IN4320 : Machine Learning Assignment 2

March 25, 2019

In this assignment, you are going to use—what we will refer to as—the linear regression classifier [LRC]. More specifically, you will be considering two semi-supervised variations on this classifier.

The regular supervised version of the LRC works as follows. For the training phase, we are given N training data points (x_i, y_i), in which the x_i are d-dimensional feature vectors and the corresponding $y_i \in \{-1, +1\}$ are the labels. Based on these N observations, we determine $w \in \mathbb{R}^d$ and $w_\circ \in \mathbb{R}$ that minimize the objective function

$$\frac{1}{N}\sum_{i=1}^{N} \left(\langle w, x_i \rangle + w_\circ - y_i\right)^2 .$$

At test time, the classifier assigns any vector $x \in \mathbb{R}^d$ to the class $\operatorname{sign}(\langle w, x_i \rangle + w_\circ)$.

When it comes to the implementation of your two semi-supervised approaches for the LRC, you are allowed to take any inspiration from other works, papers, web pages, etc., you are allowed to implement existing methods and use other people's toolboxes. In any case, do provide proper references to where you got your inspiration from!

Here are the more specific exercises for you to do[1].

1 Describe your two different ways of semi-supervised learning for the LRC *on an algorithmic level.* Keep the descriptions clear and concise. **/ 2 points**

Through the course page you can find the data for a two-class classification task in 11 dimensions that contains 10000 samples. It is named `twoGaussians.txt`. The feature vectors are stored in rows. The last element in every row contains the class label.

2 For the above data set, make learning curves against the number of *unlabeled* samples when using 8 labeled samples *per class* in the training set. Check, at least, adding 0, 8, 16, 32, 64, 128, 256, and 512 unlabeled samples and see how the expected true error rates change. Compare the two curve to the supervised error rates. Make sure you repeat your experiments sufficiently often enough to get some nice and, possibly, smooth curves. Do you get significant changes in error rates? **/ 2 points**

[1] **Note :** IN4320 assignments should be made individually. You must provide your own answers.

3 With the same data set as in **2**, make the same type of plots, but now for the expected squared loss[2] versus the number of unlabeled data. Are the changes significant? **/ 2 points**

4 Construct two artificial classification tasks with two classes [i.e., a problem from which one can draw random training and test sets]. On the one, your first semi-supervised LRC should work well *in terms of the error rate* and improve over the regular supervised LRC, but the second semi-supervised method should give deteriorated performance on this same set: its performance should be worse than the supervised classifier. On the other data set, it should be the other way around: the second semi-supervised LRC should work better than the supervised learner and the first learner should fail to do so.

Consider the setting in which you take few labeled samples and a large number of unlabeled samples [no need for learning curves]. Explain why the respective improvements and failures are expected. **/ 4 points**

My assessment : you should be able to keep your report within 800 words…

[2] Of course, like for the error rate, this should be done through the use of separate test data.