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 10 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_2 900, and test the model on the data of size N_3 100. 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

$\frac{k}{n}$

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:

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

Consider:

$$J(w) = \frac{1}{2} \sum_{i=1}^{N} \left(y_i - (w_1 x_{i,1} + w_2 x_{i,2})\right)^2$$ ⇒ we want to minimize this.

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

Notice there is an closed form solution: $X: \left[\begin{bmatrix} X_{11} \\ X_{12} \end{bmatrix} \cdots \begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix} \right]$ $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_i \end{bmatrix}$

$$w = (XX^T)^{-1} XY = (\alpha_1 + \alpha_2 + 2\alpha_{12})^{-1} XY$$

$$= (\alpha_1 + \alpha_2 + 2\alpha_{12})^{-1} \cdot (\beta_1 + \beta_2)$$

$\frac{26}{60}$ $\frac{26}{60}$

$\frac{\cancel{26}^{13}}{10} \times \frac{1}{\cancel{6}3} \Rightarrow \frac{13}{30} = \frac{26}{60}$ $\log_2 \frac{5}{6} = \log_2 5 - \log_2 6 = 2.3 - 2.6 = 0.3$ $\log_2 \frac{1}{6} \Rightarrow \underbrace{\log_2 1}_{0} - \underbrace{\log_2 6}_{\log_2 2 + \log_2 3 = 1 + 1.6 = 2.6}$

$\frac{1\cancel{3}}{26} \times \frac{31}{10.2} = \frac{1}{4} = \frac{15}{60}$

Decision Tree (15 pts)

$\frac{5}{26} \times \frac{31}{10} = \frac{1}{4}$

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

4

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

2

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

$P(\text{Diabetes}) = \frac{2}{12} = \frac{1}{6}$

$$H(\text{Diabetes}) = -\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6}$$
$$= -\frac{1}{6}(\log_2 1 - \log_2(3\times 2)) - \frac{5}{6}(\log_2 5 - \log_2(3\times 2))$$
$$= -\frac{1}{6}(0 - 1 - 1.6) - \frac{5}{6}(2.3 - 1 - 1.6)$$
$$= \frac{41}{60}$$

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

$H(\text{Age}=\text{Young}) = 0$ (All is No.) $H(\text{Age}=\text{Old}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$$\text{Gain}(\text{Diabetes}, \text{Age}) = \frac{41}{60} - 0 \times \frac{8}{12} - 1 \times \frac{[illegible]}{12} = \frac{31}{60}$$

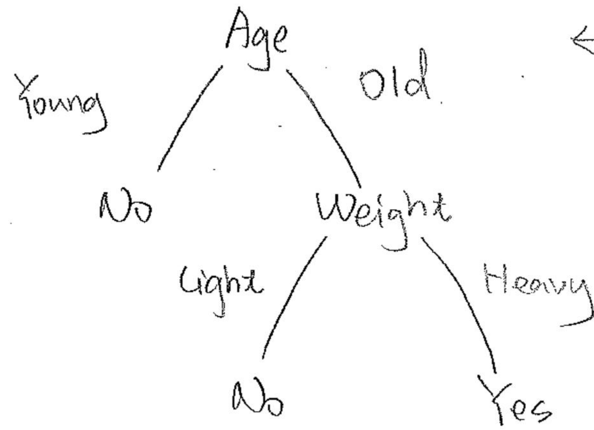
3. (3 pts) What is the information gain if we partition the data on the attribute *Weight*?

$H(\text{Weight} = \text{Light}) = 0$ (All is No.)

$$H(\text{Weight} = \text{Heavy}) = \boxed{-\frac{4}{6}\cdot\log_2\frac{4}{6} - \frac{2}{6}\cdot\log_2\frac{2}{6}} = -\frac{2}{3}\cdot\log_2\frac{2}{3} - \frac{1}{3}\cdot\log_2\frac{1}{3}$$
$$= \frac{14}{15}$$

$$\text{Gain}(\text{Diabetes}, \text{Weight}) = \frac{41}{60} - 0 \times \boxed{\frac{6}{12}} - \frac{14}{15} \times \boxed{\frac{6}{12}}$$
$$= \frac{41}{60} - \frac{28}{60} \quad 4$$
$$= \frac{13}{60}$$

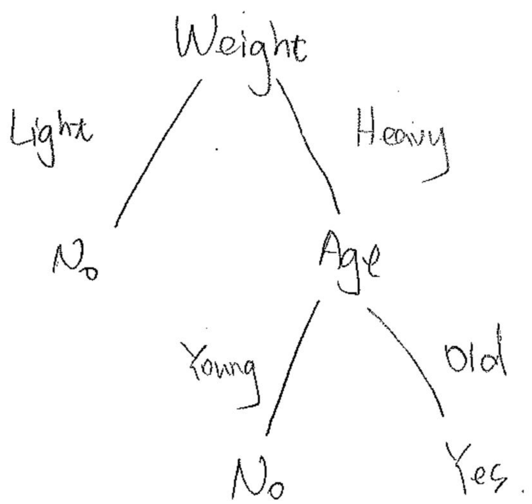
4. (3 pts) Apply the ID3 algorithm to build the tree using both features Age and Weight.



Maximize information gain.

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

Want training error = 0



This also has training error 0.

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. In that way we will never update w, so our final w returned would be $\vec{0}$.

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)

Line 3 is true:
$y_i w^T x_i \leq 0$.

After update.
$w' = w + \alpha y_i x_i$.

