The Mathematics of Machine Learning
Homework Set 7

Due 22 May 2024 before 13:00
via Canvas

You are allowed to work on this homework in pairs. One person per pair submits the answers via Canvas. Make sure to put both names on the submission.

This week, you will need to submit two files: one file containing your answers to the questions below, and one notebook.

1 K-Means Clustering for Feature Engineering

Read through the lesson "Clustering with K-Means" on Kaggle: `https://www.kaggle.com/code/ryanholbrook/clustering-with-k-means/tutorial`. It uses K-means clustering, which is a method for *unsupervised* learning, to engineer extra features in the context of regression, which is an instance of *supervised* learning. NB The notebook refers to 'means' as 'centroids'. Then answer the following questions:

1. [1 pt] The lesson suggests to tune the number of clusters K using cross-validation, but in class you were told that you cannot use cross-validation to tune K in K-means clustering, because it is not even well-defined. How can these conflicting suggestions be reconciled?

At the bottom of the Kaggle lesson, there is a link to "Add a feature of cluster labels" under the heading "Your Turn". Use this link to continue to the second part of the lesson, which consists of a notebook, and complete all questions in the notebook.

2. [1 pt] Download the completed notebook via `File` → `Download Notebook`, and submit it along with your answers to the other questions. NB. The notebook has provisions to check your answers, so make sure to do that.

3. [1 pt] The notebook emphasizes that K-means is sensitive to the features having different scales: features that have a large scale will be treated as much more important than features that have a small scale in the clustering decisions. Explain why this happens.

The notebook runs K-means on the x-values in both the train set and the test set jointly. Although there might exist applications in which the x-values for the test set are available ahead of time, this violates the normal protocol in which all training has to be done on the train set. For the next questions we therefore want to stick with the standard protocol, and assume that we run K-means only on the x-values in the train set.

4. [1 pt] What would be a natural way to assign clusters to the data points in the test set?

5. [1 pt] Under "Cluster-Distance Features" the notebook discusses alternative features based on K-means. If these are used, how many features are added to each data point?