          \
                No         ┌────────┐
                           │ Weight │
                           └────────┘
                     Heavy /        \ Light
                          /          \
                        Yes           No
```

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

```
                    ┌────────┐
                    │ Weight │
                    └────────┘
            Light  /          \  Heavy
                  /            \
                No          ┌─────┐
                            │ Age │
                            └─────┘
                       Old /       \ Young
                          /         \
                        Yes          No
```

5

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) _____ **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 we do so, our w will never be 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)

$$\begin{cases} y_i W_t^T x_i \leq 0 \\ y_i W_{t+1}^T x_i > 0 \end{cases} \Rightarrow$$

$$W_{t+1} = W_t + \alpha y_i x_i$$

$$\alpha = \frac{W_{t+1} - W_t}{y_i x_i}$$

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

$$w^t = \alpha \sum_{m=1}^{t} y_m x_m \quad \leftarrow (x_m, y_m) \text{ refers to pairs where we made mistake}$$

$$w^t \cdot u = \alpha \sum_{m=1}^{t} y_m (u^T x_m) \geq \alpha t \gamma \quad \text{since } y_m(u^T x_m) \geq \gamma , \forall m.$$

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

$$||w^t||^2 \leq \alpha^2 ||\sum_{m=1}^{t} y_m x_m||^2$$

$$\leq \alpha^2 \sum_{m=1}^{t} ||y_m x_m||^2$$

$$\leq \alpha^2 t R^2 \quad \text{since } ||y_m x_m|| \leq R, \forall m.$$

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

The mistake bound is $\frac{R^2}{\gamma^2}$

We want $(w^t \cdot u)^2 = ||w^t||^2$

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

$$t = \frac{R^2}{\gamma^2}.$$

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

   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 $\gamma \to \infty$?

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

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.

   $$\frac{1}{1 + e^{-\frac{w^T x}{\gamma}}} \geq \frac{1}{2} \qquad -\frac{w^T x}{\gamma} \leq 0$$

   $$2 \geq 1 + e^{-\frac{w^T x}{\gamma}} \qquad \frac{w^T x}{\gamma} \geq 0 \leftarrow \text{This is a}$$

   $$e^{-\frac{w^T x}{\gamma}} \leq 1 \qquad\qquad \text{linear function.}$$

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

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

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.

   $$\text{Find } w \text{ s.t. } \sum_{i=1}^{N} \sigma_\gamma(w^T x_i)(1 - y_i) + (1 - \sigma_\gamma(w^T x_i))y_i \text{ is minimized}$$

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

$$P(x_1, x_2, \cdots x_N) = \lambda^N \exp\left(-\lambda \sum_{i=1}^{N} x_i\right)$$

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

$$\log\left(\lambda^N \exp\left(-\lambda \sum_{i=1}^{N} x_i\right)\right) = N \log \lambda - \lambda \sum_{i=1}^{N} x_i$$

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

$$\frac{\partial \log \text{likelihood}}{\partial \lambda} = 0$$

$$N \log \lambda - \lambda \sum_{i=1}^{N} x_i = 0$$

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

$$\lambda e^{\frac{1}{\lambda}} = \exp\left(\frac{1}{N} \sum_{i=1}^{N} x_i\right)$$

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.