


HW #2 & 3 Due: 4/9/2021

1. A contest has 1,000 classes of images to be recognized. Between discriminative and generative models, which model is more suitable for this contest? Why?
2. 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?
3. We mention the condition number of a matrix in the lecture. Compute the condition number for the following matrix and comment if it is ill-conditioned or not.

$$A = \begin{bmatrix} 0.540 & 0.323 \\ 0.647 & 0.387 \end{bmatrix}$$

4. Follow the numerical example in GMM and complete the computation of μ_2 , σ_1^2 , σ_2^2 , α_1 , and α_2 in one step.
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.

6. If a sequence of coin tossing has the results of H,H,T,T,H,H,H, follow the MAP method given in the ppt file to **numerically** (i.e., not from the equation) find θ_{MAP} based on a likelihood plot for $a = b = 5$. You need to include your program to show how the likelihood plot is drawn.
7. We can also use the iris dataset for regression work. Consult the Internet to learn how to use a k -NN regressor, and use the first three features (sepal length, sepal width, petal length) in each sample as inputs to predict the fourth feature (petal width). To conduct one trial, again you need to divide the dataset into a training set (70%) and a test set (30%). To simplify the problem, build one model for all three classes. Repeat 10 trials and report the average MSE. Use `weights='distance'` in this problem for sk-learn.
8. Use the Naïve Bayes classifier for the classification task of the iris dataset. Do a 70/30 split for training and test set. Repeat the trials 10 times and compute the average accuracy. (Note: Use GaussianNB in sk-learn because the features are continuous numbers)
9. Follow problem 8 but use the GMM model to classify the iris dataset with 2

mixtures instead.

10. In this problem, you are asked to perform the wrapper-type feature selection using the k -NN with $k = 3$ for Breast Cancer Wisconsin (Original) Data Set, at <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29>. To simplify the problem, you just need to select 3 attributes out of 9. To begin one experiment, randomly draw 60 % of the samples 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 use imputation to deal with **missing attributes**.