DSE 210
Final Exam

Q1: Urn A contains 2 white balls and 4 red balls; Urn B contains 8 white balls and 5 red balls; Urn C contains 1 white ball and 3 red balls. P(UrnA)=0.5 and P(UrnB)=P(UrnC). We draw a red ball.

a) What is the probability P(UrnA | red)?
b) P(White)=?

Q2: Consider a set of cardinality n, i.e., the set has n elements.

a) Show that the number of subsets of this set is 2^n (including the empty set).
b) Show that $2^n - 1 = \sum_{k=1}^{n} 2^{n-k}$ (Hint: count the nonempty subsets of a set of cardinality n). You must use combinatorial counting arguments to show this result.

Q3) Let Pr(X=1) = 0.5, Pr(X=2)=0.25, Pr(X=3)=0.25.

a) Calculate E(X)
b) Calculate Var(X)
c) Plot the cumulative density function (CDF) of the random variable X.

Q4) For this problem, you'll be using the 20 Newsgroups data set. There are several versions of it on the web. You should download "20news-bydate.tar.gz" from http://qwone.com/~jason/20Newsgroups/

The same website has a processed version of the data, "20news-bydate-matlab.tgz", that is particularly convenient to use. Download this and also the file "vocabulary.txt". Look at the first training document in the processed set and the corresponding original text document to understand the relation between the two. The words in the documents constitute an overall vocabulary V of size 61188. Build a Bernoulli Naive Bayes model using the training data. Write a routine that uses this naive Bayes model to classify a new document. To avoid underflow, work with logs rather than multiplying together probabilities.

(a) Evaluate the performance of your model on the test data. What error rate do you achieve?

(b) Evaluate your final model on the test data. Construct a confusion matrix.

https://en.wikipedia.org/wiki/Confusion_matrix

Q5) Consider the linear classifier $w \cdot x \geq \theta$, where

$w = [-1\ \ 2]^T$ and $\theta=10$.

Sketch the decision boundary in R^2. Make sure to label precisely where the boundary intersects the coordinate axes, and also indicate which side of the boundary is the positive side.

Q6) Urn A contains a Gaussian pdf: N(0, σ=1); Urn B contains another Gaussian N(5,σ =2); We draw a number and it is X=2.5. P(UrnA)=2P(UrnB)

a) Determine the likelihood of (UrnA | X=2.5)?
b) Determine a decision boundary (Urn A vs Urn B) for this problem.

Q7) Consider the following observations:

X=(-0.1,-0.2, 0.1, 0.2, 0, 0.1, -0.1, 0, -0.05, 0.1, 1.05, 1.1, 0.9, 0.8, 0.9, 1, 1.2, 1.1,1.2, .9)

Cluster this data into two classes using the K-means algorithm. What are the cluster centers?

Q8) Construct a Gaussian Mixture Model for the above data. What is your estimate for the number of mixtures?

Q9) Worksheet 9, Question 10

Q10) Worksheet 10, Question 12

.