CptS 315: Introduction to Data Mining
Homework 4

(Due date: April 14th 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. (10 points) Suppose you are given the following multi-class classification training data, where each input example has three features and output label takes a value from *good*, *bad*, and *ugly*.

- x_1=(0, 1, 0) and y_1=*good*
- x_2=(1, 0, 1) and y_2=*bad*
- x_3=(1, 1, 1) and y_3=*ugly*
- x_4=(1, 0, 0) and y_4=*bad*
- x_5=(0, 0, 1) and y_5=*good*

Suppose we want to learn a linear classifier using multi-class perceptron algorithm and start from the following weights: w_{good}=(0,0,0); w_{bad}=(0,0,0); and w_{ugly}=(0,0,0). Please do hand calculations to show how weights change after processing examples in the same order (i.e., one single pass over the five training examples). See slide 88 of the Perceptron notes.

Q2. (10 points) Suppose you are given the following binary classification training data, where each input example has three features and output label takes a value *good* or *bad*.

- x_1=(0, 1, 0) and y_1=*good*
- x_2=(1, 0, 1) and y_2=*bad*
- x_3=(1, 1, 1) and y_3=*good*
- x_4=(1, 0, 0) and y_4=*bad*
- x_5=(0, 0, 1) and y_5=*good*

Suppose we want to learn a classifier using kernelized perceptron algorithm. Start from the following dual weights: α_1=0; α_2=0; α_3=0; α_4=0; and α_5=0. Please do hand calculations to show how dual weights change after processing examples in the same order (i.e., one single pass over the five training examples). Do this separately for the following kernels: (a) Linear kernel: $K(x, x')$=$x \cdot x'$; and (b) Polynomial kernel with degree 3: $K(x, x')$=$(x \cdot x' + 1)^3$, where $x \cdot x'$ stands for dot product between two inputs x and x'. See Algorithm 30 in `http://ciml.info/dl/v0_99/ciml-v0_99-ch11.pdf`. You can ignore the bias term b.

Q3. (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.

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

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?

Q4. (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.

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`

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 eas-

ier. 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 voting dataset provided for this question in ARFF format. Please use the last 100 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 (weka.classifiers.trees.j48) 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 (weka.classifiers.functions.supportVector). You will run the SVM classifier on the training data to answer the following questions.

  (a) Using a linear kernel (-t 0 option), 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 (-t 1 -d option) 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.