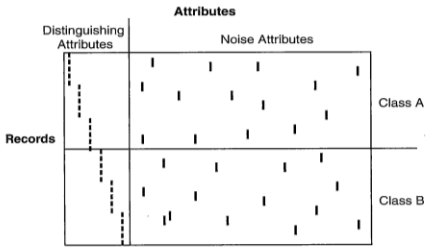
Homework 3, Due on Feb 21, 2025 (25 points in total)

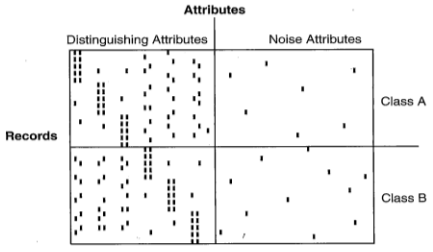
Reading Assignment for the book of Tan et al.: page 213 – 249 about Naïve Bayesian and Logistical Regression, and Page 276-296 about Support Vector Machines (SVM).

Reading and hands-on exercise assignment for the Geron's book, Chapter 4 Training Models, and Chapter 5 Support Vector Machines (page 111 – 172).

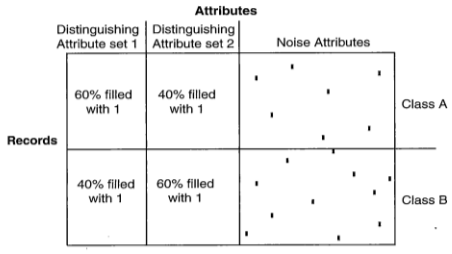
1. (3 points) Suppose the fraction of undergraduate students who smoke is 15%, and the fraction of graduate students who smoke is 23%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who smokes is a graduate student?
2. (3 points) Problem 7 pages 348 (old book Page Problem 7, page 318)
3. (3 points) Problem 14 pages of 352 (old book Page 322 of the textbook, #15)
4. (5 points) Given the descriptive meta data sets shown in the figure below (if you have trouble understand the distribution patterns of the hypothetic features, please ask me at office hours), explain how the decision tree, naïve Bayes, and k-nearest neighborhood, and logistical regression classifers would perform on these data sets (i.e., which one you think will have the best, which one has the worst performance and why based on the summary of the strength and weakness of the characteristics of the six synthetic datasets).



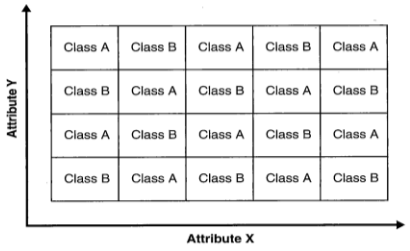
(a) Synthetic data set 1.



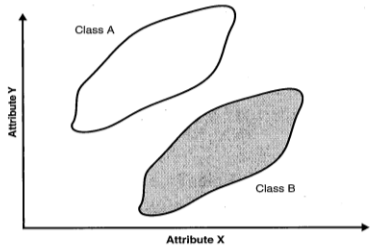
(b) Synthetic data set 2.



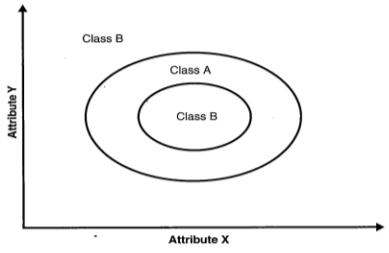
(c) Synthetic data set 3.



(d) Synthetic data set 4.



(e) Synthetic data set 5.



(f) Synthetic data set 6.

5. (4 points) The dataset of Universal Bank includes data on 5000 customers. The data includes the customer's response to the last personal loan campaign as well as the customer's demographic information (Age, Income, etc.), and the customer's relationship with the bank (mortgage, securities, account, etc.). Among the 5000 customers, only 480 (=9.6%) accepted the personal loan offered by the bank in a previous campaign. The goal is to build a model that identifies the customers who are most likely to accept the loan offer (the eighth attribute: PersonalLoan) in future mailings. If you use Weka to load the UniversalBank.csv file, you cannot see any class because every attribute is shown as numerical data. You cannot use this to do any classification unless that you change the response data PersonalLoan into nominal and class. The following four steps help you to change the PersonalLoan into nominal and class as the default of Weka.
   A) In the Preprocessing tab, under filter click "choose," under filter tree, navigate to unsupervised and search for "NumerictoNomial."
   B) Click once, and press Apply. Note: Make sure you have checked the checkbox of your desired field that you want to change.
   C) Save the file and reload it.
   D) Select (right-click) the Personal.Loan variable as class.

   Your task 1: Select 10-folder Cross-Validation, and Try to use Decision Tree and RandomForest models and compare their performances.

   Task 2 Select 10-folder Cross-Validation, and Try to use Logistic Regression under the Function Classifiers. You will find that Weka crash for memory requests exceeded its ability. Now you delete the irrelevant attributes ID and Zip and use the Logistic regression again. You will see that it works though it takes a long time to complete the 10 folders of cross-validation.

6. (3 points) Use the same data of the Universal Bank above
   Task 1: Modify the Jupyter notebook codes in subfolder Pipeline-Master under the homework folder and run Logistic Regression models with GridSearchCV for the bank data to find the best model, compare the performance of Weka in task B of problem 5.
   Task 2: Use the same Bank Data above and run the decision tree model with GridSearchCV to find the optimal levels of tree and RandomForest, and compare the performance with the results in task A of problem 5.

7. (4 points) The sonar data sets are linked at sonar_train.csv, and sonar_test.csv for training and testing.
   a) Use KNN, Naïve Bayesian, and SVM in Jupyter Notebook and the Pipeline used in problem 6 above to fit the last column of the training dataset
   b) Use the testing data set to compute and compare the accuracy of the three models above.
   c) Use the summary for the strengths and weaknesses of each model and the characteristic of the sonar dataset to explain the observed performance differences.