Methods of machine learning

Exercise sheet I

April 16th, 2025

In this exercise class we want to analyze the classical *wines data set* by k-nearest neighbor method:

1. **Task.**
   Load the data, take a look at the data documentation and display the three classes in a 2D plot using only the two features *alcohol* and *proline*.

2. **Task.**
   Display the different scales of the features by plotting parallel box plots. Over which range do the scales of the features differ? Now use a scaling (use MinMaScaler and StandardScaler) to standardize the data.

3. **Task.**
   Split the (scaled) data into training and test sets (80-20 split), but only use the two features *alcohol* and *proline* for now. Train k nearest neighbors with $k \in \{1, \dots, 10\}$ using `KNeighborsClassifier` from `sklearn.neighbors`. Compute the training and test error (or accuracy) w.r.t. 0-1 loss and display it graphically. Which choice of k performs best?

   In addition, display the decision boundaries for $k = 1$, $k = 10$, and your best choice of k in a 2D plot. Use the routine `DecisionBoundaryDisplay.from_estimator`.

4. **Task.**
   Now vary the metric. Choose k as before (optimal choice from previous task) and vary between *Minkowski, Manhatten*, and *cosine distance*. Again plot the decision boundarys and report the training and test accuracies.

5. **Task.**
   Now perform a grid search to find the best combination of $k \in \{2, 3, 4, 5, 6, 7\}$, metric (*Minkowski*, *Manhatten*, *cosine*), and weights (*uniform*, *distance*). Use `GridSearchCV` from `sklearn.model_selection` and the accuracy for evaluation. Estimate the generalization error (w.r.t. 0-1 loss) for the chosen best set of hyperparameters.

6. **Task.**
   Finally, increase the number of features starting with *alcohol* and *proline* and adding additional features one by one. Plot the achieved test error versus the number of features. Repeat the same for the original non-standardized features. What do you observe?

   Is your observation in contradiction to the curse of dimensionality?