Decision Making with Business analytics
HW 1. Due date: 01 Oct 2019

You are required to do and submit your work in groups of (at most) two. Your work should be handed in (hard copies only) no later than the due date. Late work will be accepted with a 20% penalty per day (or fraction of day) beyond the deadline. **This hw is marked over 15 points.**

This hw is a continuation of the Labs 2 and 3. See the description of those labs for more information. In particular notice that the MNIST data set is divided into training and test sets. Your are not allowed to use the test set as part of your training data.
Some of the given tasks might take long time to run (but not to code). Be clever on how do you manage this issue, as the hw has a tight deadline. Also is not ideal that you will sit in from of a computer waiting for an answer.

Support Vector Machines

In Lab3, for each pair of digits (d_1, d_2) you have created a support vector machine SVM_{d_1,d_2} to classify each digit as d_1 or d_2. This was done by picking the best combination of parameters (the value of C and the kernel used). Answer the following questions based on SVMs obtained using only the radial kernel $K(u,v) = \exp\left(-\frac{\|u-v\|^2}{2\sigma^2}\right)$.

1. **[1 pts]** Consider the digit 5. What is the most similar digit to 5? What is the least similar one?

Majority vote

As explained in Lab3, the majority vote system will look at the predictions of the 45 classifiers to predict one single digit. Majority vote is very simple. Each classifier predicts a digit, majority vote then predicts the one predicted the most.

2. **[1 pts]** What is the accuracy of the majority vote system in this case?

3. **[1 pts]** Looking at each digit separately, which ones have the best/worst predictions? What are possible reasons?

We can try to improve the accuracy by using another voting system. We can look at the problem at hand as a classification problem with 10 classes and 45 features. We can tackle this problem via ANN. In this case we will use a package (e.g neuralnet in R) to train the ANN.

4. **[2 pts]** Train a neural network with 45 input nodes, one hidden layer with $H = 5, 10, 15, 20$ nodes in this layer, and 10 output nodes to obtain a voting system. For each H what is the accuracy of your prediction when using this system?. Pick the H performing best.

To solve 4, from each data point $(x_i, y_i) \in \mathbb{R}^{784\times} \times \{0, \ldots, 9\}$ you should prepare pairs $(u_i, v_i) \in \{0,1\}^{45} \times \{0,1\}^{10}$ where u_i is the result of the 45 comparisons and v_i is the binary vector that has a 1 in position y_i and a 0 everywhere else (i.e. $y_i = 0v_{i,0} + 1v_{i,1} + 2v_{i,2} + \cdots + 9v_{i,9}$). Then the data set $(u_i, v_i)_{i=1,\ldots N}$ obtained from the selected training set will be the training set for your ANN, where u_i are the input of the ANN and the v_i are the output.

Deep Learning - Auto encoders

In this part we will use auto-encoders to reduce our data 784 to 45 features. Then we can proceed as in 4 to create a "deep learning" system. I.e. we will use the output of our auto-encoder as input for a ANN. We had pick 45 as number of outputs for the auto-encoder to be able to compare to the majority vote.

5. **[3 pts]** Train a ANN with the following topology.

   [Input [784 nodes]] → [Hiden [200 nodes]] → [Hiden [45 nodes]] → [Hiden [H nodes]] → [Output [10 nodes]]

   To create this network, the following "parts" are trained separately and assembled together:

   (5.1) An auto-encoder that reduces from 784 variables to 200 'basic' features. To do this construct a neural network with 784 input nodes, one hidden layer with 200 nodes, and 784 output nodes (remember that for training you use output = input).

   (5.2) An auto-encoder that reduces from 200 to 45 'complex' features.

   (5.3) A ANN constructed in a similar way to 4. Use the value you have picked for H in 4.

   What is the accuracy of your prediction when using this topology? Relate the accuracy of the whole system to the accuracy of the parts.

6. **[1 pts]** Look at a few of the 200 nodes in the first hidden layer, by plotting the corresponding weights in a 28 by 28 format. What features are they encoding?

7. **[1 pts- Bonus point]** Look at a few of the 45 nodes in the second hidden layer. What features are they encoding?

Dimension Reduction

Our images are 28×28, even though this is small for today's standards, this generates samples with 784 features (pixels). In this part we will reduce the number of features used and evaluate the efficiency gain vs accuracy lost of the SVMs and ANNs.

8. **[4 pts]** Pick one of the methods used before and evaluate the effect of dimension reduction. Consider (separately) at least two of the following methods for dimension reduction:

   (8.1) PCA

   (8.2) Random projection

   (8.3) Projecting each 28×28 image to a 7×7 image by averaging the grey tone in each 4×4 sub-square.

Conclusions

9. **[1 pts]** Explain what is the best performing predictor and why do you think this is the case.

10. **[1 pts]** Pick another of the predictors constructed in this homework (different to the 'best performing' one). How could the performance of this predictor be improved? How difficult will be to do this?

Deliverables

You should deliver a written report. It should be no longer than 5 pages.

- There is not need for you to submit your code. But we may ask for it if we consider this necessary. We should be able to run you code in a simple way, and to obtain the same results.

  - In particular, you should seed your random number generator, to ensure that you are using the same random number in each run.

- You need to **clearly justify all your answers**. In particular (for the corresponding task) it should be clear:
  - what data set are you using (sample size, how did you obtain it (e.g. 10.000 random observations from the data labeled "train data"),
  - what methods are you using (in particular which function in R or matlab),
  - how did you chose the parameters of the method (e.g. the number of hidden units),
  - what formulas you used (e.g. for accuracy), and to what data-sets are you applying the formula.
  - any assumption you are making.
- If you are worried that tables and plots are consuming too much space, you can use an appendix for them. The appendix does not count towards the 5 page limit, but it should be limited to no more than 5 pages itself.
- Illegible answers will be ignored (i.e. you do not get any point if we can not understand what you wrote/typed).