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 & 2.

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

Probability

Week 3 & 4.

Naïve Bayes Classifier

Decision Tree, Mutual Information, Random Forest

Week 5 & 6 & 7.

Classification：
KNN.
Centroid Method.

Support Vector Machine (SVM):
Margin, Primal-Dual Problems, KKT Condition.
Hard SVM, Soft SVM, Kernel SVM.

Week 8-9.

Spring break (no class).

Week 10.

March 24th: **first computer quiz** (polynomial curve fitting & naïve Bayes classifier)

Week 11.

March 31th: **first written quiz** (including all the lectures before spring break)

Week 12-15.

Majority Voting, Ensemble Learning, Logistic Regression

Feature Selection, Sparse Coding

K-means Clustering

Principal Component Analysis

Week 16.

May 5th: **second computer quiz** (computer quiz will depend on the codes you write for Project 2 & 3.)

Week 17.

May 12th: **second written quiz** (including all the lectures after spring break)

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.

Homework 6-7

HW6

Solve SVM for a data set with 3 data instances in 2 dimensions: (1,1,+), (0,-1,+), (-1,1,-). 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: (1) write down dual-problem using those 3 data instances. (2) compute alpha's. (3) compute w and b. (4) compute margin.

HW7

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

$$\min_{\mathbf{w}} \ \frac{1}{2}||\mathbf{w}||_2^2 + C\Big(\sum_{i=1}^{n}\xi_i\Big)^k$$

$$\text{s.t.} \ \ y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \cdots, n.$$

$$\xi_i \geq 0, \quad i = 1, \cdots, n.$$

Your task is to: derive the dual-problem of the above primal-problem when k=2.

HW6, HW7 are due on Mar 31th, 7:00pm.

Homework 8-9

Homework 9.

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).

(A) Generate Three Gaussian distributions, each with 100 data points in 2 dimensions, with centers at (3,3), (-3,3), and (0,-3) and standard deviation \sigma = 2. Draw them in a Figure. Set K=3, do K-means clustering. Show the clustering results in the same Figure and compute the converged K-mean loss. Repeat this 5 times. Submit the 5 figures and losses, each represent the result of each K-means clustering.

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

HW8, HW9 are due on May 5th, 7:00pm.

Computer Project 1, due on Mar 24th, 7:00pm. -------------------------------

You MUST write the codes YOURSELF.

For the implementation of "Polynomial Curve Fitting" and "Naïve Bayes Classifier", you are NOT allowed to use any Python Library.

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, t_i), a random noise e_i is added to obtain $t_i = \sin(2\pi x_i) + e_i$.
Details can be found in the textbook page #4.

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 May 5th, 7:00pm. -------------------------------

You MUST write the codes YOURSELF.

For the implementation of "K Nearest Neighbor" and "Centroid", you are NOT allowed to use any python machine learning library.

For the implementation of "SVM", you are allowed to use python machine learning library, such as Scikit-Learn.

Computer Project 2A: Write a subroutine **"distance(x1, x2)"** to compute the distance between two data instances x1, x2. Inside distance(), you should group numerical attributes together as num-set, group categorical attribute together as cat-set, etc. On num-set, you use **Euclidean distance**. On cat-set, you use **Hamming distance**.

For example, we have x1 = [11, "warm-blooded", 2] and x2 = [13, "cold-blooded", 3], distance(x1, x2) will be $(11-13)^2 + 4 + (2-3)^2 = 9$.

For categorical attributes, we use Hamming distance. The Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different, for example:
dist("warm-blooded", "cold-blooded") = 4;
dist("small", "large") = 5.
For two strings of unequal-length, we can pad the string of shorter length with empty characters, for example:
dist("small", "medium") = dist("small ", "medium") = 6;
dist("war", warm") = dist("war ", warm") = 1.

Computer Project 2B: Write a computer program to implement **KNN method**.
(i) The program can represent each data instance --- these are "stored data".
(ii) When a new data instance is presented, the program compares the new data instance to every "stored data instances" to compute the distance via subroutine "distance(x1, x2)", and find out the k nearest neighbors, and predict the class label for the new data instance.

Computer Project 2C: Write a subroutine **"convert(X)"** to convert both numerical features and categorical features in data instances.
(i) Convert categorical features using **One-hot Encoding** representation.
(ii) Convert numerical features using **Z-score** representation.

For example, we have data instances with one categorical feature named "body-temperature", then we convert "warm-blooded" and "cold-blooded" to [1,0] and [0,1].

For example, we have data instance with one numerical feature named "salary". First, we compute the mean μ and the standard deviation σ of all the sample values in the "salary" feature. Then, we replace each value x in “salary” with $\frac{x-\mu}{\sigma}$.

Computer Project 2D: Write a computer program to implement **Centroid method**.
(i) Preprocess the data instances via subroutine convert(X).
(ii) The program can represent all the data instances as “k centroids” (assuming the number of classes is k) --- these are “stored data”.
(iii) When a new data instance is presented, the program compares the new data instance to “k stored centroids” to compute the distance via subroutine “distance(x1, x2)”, and find out the nearest neighbor, and predict the class label for the new data instance.

Computer Project 2E: Write a computer program to implement **SVM method.**
(i) Preprocess the data instances via subroutine convert(X).
(ii) Given the training data instances, your program should be able to compute the w, b, alpha, margin.
(iii) When a new data instance is presented, the program uses pre-trained SVM model to predict the class label for the new data instance.

Computer Project 2F: Write a computer program to preprocess the given dataset “american_people_1000.txt”. Data file contains 15 attributes and last attribute “Income” is the class label.
(i) Following attributes: “fnlwgt”, “EducationNum”, “CapitalGain”, “CapitalLoss” will not be used. You should drop those columns.
(ii) Partition the dataset into training set (1st-900th instances) and testing set (901st-1000th instances).
(iii) Train your implemented KNN, Centroid, and SVM classifiers on those training data instances, and report their classification accuracy on testing data instances.
(iv) Check if SVM has different classification performances using linear/Gaussian kernel.

Computer Project 3, May 5th, 7:00pm. -------------------------------

You MUST write the codes YOURSELF.

For the implementation of project 3, you are allowed to use python machine learning library, such as Scikit-Learn.

In “AT&T” and “Hand-written-letters” datasets, each column represents one data instance. For each data instance, first element is class label and remaining elements are numerical features.

Computer Project 3A. Run k-means on AT&T face images dataset, set K=40. (i) Return the clustering result and compute the k-means loss. (ii) Obtain confusion matrix and compute accuracy. (iii) Re-order the confusion matrix using bipartite graph matching and compute accuracy.

Computer Project 3B. Run k-means on Hand-written-letters dataset, set K=26, as above.