We want $y_i w'^T x_i > 0$.

$$\Rightarrow y_i (w + \alpha y_i x_i)^T x_i > 0$$
$$\Rightarrow y_i w^T x_i + \alpha y_i^2 x_i^T x_i > 0$$
$$\Rightarrow y_i w^T x_i + \alpha x_i^T x_i > 0 \qquad (y_i^2 = 1)$$
$$\Rightarrow \alpha \|x_i\|^2 > -y_i w^T x_i$$
$$\Rightarrow \boxed{\alpha > \frac{-y_i w^T x_i}{\|x_i\|^2}}.$$

Prove of Correctness:
Consider $y_i \left(w + \frac{-y_i w^T x_i}{\|x_i\|^2} \cdot y_i x_i\right)^T x_i = y_i w^T x_i + 6 y_i \frac{-y_i^2 w^T x_i}{\|x_i\|^2} \cdot x_i^T x_i = y_i w^T x_i - y_i^3 w^T x_i$
$= y_i w^T x_i - y_i w^T x_i = 0$. So any $\alpha > \frac{-y_i w^T x_i}{\|x_i\|^2}$ will make
$y_i (w + \alpha y_i x_i)^T x_i > 0$.

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

Consider $w^{t+1} = w^t + \alpha \cdot y_i \cdot x_i \Rightarrow w^{t+1} \cdot \vec{u} = w^t \cdot u + \alpha \cdot y_i (u^T x_i)$

$\Rightarrow w^{t+1} \cdot u \geq w^t \cdot u + \alpha\gamma$. Since $w^0 = 0$, by recursion

we have: $w^t \cdot u \geq \vec{0} + t \cdot \alpha\gamma \Rightarrow w^t \cdot u \geq t\alpha\gamma$.

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

$||w^{t+1}||^2 = ||w^t + \alpha \cdot y_i x_i||^2 = ||w^t||^2 + 2\alpha \cdot y_i w^{t^T} x_i + ||\alpha \cdot y_i x_i||^2$

$= ||w^t||^2 + 2\alpha \cdot y_i w^{t^T} x_i + \alpha^2 ||x_i||^2$ Since $2\alpha \cdot y_i w^{t^T} x_i < 0$:

$||w^{t+1}||^2 \leq ||w^t||^2 + \alpha^2 R^2$. Since $w^0 = 0$, by recursion

We have: $||w^t||^2 \leq ||0||^2 + t \cdot \alpha^2 R^2 \Rightarrow ||w^t||^2 \leq t\alpha^2 R^2$

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

$w^t \cdot u \geq t\alpha\gamma \Rightarrow ||w^t \cdot u||^2 \geq t^2\alpha^2\gamma^2 \Rightarrow ||w^t||^2 \geq t^2\alpha^2\gamma^2$

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

$\Rightarrow t \leq \frac{R^2}{\gamma^2}$

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

$w^T x \to \infty$ $\quad -w^T x \to -\infty$ $\quad (-w^T x/r) \to -\infty$ $\quad \exp(-w^T x/r) \to 0.$

$\sigma_r(w^T x) \to 1$

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

$w^T x \to -\infty$ $\quad -w^T x \to \infty$ $\quad (-w^T x/r) \to \infty$ $\quad \exp(-w^T x/r) \to \infty$

$\sigma_r(w^T x) \to 0.$

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

$r \to \infty.$ $\quad \exp(-w^T x/r) \to 1$ $\quad \sigma_r(w^T x) \to \frac{1}{2}$

$-w^T x/r \to 0$

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

$r \to 0$ $\quad \exp(-w^T x/r) \to \infty$ $\quad \sigma_r(w^T x) \to 0.$

$-w^T x/r \to \infty$

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

Decision boundary: $\sigma_r(w^T x) = \frac{1}{2}$

$\Rightarrow 1 + \exp(-w^T x/r) = 2 \Rightarrow \exp(-w^T x/r) = 1$

$\Rightarrow \frac{-w^T x}{r} = 0.$ Since $r > 0$, we have $-w^T x = 0$.

The boundary is linear: $-w^T x = 0$.

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

$P(y = -1|x) = 1 - \sigma_r(w^T x) = \dfrac{1}{1 + \exp(w^T x/r)}$

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.

Let $\sigma'_r(w^T x)$ be $\dfrac{1}{1 + \exp(-y w^T x/r)}$

$\Rightarrow \text{Max} \prod_{i=1}^{N} P(y = y_i | x_i) \Rightarrow \text{Max} \log \prod_{i=1}^{N} (\sigma'_r(w^T x))$

$\Rightarrow \text{Max} \sum_{i=1}^{N} \log(\sigma'_r(w^T x))$

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

$$P(x_1, x_2, \cdots, x_N) = \lambda^N \exp(-\lambda x_1) \cdots \exp(-\lambda x_N)$$
$$= \lambda^N \exp(-\lambda(x_1 + x_2 + \cdots + x_N))$$

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

$$\log(P(x_1, x_2, \ldots, x_N)) = \log \lambda^N + \log(\exp(-\lambda(x_1 + \cdots + x_N)))$$
$$= \log \lambda^N - \lambda(x_1 + x_2 + \cdots + x_N)$$

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

$$\frac{\partial \log(P(x_1, x_2, \ldots, x_N))}{\partial \lambda} = \frac{\partial \log \lambda^N}{\partial \lambda} - \frac{\partial \lambda(x_1 + x_2 + \cdots + x_N)}{\partial \lambda}$$
$$= \frac{N \cdot \lambda^{N-1}}{\lambda^N} - (x_1 + x_2 + \cdots + x_N)$$
$$= \frac{N}{\lambda} - (x_1 + x_2 + \cdots + x_N) = 0$$

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

$$\Rightarrow N = \lambda \cdot \sum_{i=1}^{N} x_i \quad \Rightarrow \lambda = \frac{N}{\sum_{i=1}^{N} x_i}$$

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.