Final exam, ML (MDS + BDMA + MIRI)

11th of June, 2024

Question 1

Choose correct alternative among (is / is not) and complete the following sentences:

1. Feature scaling (is / is not) important in gradient descent because ...
2. Feature scaling (is / is not) important in k-Nearest Neighbors because ...
3. Feature scaling (is / is not) important in the Multilayer Perceptron because ...
4. Feature scaling (is / is not) important in Random Forests because ...
5. Feature scaling (is / is not) important in least squares linear regression because ...
6. Feature scaling (is / is not) important in logistic regression because ...

1. Feature scaling is important in gradient descent because **the learning rate is unique for all weights**.
2. Feature scaling is important in k-Nearest Neighbors because **typically it uses Euclidean or Manhattan distance to locate nearest neighbors and if unscaled those features with bigger absolute values would dominate distances**.
3. Feature scaling is important in the Multilayer Perceptron because **it uses gradient descent to learn its weights**.
4. Feature scaling is not important in Random Forests because **the nature of the decisions made in internal nodes in trees apply to single features which are compared to thresholds that are set depending on values of the feature; scaling would not have any effect since other thresholds would be chosen**.
5. Feature scaling is not important in least squares linear regression because **we have a closed formula that gives optimal parameter values**.
6. Feature scaling is important in logistic regression because **typically gradient descent is used to learn its parameters**.

Question 2

In cross-validation, we use the mean validation error across folds as a metric to optimize hyper-parameters. Apart from computing the mean, we could look into other statistics or aggregates of these validation errors. Can you think of any suitable ones? For each metric you list, please state in what way and why it would be a reasonable one.

We have many options:

- Harmonic mean of the validation errors; this type of mean does not allow very high values to be compensated by very low values and could be a more robust measure if we want to choose a model that has better performance across all "folds".
- Median of the validation errors; as a more robust measure in the presence of extreme cases such as an unrepresentative fold.
- Maximum of the validation errors: this statistic prioritizes models that always have good performance, as we would choose the model with the smallest maximum error.
- Variance of the validation errors (less variance is better): this statistic prioritizes robust models or those with low variance. In the bias/variance trade-off, we would be choosing options with low variance and therefore it makes sense to minimize variance to minimize the generalization error.

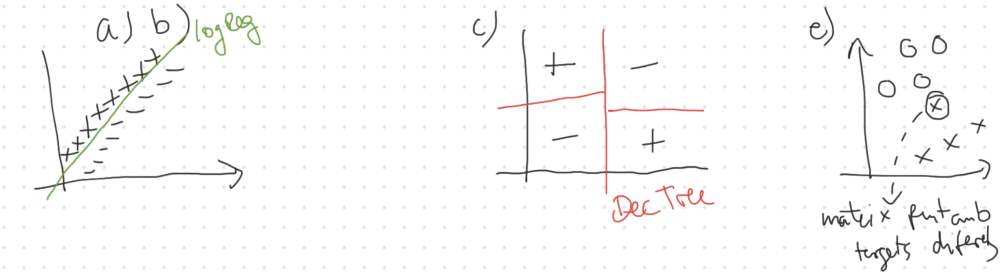
We could also consider combinations of these statistics.

In general, I have accepted responses where there is a more robust measure than the arithmetic mean (median, maximum, harmonic, geometric, and similar), and measures that somehow check the consistency of errors accross folds (variance, stdev, max-min).

Question 3

Whenever possible, invent a binary classification dataset with two numeric input features (you may draw it if you want), such that:

a) Logistic regression achieves perfect classification but a decision tree with maximum depth 1 cannot.
b) Logistic regression achieves perfect classification but a decision tree with maximum depth 2 cannot.
c) Logistic regression cannot separate the two classes, but a decision tree can.
d) Logistic regression separates the two classes, but a decision tree cannot.
e) Neither logistic regression nor a decision tree can separate the classes.



Things to note are:

- Cases a) and a) is linearly separable and thus logistic regression can find a separating line (green line). No tree can separate perfectly this dataset because it requires depth greater than 2.
- Case c) is not linearly separable thus a linar method such as logistic regression cannot perfectly separate the datapoints from different classes. A decision tree of depth 2 can do so however. In red you can see the decision boundary for a tree that achieves perfect classification.
- Case d) is not possible. If logistic regression can separate both classes, then a tree with enough depth will do so, too. (If there are no inconsistencies in the data such as duplicated points with different labels, then a tree with enough depth will find the separation).
- In case e) we have a duplicated example, with different labels, and thus no tree (nor deterministic method) will be able to separate them.