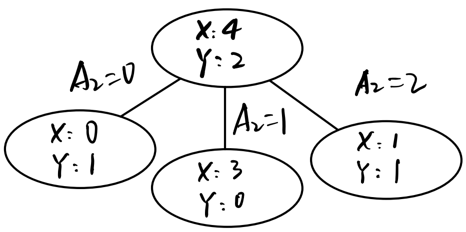
1. Consider the following training set, in which each example has two tertiary attributes (0, 1, or 2) and one of two possible classes (*X* or *Y*).

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, $A_1=0, A_1=1, A_1=2$)



Split by A_2:

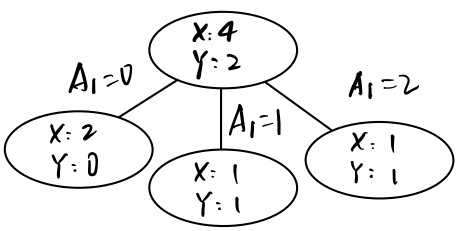
Root entropy: $H(D) = -\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6} = 0.92$

Leaves entropy: $H(D|A_2=0) = 0$, $H(D|A_2=1) = 0$

$H(D|A_2=2) = 1$.

$H(D|A_2) = \frac{1}{6}\times 0 + \frac{1}{2}\times 0 + \frac{2}{6}\times 1 = 0.33$

$IG(D|A_2) = 0.92 - 0.33 = 0.59$



Split by A_1:

Leaves entropy: $H(D|A_1=0) = 0$, $H(D|A_1=1) = 1$

$H(D|A_1=2) = 1$.

$H(D|A_1) = \frac{1}{3}\times 0 + \frac{1}{3}\times 1 + \frac{1}{3}\times 1 = 0.67$

$IG(D|A_1) \approx 0.92 - 0.67 = 0.25 < 0.59$

Thus, A_2 would be chosen for split the root.

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

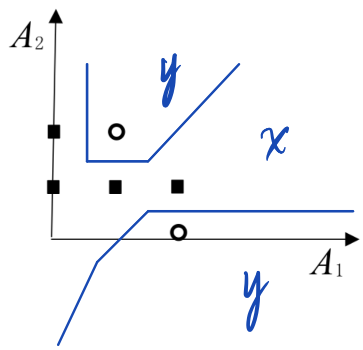
Example	A_1	A_2	Class
7	2	2	?

$$P(X|A_1=2, A_2=2) = \frac{P(A_1=2, A_2=2|X)\cdot P(X)}{P(A_1=2, A_2=2)} = \frac{P(A_1=2|X)\cdot P(A_2=2|X)\cdot P(X)}{P(A_1)\cdot P(A_2)} = \frac{\frac{1}{4}\times\frac{1}{4}\times\frac{2}{3}}{\frac{2}{3}\times\frac{1}{4}\times\frac{1}{4}+\frac{1}{3}\times\frac{1}{2}\times\frac{1}{2}} = 0.33$$

$$P(Y|A_1=2, A_2=2) = \frac{P(A_1=2, A_2=2|Y)\cdot P(Y)}{P(A_1=2, A_2=2)} = \frac{\frac{1}{2}\times\frac{1}{2}\times\frac{1}{3}}{\frac{2}{3}\times\frac{1}{4}\times\frac{1}{4}+\frac{1}{3}\times\frac{1}{2}\times\frac{1}{2}} = 0.67$$

Hence, this example will be predicted as Y.

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

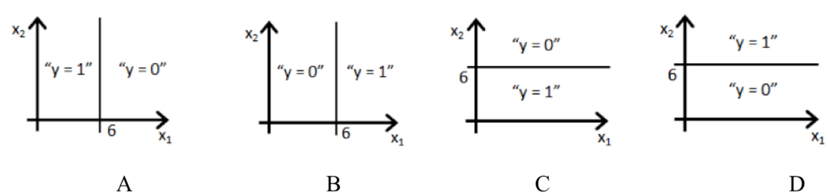


4) Which of these classifiers will be the least likely to classify the following data points correctly? Please explain the reason.

a. ID3.
b. Naïve Bayes
c. Logistic Regression
d. KNN

c、因为这是一个线性不可分问题(从上一题的图像可以看出来)。逻辑回归没有办法处理线性不可分问题。

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?



$y = \text{sigmoid}(6 - x_1)$

if $6 - x_1 > 0 \rightarrow x_1 < 6,\ y = 1$

if $6 - x_1 < 0 \rightarrow x_1 > 6,\ y = 0$

So A is right.

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

A: $y = \frac{1}{1+\exp(-w^T x)}$

B: $\theta_j = \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})\, x_j^{(i)}$ (j for $0 \sim n$)