CptS 315: Introduction to Data Mining
Homework 5

(Due date: April 29th via Dropbox)

Instructions

- Please use a word processing software (e.g., Microsoft word) to write your answers and submit a PDF via the dropbox link. The rationale is that it is sometimes hard to read and understand the hand-written answers.
- All homeworks should be done individually.

Analytical Part (60 points)

Q1. (20 points) Suppose you are given 7 data points as follows: A = (1, 1); B = (1.5, 2.0); C = (3.0, 4.0); D = (5.0, 7.0); E = (3.5, 5.0); F = (4.5, 5.0); and G = (3.5, 4.5). Manually perform 2 iterations of K-Means clustering algorithm (slide 22 on clustering) on this data. You need to show all the steps. Use Euclidean distance (L2 distance) as the distance/similarity metric. Assume number of clusters k=2 and the initial two cluster centers C_1 and C_2 are B and C respectively.

Q2. (20 points) Please read the following two papers and write a brief summary of the main points in at most FOUR pages.

Matthew Zook, Solon Barocas, danah boyd, Kate Crawford, Emily Keller, Seeta Pea Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara Knig, Jacob Metcalf, Arvind Narayanan, Alondra Nelson, Frank Pasquale: Ten simple rules for responsible big data research. PLoS Computational Biology 13(3) (2017)
`https://www.microsoft.com/en-us/research/wp-content/uploads/2017/10/journal.pcbi_.1005399.pdf`

Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, Jonathan Zittrain: Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. Proceedings of Machine Learning Research (PMLR), 81:62-76, 2018
`http://proceedings.mlr.press/v81/barabas18a/barabas18a.pdf`

Q3. (20 points) Please go through the excellent talk given by Kate Crawford at NIPS-2017 Conference on the topic of "Bias in Data Analysis" and write a brief summary of the main points in at most FOUR pages.

Kate Crawford: The Trouble with Bias. Invited Talk at the NIPS Conference, 2017. Video:
`https://www.youtube.com/watch?v=fMym_BKWQzk`

Empirical Analysis (40 points)

You will use the Weka: `http://www.cs.waikato.ac.nz/ml/weka/downloading.html` software. You can use the Graphical Interface to answer all the questions below – It is easier. Weka employs the ARFF (https://www.cs.waikato.ac.nz/ml/weka/arff.html) format for datasets. All the specific details provided below are for Weka.

Please use the "ionosphere" dataset provided for this question in ARFF format. Please use the last 25 percent examples for testing and the remaining examples for training.

You can also use Scikit-learn `http://scikit-learn.org/stable/` software if you are more comfortable with Python.

- Bagging (weka.classifiers.meta.Bagging). You will use decision tree as the base supervised learner. Try trees of different depth (1, 2, 3, 5, 10) and different sizes of bag or ensemble, i.e., number of trees (10, 20, 40, 60, 80, 100). Compute the training accuracy and testing accuracy for different combinations of tree depth and number of trees; and plot them. List your observations.

- SVM Classification learner. You will run the SVM classifier on the training data to answer the following questions.

  (a) Using a linear kernel, train the SVM on the training data for different values of C parameter(-c option): $10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}$. Compute the training accuracy, and testing accuracy for the SVM obtained with different values of the C parameter. Plot the training accuracy and testing accuracy as a function of C (C value on x-axis and Accuracy on y-axis) – one curve each for training, validation, and testing data. List your observations.

  (b) Repeat the experiment (a) with polynomial kernel of degree 2, 3, and 4. Compare the training and testing accuracies for different kernels (linear, polynomial kernel of degree 2, polynomial kernel of degree 3, and polynomial kernel of degree 4). List your observations.

Grading Rubric

Each question in the students work will be assigned a letter grade of either A,B,C,D, or F by the Instructor and TAs. This five-point (discrete) scale is described as follows:

- **A) Exemplary (=100%).**
  Solution presented solves the problem stated correctly and meets all requirements of the problem.
  Solution is clearly presented.
  Assumptions made are reasonable and are explicitly stated in the solution.
  Solution represents an elegant and effective way to solve the problem and is not overly complicated than is necessary.

- **B) Capable (=75%).**
  Solution is mostly correct, satisfying most of the above criteria under the exemplary category, but contains some minor pitfalls, errors/flaws or limitations.

- **C) Needs Improvement (=50%).**
  Solution demonstrates a viable approach toward solving the problem but contains some major pitfalls, errors/flaws or limitations.

- **D) Unsatisfactory (=25%)**
  Critical elements of the solution are missing or significantly flawed.
  Solution does not demonstrate sufficient understanding of the problem and/or any reasonable directions to solve the problem.

- **F) Not attempted (=0%)**
  No solution provided.

The points on a given homework question will be equal to the percentage assigned (given by the letter grades shown above) multiplied by the maximum number of possible points worth for that question. For example, if a question is worth 6 points and the answer is awarded a B grade, then that implies 4.5 points out of 6.