CptS 315: Introduction to Data Mining
Homework 4
(Due date: April 22nd)

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.

Q1. (10 points) Suppose $x = (x_1, x_2, \cdots, x_d)$ and $z = (z_1, z_2, \cdots, z_d)$ be any two points in a high-dimensional space (i.e., d is very large). Suppose you are given the following property, where the right-hand side quantity represents the standard Euclidean distance.

$$\left(\frac{1}{\sqrt{d}}\sum_{i=1}^{d} x_i - \frac{1}{\sqrt{d}}\sum_{i=1}^{d} z_i\right)^2 \leq \sum_{i=1}^{d}(x_i - z_i)^2 \quad (1)$$

We know that the computation of nearest neighbors is very expensive in the high-dimensional space. Discuss how we can make use of the above property to make the nearest neighbors computation efficient?

Q2. (10 points) We know that we can convert any decision tree into a set of if-then rules, where there is one rule per leaf node. Suppose you are given a set of rules $R = \{r_1, r_2, , r_k\}$, where r_i corresponds to the i^{th} rule. Is it possible to convert the rule set R into an equivalent decision tree? Explain your construction or give a counterexample.

Q3. (10 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.

Q4. (25 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

Q5. (25 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

Q6. (20 points) Please read the following two papers and write a brief summary of the main points in at most THREE pages.

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Franois Crespo, Dan Dennison: Hidden Technical Debt in Machine Learning Systems. NIPS 2015: 2503-2511
https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

Eric Breck, Shanqing Cai, Eric Nielsen, Michael Salib, D. Sculley: The ML test score: A rubric for ML production readiness and technical debt reduction. BigData 2017: 1123-1132
https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/46555.pdf

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.