Answers ML exam (Exercise 1-5)

Jonas Nordqvist

May 24, 2023

Exercise 1

Part the data into two disjoint datasets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, where the former consists of 80% of the total datapoints. Decide on a number of relevant hyperparameters for each model, and a corresponding relevant interval for these. Perform a hyperparameter search using cross-validation accuracy as the performance metric. Compare all models given by the different settings. Finally select the model which minimize the validation error. Then – only once a single model is left – you can use the test dataset to estimate the generalization error which you can present as the performance of your selected model.

The get full credit you should make a complete description of dividing your datasets, and discuss what the validation set/cross-valdiation may be used for (model selection and/or hyperparameter tuning). Finally, the test set may only be applied to the 'winning' model, no other in order for the estimate to be correct.

Exercise 2

To answer a) and b) right away we have for instance

$$W^{(2)} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad b^{(2)} = \begin{pmatrix} -2.5 \\ -1.5 \end{pmatrix}$$

and

$$W^{(3)} = \begin{pmatrix} -1 & 1 \end{pmatrix}, \quad b^{(3)} = -0.5.$$

Regarding c) the answer is *no* as the 'true' outputs of the NN should be the half plane in the unit cube of all points with at least two active signals except for $(1, 1, 1)$. This can be compared to the XOR function.

The last subexercise yields 1 of the 6 points. The other 5 are given if a) and b) are correct, and points are removed for mistakes.

Exercise 3

At the parent node the Gini index is given by

$$G_{\text{parent}} = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{3}{10}\right)^2 - \left(\frac{1}{5}\right)^2 = 1 - 0.25 - 0.03 - 0.04 = 0.62.$$

We have

$$\begin{aligned}\Delta_{S_1} &= 0.62 - \left(\frac{4}{5}\left(1-\left(\frac{5}{8}\right)^2-\left(\frac{3}{8}\right)^2\right)+\frac{1}{5}(1-1)\right)\\ &= 0.62 - \frac{4}{5}\left(1-\frac{34}{64}\right)\\ &= 0.62 - \frac{4}{5}\cdot\frac{30}{64}\\ &= 0.62 - \frac{6}{16} = 0.245.\end{aligned}$$

Also, considering S_2 we have

$$\begin{aligned}\Delta_{S_2} &= 0.62 - \left(\frac{1}{2}\left(1-\left(\frac{1}{5}\right)^2-\left(\frac{2}{5}\right)^2-\left(\frac{2}{5}\right)^2\right)+\frac{1}{2}\left(1-\left(\frac{4}{5}\right)^2-\left(\frac{1}{5}\right)^2\right)\right)\\ &= 0.62 - \left(\frac{1}{2}\left(1-\frac{9}{25}\right)+\frac{1}{2}\left(1-\frac{17}{25}\right)\right)\\ &= 0.62 - \frac{1}{2}\cdot\frac{24}{25}\\ &= 0.62 - \frac{12}{25} = 0.14.\end{aligned}$$

We conclude that the preferred split is S_1.

Computing the correct impurity for the parent node yields 2 points, and then 2 points per split. Points are removed for mistakes.

Exercise 4

In a) three classifiers each in b) (ii) would yield 4 classifiers and (i) would give $\binom{4}{2} = 6$.

Regarding c) the answer is no in (ii) there is for instance a small (upside down) triangle in the middle of the points which in which no class is dominating, but for (i) the different decision boundaries meet in the middle of the triangle, see Figure 1. Hence, for d) one could argue that (i) is better as in the entire feature space there is no ambiguity for the algorithm. We also note that the decision regions may vary from that of the figure slightly

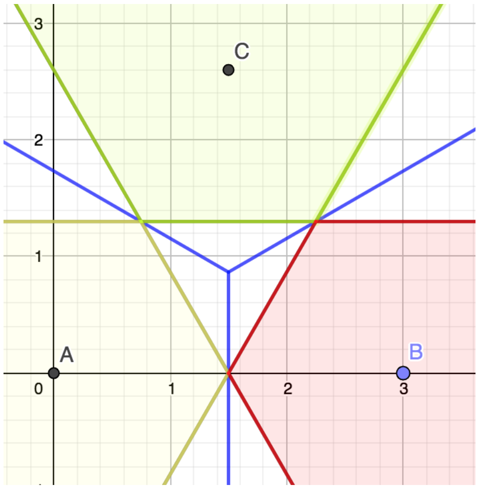

Figure 1: The blue lines are the decision boundaries for (i) and the regions are the decision boundaries for (ii)

depending on the implementation. But it is clear that in the region of the upside-down triangle, none of the classifiers in (i) are voting for these points.

The points are distributed as follows: a) and b) one point each, c) and d) two points each. You may get partial credit on d) based on your answer in c) even though your answer in c) was completely wrong.

Exercise 5

The distance between each point and $(150, 5, 4)$ is given by the vector

$$\text{distances} = (1.9, 23.8, 21.4, 52.4, 68.1, 11, 7.3, 13.7, 30.3, 17.5, 10.8, 53.1).$$

Hence, the three closest points are the first, the seventh and the eleventh. The prediction is thus $(4.3{+}3{+}4.1)/3 = 3.8$.

One alternative to the feature x_i with the desired properties is $u_i = |x_i - \bar{x}|$. We have $\bar{x} = 4.775$. Applying the algorithm on the 'new' matrix yields that the closest points are the first, sixth and seventh, *i.e* the prediction is $(4.3 + 3 + 3)/3 \approx 3.43$. Note that the point $(150, 5, 4)$ also have to be rescaled.

The major impact of the non-normalized data is that the size of the house is of (disproportional) importance as its range of values is much greater in absolute terms than the other features. The number of rooms has almost no bearing at all. Note that this is not necessarily something negative, and not normalizing is the same as assigning a certain weight to each feature, the problem is however, that you are not in control of this weighting procedure (or you may argue that all features are equally important).

Each subexercise give two points each. Points are removed for mistakes or incomplete information, *e.g.* the details on what kind of feature change was done in b).