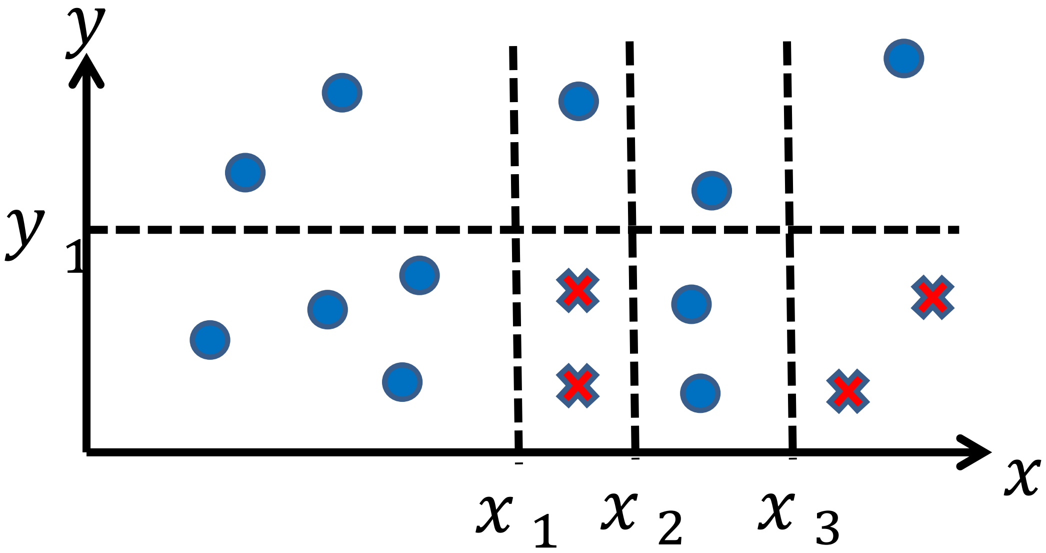
HW #2 Due: 11/8/2020

Problems with paper and pen

1. Prove that a rectangle decision function has a VC dimension of 4 by enumerating all possible sample distributions. You can reasonably assume no two (or more) samples are on the same vertical or horizontal line.
2. We mention the AIDS detection problem in Bayesian decision theory. Use Bayes theorem to confirm the given answer. To answer this problem, you need to distinguish two different conditions:
    - False positive is a conditional probability *P*(reagent is negative | patient is infected). Same argument for false negative.
    - When a patient is given a positive test result, it is actually *P*(patient is infected | reagent is positive)
3. We mention an example to use Naïve Bayesian classifier for classifying colored squares and circles in the lecture. Following the example, which class will a red circle ● be assigned to?
4. Plot a decision tree for the following data points. If you carefully design your tree, you will just need to use one “>” or “<” in a vertex.



5. We mentioned the gambler’s ruin chain in the lecture. If the gambler decides to bet different amount of money on each bet, which of the following is a better strategy to survive longer (assuming the gambler has a finite amount of money):

    (a) Bet more money next time if he/she won last time, and bet less money next time if he/she lost last time.

    (b) Bet less money next time if he/she won last time, and bet more money next time if he/she lost last time. Hint: If you are unable to figure out the answer, follow the concept of the Kelly Criterion.

Computer-based problems

Problems 6 to 8 are based on the following dataset: Breast Cancer Wisconsin (Original) Data Set, available at https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

6. Use the Naïve Bayes classifier for the classification task of the cancer dataset. Do a 70/30 split for training and test set. Repeat the trials 10 times and compute the average accuracy. (Note: Use GaussianNB because the features are continuous numbers)
7. Repeat problem 6, but use GMM with 3 mixtures instead.
8. In this problem, you are asked to perform the wrapper-type feature selection using the *k*-NN with *k* = 3 for cancer dataset. To simplify the problem, you just need to select 3 attributes out of 9. To begin one experiment, randomly draw 60 % of the instances from each class for training, and 20% from each class for finding the best 3 attributes. Once the feature selection is complete, use the rest 20% for testing to obtain the accuracy. Remember to use only the three chosen attributes for *k*-NN (*k* = 3) to classify. Repeat the experiments 10 times and report the average accuracy. You need to deal with **missing attributes**.
9. Write a program for multinomial HMM classification. The training and test sequences are given with the sample program to show how to read the data.
10. Use the data below to construct a CART tree and plot the resultant tree using sklearn. You need to think a way to deal with categorical data. If a particular day is sunny, high temperature, low humidity, and no wind, what is the decision based on your plotted tree?

Outlook	Temperature	Humidity	Windy	Decision
Sunny	Hi	Hi	No	No play
Sunny	Hi	Hi	Yes	No play
Overcast	Hi	Lo	No	Play
Overcast	Lo	Lo	Yes	Play
Rain	Lo	Hi	No	Play
Rain	Lo	Hi	Yes	No play