Compulsory Assignment #2

Data Mining, Machine Learning and Deep Learning [KAN-CDSCO1004U]

Somnath Mazumdar
sma.digi@cbs.dk
Copenhagen Business School, Denmark

Deadline: {Refer digital exam for exact date and time}

Instructions

1. This compulsory assignment contains four questions. Some questions may have more than one question. Please answer all the questions.
2. You must upload your solutions before the deadline to the digital exam http://exam.cbs.dk/.
3. If your answers involve any math, please feel free to use whatever format you like [Latex, Word, plain paper and pen], but make sure it is readable.
4. It is always a good practice to use comments extensively in your code so that it will be easy for other people to understand it.
5. The code for answering the questions **should be submitted as one single *jupyter* notebook.**

Assignments

Question 1 : **Gradient Descent and Perceptron**

1. What is a learner said to do when it outputs a classifier that is 100% accurate on the training data but only 50% accurate on test data, when in fact it could have output one that is 75% accurate on both?
2. In stochastic gradient descent, each pass over the dataset requires the same number of arithmetic operations, whether we use mini-batches of size one or size 1000. Why can it nevertheless be more computationally efficient to use mini-batches of size 1000?
3. Below Figure 1 shows the level curves in the weight space of a cost function C which we are trying to minimize. The current weight vector is marked by an x. Sketch the gradient descent update.

   Hint: We haven't given you enough information to determine the magnitude, so we want you to correct the direction.

4. Suppose we want to train a perceptron (refer to Figure 2) with weights w_1 and w_2 and a fixed bias $b = -1$ Sketch the constraints in weight space corresponding to the following training cases.

   Hint: The decision boundaries have already been drawn for you, so you only need to draw arrows to indicate the half-spaces.) Shade the feasible region or indicate that none exists.

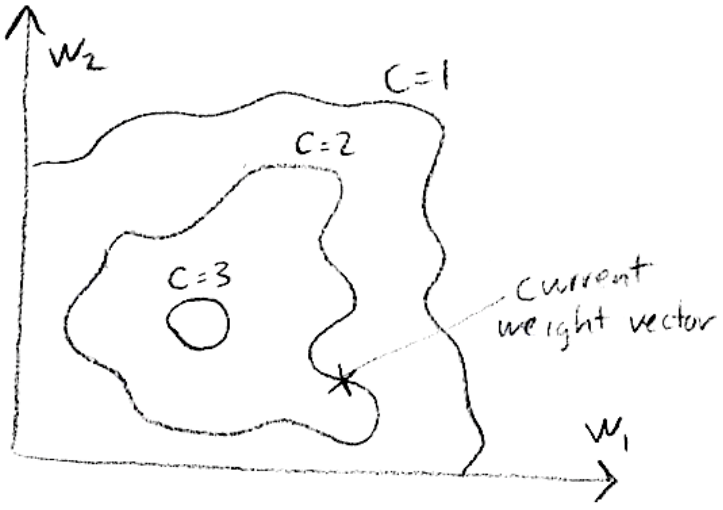

Figure 1: Use this diagram to sketch the gradient descent update

$\mathbf{x} = (1, -1),\ t = 1$

$\mathbf{x} = (-1, -1),\ t = 0$

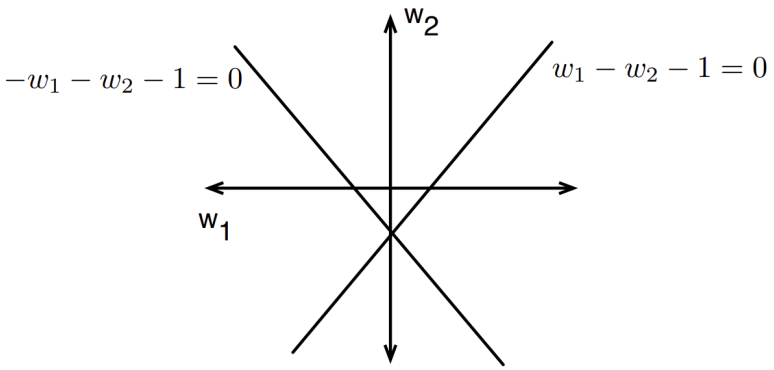

Figure 2: Use this perceptron to shade the feasible region

Question 2 : **Neural Networks**

The following is a network (refer to Figure 3) of linear neurons, that is, neurons whose output is identical to their net input. The numbers in the circles indicate the outcome of a neuron, and the numbers at connections indicate the value of the corresponding weight.

1. Compute the output of the hidden layer and the output-layer neurons for the given input (0.5, 1) and enter those values into the corresponding circles.
2. What is the network output for the input (1, 2) (i.e., the left input neuron having the value one and the right one having the value 2)? Do you have to do all the network computations once again to answer this question? Explain why you do or do not have to do this.

