CSE 151A - Homework 06

Due: Wednesday, May 13, 2020

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Unless otherwise noted by the problem's instructions, show your work or provide some justification for your answer. Homeworks are due via Gradescope on Wednesday at 11:59 p.m.

Essential Problem 1.

A vector $\vec{a}$ is a linear combination of vectors $\vec{v}_1, \ldots, \vec{v}_k$ if there exist scalars $c_1, \ldots, c_k$ such that $\vec{a} = \sum_{i=1}^{k} c_i \vec{v}_i$. Suppose we run the perceptron train algorithm on training data $(\vec{x}^{(1)}, y_1), \ldots, (\vec{x}^{(n)}, y_n)$ using an initial weight vector of $\vec{0}$, and we get an output $\vec{w}$ after a single pass on the data. Is $\vec{w}$ is a linear combination of $\text{Aug}(\vec{x}^{(1)}), \ldots, \text{Aug}(\vec{x}^{(n)})$? Justify your answer.

Essential Problem 2.

The perceptron learning algorithm is run on a data set. It converges after performing $n = a + b$ updates. Of these n updates, a are on data points with label $+1$ and b are on points whose label is -1. What is the final value of the "bias weight", w_0? Assume that the initial value of $\vec{w}$ is $\vec{0}$, and the learning rate α is set to 1.

Essential Problem 3.

An SVM classifier is learned for a data set in $\mathbb{R}^2$. The weight vector of this classifier is $\vec{w} = (-12, 3, 4)^T$.

a) Draw the decision boundary and mark where it intersects the axes.

b) Draw the margin boundaries (the lines where the dot product is 1 and -1). Also mark where these intersect the axes.

c) What is the (shortest) distance between the margin lines? Show your work.

d) How would the point $(2, 2)$ be classified by the SVM? Show your work.

e) It turns out that the there were only two support vectors, and they both have the form $(1, ?)$. What are they? Describe how you know.

Essential Problem 4.

Consider the following sentences:

- That is a big bug.
- No, seriously, that is a really big bug.
- Everybody run!

Encode each of the above sentences as bag-of-words feature vectors, using the set of all words in the three sentences as your dictionary. Sort the words alphabetically when creating your dictionary, and disregard capitalization and punctuation. You do not need to show your work (but doing so can help you get credit if you make a simple mistake).

Plus Problem 1. (16 plus points)

This is a legitimate machine learning mini-project that ties together several ideas, so it is worth 16 plus points.

The file `http://cse151a.com/data/yelp/train.csv` contains 10,000 Yelp reviews along with the score the user left (from 1 to 5, with 5 being the best). In this plus problem, you'll train an SVM to do sentiment analysis on these reviews and predict the sentiment of an unlabeled piece of text.

a) Split the data 75%/25% into training and validation sets, encode the training data using a bag of words feature representation, and train a (linear, soft-margin) support vector machine. When training, consider any review with a score or 4 or higher to be a positive review, and anything with a smaller score to be a negative review. Find the value of C that minimizes the error of your classifier on the validation set and make a plot of the validation error as a function of C.

For this part, turn in four things:

1. the value of C that was best,
2. the training and validation error that corresponded to this choice of C,
3. your plot, and
4. your code.

You can use whatever machine learning libraries you like in whatever language you'd like. Note that most languages have libraries which will do the bag-of-words encoding for you. For instance, `sklearn` has this feature (but I'll let you Google for it!).

b) Is the data in `train.csv` linearly separable? How do you know?

c) Give an example of:

- A sentence that you think is positive that your predictor got right.
- A sentence that you think is negative that your predictor got right.
- A sentence that you think is positive that your predictor got wrong.