1. Consider the following training set, in which each example has two tertiary attributes (0, 1, or 2) and one of two possible classes (X or Y).

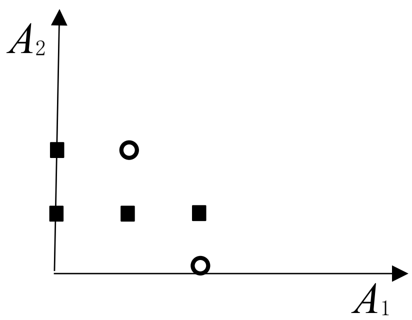
Example	A_1	A_2	Class
1	0	1	X
2	2	1	X
3	1	1	X
4	0	2	X
5	1	2	Y
6	2	0	Y

1) What feature would be chosen for the split at the root of a decision tree using the information gain criterion? Show the details. (Note: we split attributes at each value of the attributes, for example, A_1=0,A_1=1,A_1=2)

2) What would the Naïve Bayes algorithm predict for the class of the following new example? Show the details of the solution.

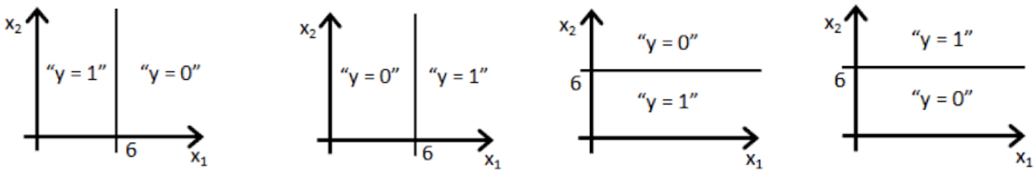
Example	A_1	A_2	Class
7	2	2	?

3) Draw the decision boundaries for the nearest neighbor algorithm assuming that we are using standard Euclidean distance to compute the nearest neighbors.



4) Which of these classifiers will be the least likely to classify the following data points correctly? Please explain the reason.
   a. ID3.
   b. Naïve Bayes
   c. Logistic Regression
   d. KNN

2. You have trained a logistic classifier y=sigmoid($w_0+w_1x_1+w_2x_2$). Suppose w_0=6, w_1=-1, and w_2=0. Which of the following figures represents the decision boundary found by your classifier?



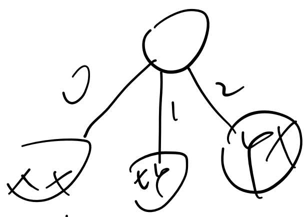
A B C D

3. Suppose we are given a dataset $D=\{(x^{(1)},r^{(1)}),\ldots,(x^{(N)},r^{(N)})\}$ and aim to learn some patterns using the following algorithms. Match the update rule for each algorithm.

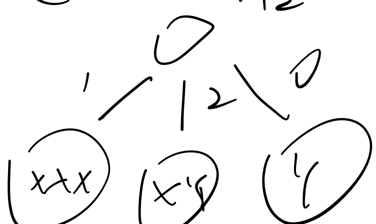
Algorithms:	**Update Rules:**
A: SGD for Logistic Regression $y = \text{sigmoid}\,(\boldsymbol{w}^T x)$	***1.*** $w_t \leftarrow w_t + (w_t^T x^{(l)} - r^{(l)})$
B: Least Mean Squares for Linear Regression $y = \boldsymbol{w}^T x$	***2.*** $w_t \leftarrow w_t + \dfrac{1}{1+\exp \eta(y^{(l)}-r^{(l)})}$
C: Perceptron $y = \text{sign}(\boldsymbol{w}^T x)$ (where sign(a)=1 if a>0 else -1)	***3.*** $w_t \leftarrow w_t + \eta(y^{(l)} - r^{(l)})x_i^{(l)}$

1. (1) we would choose A_2.

if we choose A_1



choose A_2



$H(D) = -(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3})$

$= \frac{2}{3}(\log 3 - \log 2) + \frac{1}{3}\log 3 = \log 3 - \frac{2}{3}\log 2$

$H(D|A_1) = -\frac{1}{3}\cdot 0 - \frac{1}{3}\cdot(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}) - \frac{1}{3}(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2})$

$= \frac{2}{3}\log 2$

$H(D|A_2) = -\frac{1}{2}\cdot 0 - \frac{1}{6}\cdot 0 - \frac{1}{3}(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}) = \frac{1}{3}\log 2$

Example	A_1	A_2	Class
1	0	1	X
2	2	1	X
3	1	1	X
4	0	2	X
5	1	2	Y
6	2	0	Y

$Gain(D, A_1) = \log 3 - \frac{2}{3}\log 2 - \frac{2}{3}\log 2 = \log 3 - \frac{4}{3}\log 2$

$Gain(D, A_2) = \log 3 - \frac{2}{3}\log 2 - \frac{1}{3}\log 2 = \log 3 - \log 2$

Therefore, we should split based on A_2.

(2)

A_1	A_2	Class
2	2	

$$P(X \mid A_1 = 2, A_2 = 2) = \frac{P(A_1=2, A_2=2 \mid X)\, P(X)}{P(A_1=2, A_2=2)}$$

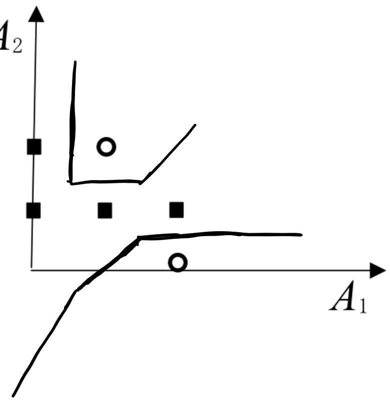
$$P(Y \mid A_1 = 2, A_2 = 2) = \frac{P(A_1=2, A_2=2 \mid Y)\, P(Y)}{P(A_1=2, A_2=2)}$$

$$P(A_1=2 \mid X)\, P(A_2=2 \mid X) P(X) = \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{2}{3} = \frac{1}{24}$$

$$P(A_1=2 \mid Y)\, P(A_2=2 \mid Y)\, P(Y) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{12}$$

Therefore, example 7 is Y based on Naive Bayes.

(3):



(4): Logistic Regression will be least likely to classify the points. Because the data cannot be classified by a linear classifier.

2. $y = \text{Sigmoid}(6 - x_1 + 0 \cdot x_2) = \frac{1}{1+\exp[-(6-x_1)]}$

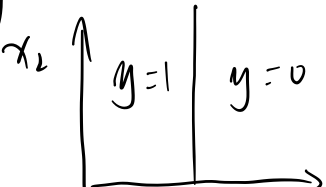
Decision boundary is $\log \frac{y}{1-y} = 6 - x_1$

$x < 6 \quad \log \frac{y}{1-y} > 0 \Rightarrow \frac{y}{1-y} > 1$

$\Rightarrow y > 1-y$

$\Rightarrow y > 0.5$

$\log(C_1 \mid x) > 0.5$



"A"

x_1

3. $y = \text{Sigmod}(w^T x)$ update rules: $w_t \leftarrow w_t - \eta (r^{(l)} - y^{(l)}) x_i^{(l)}$

A - 3

$y = w^T x$ if we use SGD: $w_t \leftarrow w_t + \eta (r^{(l)} - y^{(l)}) x_i^{(l)}$

B - 3

C - 3

$w \leftarrow w + \eta r^{(l)} x^{(l)}$