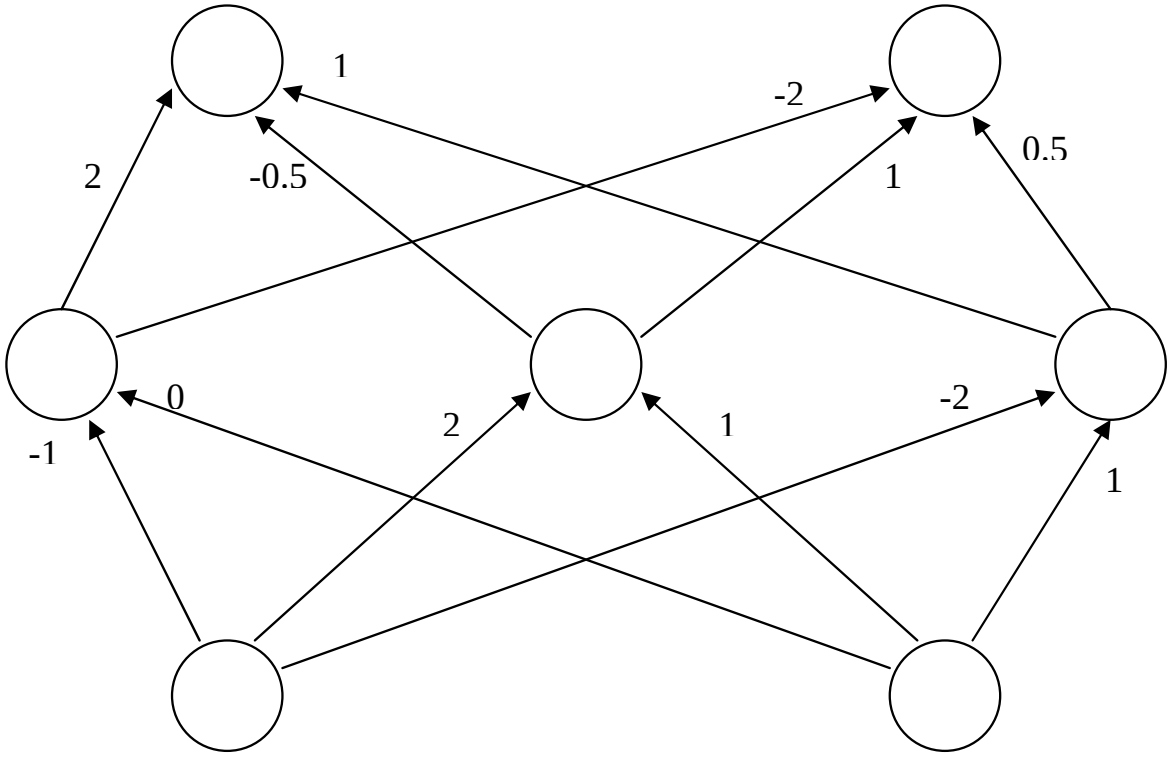

Figure 3: Network of linear neurons

Question 3 : **SVM and Random Forest**

Please download the attached dataset (data.csv), a labelled dataset from the marketing domain of a fictitious company ABC Limited. The dataset contains the number of impressions on ad platforms across 81 digital marketing channels. To clarify impressions in this context, assume that impressions mean the number of times an advertisement is displayed to users. The target label in this dataset is the last column (*Click*), which contains a boolean value whether the user has clicked on an advertisement or not.

Use this data to build a predictive model for deciding whether a user is likely to click on an advertisement based on the number of impressions on the marketing channels. Use ***Support Vector Machine*** (SVM) and ***Random Forests*** for prediction, which can be further used to make better market predictions in this domain. You might be interested in exploring the data for discrepancies (such as non-numeric values and some corrupt data, replace them with NA values). You also can check if there are any outliers (e.g. the number of clicks that are < 0 or ≥ 1000).

1. Develop supervised machine-learning classifiers as mentioned above, using an 80/20 split for training to test sets. Use the confusion matrix, precision, recall, f1, support and accuracy to compare the performance of the algorithms. Apart from that, also carry out Cross-Validation.
2. Next, create a function to find the five most influential features and develop a function to plot the most significant predictors by considering the size of their coefficients.
3. Based on the findings, make a recommendation to the company on which marketing channels you think the company should use if they would like to use only a few (one or two channels).
4. Furthermore, assuming a linear cost of impression (i.e. pay per impression) is the same across all channels (e.g. 0.10 DKK per impression), then considering the average costs spent per each channel, which channel(s) would you recommend if the company would like to minimize their marketing expenses?

Question 4 : **Written Assignment**

Write an approximately 2-pages extended abstract to discuss your reflections on:

The role of machine learning in the development of robotics design.

To write an essay on the above topic, you can do a tiny literature review to find more information. Alternatively, you can also write a report on the basic principles of machine learning and explain how it contributed to the overall state of the art. Therefore you have several choices to write/shape your extended abstract and feel free to choose whatever direction you want to explore. Finally, we would like to see more of your reflections and critical comments rather than just reproducing/reporting from what you have found in the literature review. Your report should conform to academic standards and use APA referencing/citation style for your essay.