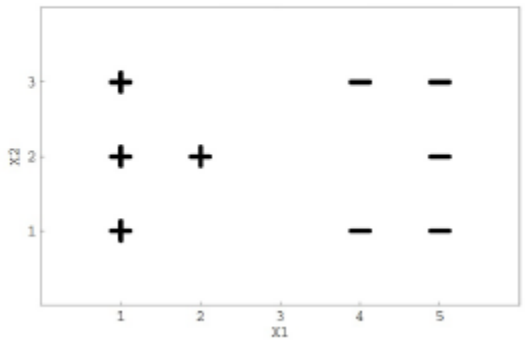
Machine Learning, Spring 2020: Homework 3

1. **[20 pts/2pts each] True/False** (Please add a 1 sentence providing justification for your answer)
   a. True/False. When we have less data and the model complexity stays the same, overfitting is more likely.
   b. True/False. When our data points have fewer features, overfitting is more likely.
   c. True/False. Nearest Neighbors is more efficient at test time than logistic regression.
   d. True/False. Nearest Neighbors is more efficient at training time than logistic regression.
   e. True/False. The depth of a learned decision tree can be larger than the number of training examples used to create the tree.
   f. True/False. A large learning rate is always preferred.
   g. True/False. Information gain is the only way to calculate feature importance.
   h. True/False. A benefit of a decision tree is it can handle noisy or missing values within training data.
   i. True/False. Decision trees can produce non-linear decision boundaries.
   j. True/False. It is not possible to overfit a model if we use cross-validation.

2. **[20 pts] SVM**

1. [10 points] Suppose we are using a linear SVM, with some large margin value, and you are given the following data set.



Draw the decision boundary of linear SVM. Give a brief explanation.

2. [10 points] In the above image, circle the point such that removing that example from the training set and retraining SVM, we would get a different decision boundary than training on the full sample.

3. [30 pts] Naïve Bayes

1. [5 points] Why is *naive Bayesian classification* called "naive"?
2. [25 points] Use a Naive Bayesian classifier to classify, the input

$$X = (age = youth, income = low, student = yes, credit\_rating = fair),$$

given the following table:

ID	age	income	student	credit-rating	Class:buys-computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle-aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle-aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle-aged	medium	no	excellent	yes
13	middle-aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

4. [30 pts] K-Means

Suppose that the task is to cluster points (with (x, y) representing location) into three clusters, where the points are:

$$A_1(2, 10), A_2(2, 5), A_3(8, 4), B_1(5, 8), B_3(6, 4), C_1(1, 2), C_2(4, 9).$$

The distance function is Euclidean distance. Suppose initially we assign A_1,B_1, and C_1 as the center of each cluster, respectively.

1. [10 points] Use the k-means algorithm to show only the three cluster centers after the first round of execution.
2. [20 points] The final three clusters.