Quiz 1 Solutions

CS 349-02

April 10, 2017

Name: ____

Write in the space provided. You need not simplify arithmetic such as scalar addition or multiplication. Be **brief** with answers involving explanations. If you get stuck, move on.

1. Consider the following training data-points.

   $\boldsymbol{a} = [1, 0]$, labeled -1

   $\boldsymbol{b} = [0, 1]$, labeled $+1$

   $\boldsymbol{c} = [1, 2]$, labeled -1

   We've currently learned a hyperplane centered at the origin, given by $\boldsymbol{w} = [-1, 1]$

   (a) Compute the dot product of $\boldsymbol{w}$ with each of the three points.

   $\boldsymbol{w}.\boldsymbol{a} = -1 \cdot 1 + 1 \cdot 0 = -1$

   $\boldsymbol{w}.\boldsymbol{b} = -1 \cdot 0 + 1 \cdot 1 = +1$

   $\boldsymbol{w}.\boldsymbol{c} = -1 \cdot 1 + 1 \cdot 2 = +1$

   (b) What is the distance between point $\boldsymbol{b}$ and the hyperplane?

   $$\frac{\boldsymbol{w}.\boldsymbol{b}}{\|w\|} = \frac{1}{\sqrt{2}}$$

   (c) Does the current hyperplane mis-classify any of the points? If so, update $\boldsymbol{w}$ in response to this error. (Make only this update; don't check the other points again.)

   It misclassifies $\boldsymbol{c}$. Update:

   $$\boldsymbol{w} - \boldsymbol{c} \text{ (since } \boldsymbol{c} \text{ is labeled -1)} = [-1, 1] - [1, 2] = [-2, -1]$$

   (d) I re-represent this data using 6 features instead of 2. If an epoch through the original 2-dim data takes 1 second, how long will it take on the new 6-dim data?

   Since the algorithm is linear in the number of features, it will take $6/2 = 3$ seconds.

2. The perceptron algorithm is guaranteed to find a separating hyperplane

   provided that the data is linearly separable.

3. When is it helpful to have a bias term b in addition to $\boldsymbol{w}$?

   When the data is linearly separable, but not by a hyperplane centered at the origin.

4. The last few perceptron updates have an extreme influence on the final hyperplane. Name one modification to the perceptron algorithm that mitigates this.

   Averaged perceptron (or voted perceptron).

5. The perceptron learning algorithm may converge very slowly under a given ordering of training points. What is the workaround?

   Shuffle the order of the points on each iteration.

6. Here is a 3x3 black and white (0=black, 1=white) image; each cell is a pixel.

0	1	1
0	1	0
0	0	0

   (a) Represent this image as a high dimensional vector like we did in PS1.

   [0, 1, 1, 0, 1, 0, 0, 0, 0]

   (b) Write the above vector in sparse representation like the movie review data provided in PS2.

   {1: 1, 2: 1, 4: 1}

7. You have several decisions to make when building a supervised classifier for a task. One is which algorithm to use. Name another.

   Which features to use to represent the data.

8. Alice is building a classifier to detect which language a word belongs to. She represents each word in terms of these features:

   - Length
   - Number of vowels
   - Whether or not the word ends in the letter `n` (0 for no, 1 for yes)

   Here's the training data she's given.

`regeln`	German
`pido`	Spanish

   (a) The dimensionality of the data under this feature representation is 3

   (b) Write out the feature vectors of the training words.

   `regeln`: $[6, 2, 1]$

   `pido`: $[4, 2, 0]$

   (c) Using the kNN algorithm with $k = 1$ and Manhattan (L1) distance, what should she predict is the language for the word '`neben`'? Show your calculations.

   `neben`: $[5, 2, 1]$

   L1 distance between `neben` and `regeln` $= |[5, 2, 1] - [6, 2, 1]| = 1$
   L1 distance between `neben` and `pido` $= |[5, 2, 1] - [4, 2, 0]| = 2$

   The label of the nearest training example, `regeln`, is German. Therefore, Alice should predict that `neben` is German.

   (d) Alice does not know if the Manhattan metric is the best choice for her classifier. What should she do to decide on the best metric?

   Tune the metric hyperparameter on development data: Run the kNN algorithm with different candidate metrics, training on the training data and evaluating the error on the development set. Pick the metric with the highest development accuracy.

9. What is one drawback of the k nearest neighbors algorithm?

   Slow testing time, does not take into account the overall distribution of the data.

10. Name one drawback of the perceptron algorithm compared to k nearest neighbors.

    Slow training time, only works for binary classification, only works when the data is linearly separable.

11. Write the two text messages below as vectors using the word-count representation as in PS2. The vocabulary is the set of all words present in either of the two messages.

    `good luck`: [1, 1, 0, 0]

    `i am good`: [1, 0, 1, 1]

    (With the features being, in order, `good, luck, i, am`.)

12. A company is selling a kNN classifier that gets 100% accuracy on training data. What is the danger of using it on your own data?

    It has probably overfit to the training data.

13. You are classifying several testing points, as in the problem sets. Assume your model is already trained. What is the runtime complexity of this testing step given m training points, p testing points, and d dimensions, using

    (a) kNN?

    $O(pmd) + O(pm \log m)$, or $O(pmd) + O(pmk)$

    (b) Perceptron?

    $O(pd)$

This is the end of the quiz.