CSE Ph.D. Qualifying Exam, Spring 2021

Data Analysis

Please answer three of the following four questions. All questions are graded on a scale of 10. If you answer all four, all answers will be graded and the three lowest scores will be used in computing your total. Show all your work and write in a readable way.

1. **Streaming K-means**

   In this problem, we consider how to extend the k-means algorithm to process streaming data. The standard k-means algorithm loads all data points into memory. In practice, data may come in a stream, such that they are sequentially processed and dropped. The advantage of streaming algorithms is in processing data that cannot fit into the memory.

   Now consider how to extend the k-means algorithm to the streaming version. Suppose that there are k clusters. The cluster centers are randomly initialized. Once the processor receives a data point $x \in R_d$, it does the following: 1) Find the cluster whose center is the closest to x (in Euclidean distance), then add x to the cluster; 2) Adjust the cluster center so that it equals the mean of all cluster members. The algorithm outputs the k cluster centers after processing all data points in the stream. According to the above algorithm specification, complete the streaming algorithm for k-means:

   (a) List the variables that are stored in the memory and their initial values. Which variables should be the output of the algorithm?

   (b) When the processor receives a data point x, state the updates that are made on the variables.

   (c) In each iteration, suppose the processor receives a data point x along with its weight $w > 0$. We want the cluster center to be the weighted average of all cluster members. How do you modify the updates in question (b) to process weighted data?

2. **Cars and Clusters**

   (a) Imagine n cars, each of which travels at a different maximum speed. Initially, the cars are queued up in uniform random order at the starting point of a semi-infinite, one lane highway. Each car drives at the minimum of its maximum speed and the speed at which the car in front of it is driving. The cars will form 'clumps'/clusters. What is the expected number of clumps? Prove your answer.

   (b) Consider the following random graph model with clustering. For n nodes, we have $\binom{n}{3}$ distinct 'triplets'. For *each* triplet, with independent probability p we connect the nodes belonging to this triplet in the graph using three edges to form a triangle, where $p = \frac{c}{\binom{n-1}{2}}$, where c is a constant. Assume n is very large. Show that the expected degree of a node in this graph model is $2c$.

3. **Support Vector Machine**

Given 2-dimensional input data points $S_1 = \{(1,4),(1,5),(2,4),(2,5),(3,3)\}, S_2 = \{(3,2),(3,1),(4,1),(5,1),(6,1),(6,2)\}$, where S_1 has the data points from the positive class and S_2 has data points from the negative class:

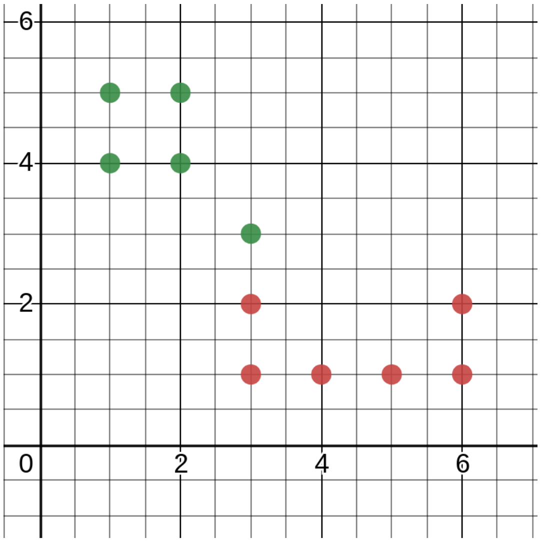

Figure 1: Plot of points $S_1 = \{(1,4),(1,5),(2,4),(2,5),(3,3)\}, S_2 = \{(3,2),(3,1),(4,1),(5,1),(6,1),(6,2)\}$

(a) Suppose you are using a linear SVM with no provision for noise (i.e. a Linear SVM that is trying to maximize its margin while ensuring all data points are on their correct sides of the margin). Draw three lines on the above diagram, showing classification boundary and the two sides of the margin. Circle the support vector(s).

(b) Using the familiar Linear SVM classifier notation of the classifier $\text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b)$, calculate the values of $\boldsymbol{w}$ and b learned for part (a).

(c) Assume you are using a noise-tolerant Linear SVM which tries to optimize

$$\min_{w,b,\epsilon} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\Sigma_i\epsilon^i$$
$$\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x}^i + b) \geq 1 - \epsilon^i, \epsilon^i \geq 0, \forall i$$

Question: is it possible to invent a dataset and a positive value of C in which the dataset is linearly seperable but the linear SVM classifier would never the less misclassify at least one training point? If it is possible to invent such an example, please sketch the example and suggest a value for C. If it is not possible, explain why not.

4. **Decision Tree Classifier**

   Alice is a cyber analyst designing a binary classifier to detect network intrusions in at a large technology company. She is considering using a decision tree (classification tree) for this task.

   (a) In Alice's context, what would the *positive* class typically refer to?

   (b) Alice is considering three common approaches to measure her tree's classification error. Briefly describe each approach, and state at least one drawback for each approach.

      i. Misclassification rate

      ii. Average loss

      iii. Normalized negative log-likelihood (or cross-entropy)

   (c) Alice is consider using a ROC (receiver operating characteristic) curve to visualize her classifier's performance. Her colleague Bob suggests she use AUC (area under an ROC curve) to summarize each ROC into a single AUC value instead, so the AUC values may be more easily compared.

      i. Briefly explain why Bob's suggestion of using AUC may be problematic.

      ii. Alice finds that one of her trees has an AUC score of 0. Her colleague Bob notices this and is very happy with the score — why?