CSE 6363： Machine Learning, Spring 2020

Time: Tuesday 7-9:50pm Location: NH 202

Instructors: Dr. Chris Ding, chqding@uta.edu
Dr. Di Ming, di.ming@mavs.uta.edu
Office Hour: Tuesday, 10:30am-12:30pm, ERB 424. (or by appointment)

TA: Qicheng Wang, qicheng.wang@mavs.uta.edu
Office Hour: Tuesday & Thursday, 3:00pm-5:00pm, ERB 204.

Textbook:

Pattern Recognition and Machine Learning
Christopher Bishop

Course Schedule

Week 1.

Introductions
Three concrete examples:
1. Data Mining example: Market basket Data analysis
2. Pattern Recognition example: Handwritten letters recognition
3. Cancer prediction using DNA expressions recorded on microarrays

Fitting Curve to Data (textbook sec. 1.1)

Linear Regression

Homework 1-5

HW1: Textbook Exercise 1.1(p.6, p.58)

HW2: Show that when M=1, the results of HW1 is identical the results of linear regression.

HW3: Textbook Exercise 1.2.

HW4

A problem on a multiple-choice quiz is answered correctly with probability 0.9 if a student is prepared. An unprepared student guesses between 4 possible answers, so the probability of choosing the right answer is 1/4. Seventy-five percent of students prepare for the quiz. If Mr. X gives a correct answer to this problem, what is the chance that he did not prepare for the quiz?

HW5

At a plant, 20% of all the produced parts are subject to a special electronic inspection. It is known that any produced part which was inspected electronically has no defects with probability 0.95. For a part that was not inspected electronically this probability is only 0.7. A customer receives a part and finds defects in it. What is the probability that this part went through an electronic inspection?

HW1, HW2, HW3, HW4, HW5 are due on Feb 11th, 7:00pm.

Computer Project 1, due on Mar 3th, 7:00pm. ----------

You MUST write the codes YOURSELF.

Computer Project 1A: Write a computer program to generate the 10 data points as shown in Figure 1.2.
Hint 1A: for each data point (x_i, y_i), a random noise e_i is added on x_i.

Computer Project 1B: Write a computer program to solve the equations of Exercise 1.1, for the 10 data points you generated in part 1A. Plot the fitted curves and original data points as Figure 1.4, for M=0, 1, 3, 9.
Hint 1B: (1) transform original 1-dimensional feature to multi-dimensional features;
(2) use linear regression to solve equation (1.2).

Computer Project 1C: Write a computer program to solve the equations of Exercise 1.2, for the 10 data points you generated in part 1A. Here, M is fixed as 9. Show that as the lambda of Equation (1.4) increases, the overfitting of Figure 1.4 (the right-bottom figure) is reduced significantly, see Figure 1.7.
Hint 1C: (1) transform original 1-dimensional feature to multi-dimensional features;
(2) use linear regression with L_2-regularization to solve equation (1.4).

Computer Project 1D: Write a computer program to implement naïve Bayes classifier with Laplacian (add-1) smoothing, for the given "vertebrate.txt" dataset. Compute the multinomial distribution for each attribute of the data instances and prior probability. When a new data instance is presented, compute the class label using NAÏVE Bayes classification method.
Hint 1D: (1) training stage: compute the prior probability and likelihood.
(2) testing stage: compute the posterior probability using Bayes theorem.

Computer Project 2, due on TBD. ----------

Write a computer program that can read in data such as those in Table 4.1. (A) The program can represent each data instance --- these are "stored data" (B) When a new data instance is presented, the program compares the new data instance to every "stored data instances" to compute the distance, and find out the k nearest neighbors, and predict the class label for the new data instance. (C) Compute the multinomial distribution for each attribute of the data instances and prior probability. When a new data instance is presented, compute the class label using NAÏVE Bayes classification method.

You MUST write the codes YOURSELF.

In the exam, we will ask you modify the codes slightly to show more information (beyond the class labels) on KNN or Naïve Bayes; or to improve the algorithm.

Hint. You need to write a subroutine distance(x1,x2) to compute the distance between data instances x1,x2. Inside distance(), you should group numerical attributes together as num-set, and group categorical attribute together as Cat-set, etc. On num-set, you use Euclidean distance. On Cat-set, you use Hamming distance. In the Exam, we might require you to change how distance are computed.

Computer Project 3, due on TBD. ----------

Data Cluster using K-means algorithm provided by the system.

1. Run k-means on AT&T 100 images, set K=10. Obtain confusion matrix. Re-order the confusion matrix using bipartite graph matching and obtain accuracy.

2. Run k-means on AT&T 400 images, set K=40. Obtain confusion matrix. Re-order the confusion matrix and obtain accuracy.

3. Run k-means on Hand-written-letters data, set K=26, as above.

Computer Exam3 will depend on the codes you write for Project 3.

---------- Exams ----------

First computer quiz, on TBD**,** will use the program you developed in Project 1 to solve a problem similar to HW3.

Second computer quiz, on TBD, will use the program you developed in Project 2.

Third computer quiz will be on final exam day.

First written quiz will be on TBD. [Openbook: you can bring anything, Class lectors printed out. Textbook, etc. No computer access].

Second written quiz will be on final exam day.

No third written quiz/exam

Week 2-3:

Probability, Binomial distribution, Multinomial distribution, Gaussian distribution, Bayes classifier, Bayes Error rate, Bayes theorem, Naive Bayes Classification, Linear regression from the Gaussian Distribution point of view (Sec. 1.2.5)

Read: Textbook Sections 1.2.1, 1.2.2, 1.2.3, 1.2.4, 1.2.5. Sec.1.2.5 is one focus subject. Sections 2.2,2.3,

Week 4-5:

More probability

Mutual information

Decision Trees

Week 6

Feature Selection
- t-statistic, f-statisic
- mutual information
- mininum redundency, maximun relevance
- filters, wrappers, feature set selection

Week 7:

K-nearest Neighbor

Nearest Centroid Classification Method

Decision boundary

Week 8-9.

Read Textbook: Section 7.1, 7.2, 7.3. "A Tutorial on Support Vector Machines for Pattern Recognition" by Christopher Burges.

- Support Vector Machine
- Multi-class classification using binary classifiers
- Kernels (Gaussian, polynomial)

Homework. Due on TBD.

HW4: Solve SVM for a data set with 3 data instances in 2 dimensions: (1,1,+), (-1,1,-), (0,-1,-). Here the first 2 number are the 2-dimension coordinates. '+' in 3rd place is positive class. And '-' in 3rd place is negative class. Your task is to compute alpha's, w, b.

HW5. Solve SVM when data are non-separable, using k=2 when minimizing the violations of the mis-classification, i.e., on those slack variables.

Week 10

K-means clustering (Textbook: Section 9.1)

Homework 9. Due TBD

Explain why the K-means objective function decreases in each of the two steps in K-mean algorithm: (a) re-assign every data points to their nearest cluster centroids. (b) Given the grouping (or clustering), re-computer the cluster centroids.

Homework 10 (Computer) . Due TBD.

(A) Generate Three Gaussian distributions, each with 100 data points in 2 dimensions, with centers at (5,5), (-5, 5), and (-5,-5) and standard deviation \sigma = 2. Draw them in a Figure. Set K=3, do K-means clustering. Show the results in the same Figure. Repeat this 5 times. Submit the 5 figures, each represent the results of each K-means clustering.

(B) Everything are same as (A), but with \sigma=4. Submit the 5 figures.

NOTE: The third computer exam could use the codes you developed for Computer Homework 10.

Week 11

Dimension Reduction
- principle component analysis
- linear discriminant analysis