Machine Learning

Exercise Sheet 7

Winter Term 2023/2024
Prof. Dr. Niels Landwehr
Dr. Ujjwal

Available: 21.12.2023
Hand in until: 11.01.2024 11:59am
Exercise sessions: 15.01.2024/17.01.2024

Task 1 – Nearest Neighbor Regressor [15 points]

In this task, we assume that an instance $\mathbf{x} \subseteq \mathcal{A}$ is a subset of a set of possible elements $\mathcal{A} = \{a, b, c, d, e, f, g\}$. That is, $\mathbf{x} \in \mathcal{P}(\mathcal{A})$, where $\mathcal{P}(\mathcal{A})$ is the set of all subsets of the set $\mathcal{A}$ (the so-called *power set*). We study regression models of the form $f_{\mathcal{D}} : \mathcal{P}(\mathcal{A}) \rightarrow \mathbb{R}$. The model $f_{\mathcal{D}}$ is given by a K-nearest neighbor regressor using a distance function or similarity function as discussed below.
Assume the following set $\mathcal{D}$ of $N = 5$ training instances $\mathbf{x}_1, ..., \mathbf{x}_N$:

n	$\mathbf{x}_n$	y_n
1	{a,b,d,e}	3
2	{c,d}	1
3	{a}	2
4	{a,f,g}	3
5	{e,f}	4

a) Assume the K-nearest neighbor regressor is based on the Hamming distance, and produces the following predictions on novel test instances: $f_{\mathcal{D}}(\{a, b, d\}) = 2$ and $f_{\mathcal{D}}(\{f\}) = 3$. Which parameter K was used in the model? Show that with this parameter K these predictions are indeed obtained.

b) Now assume the K-nearest neighbor regressor is based on the Jaccard similarity, and produces the following predictions on novel test instances: $f_{\mathcal{D}}(\{a, b, d\}) = 2.5$ and $f_{\mathcal{D}}(\{f\}) = 3.5$. Which parameter K was used in the model? Show that with this parameter K these predictions are indeed obtained.

Task 2 – Visualizing Decision Surface of K-NN (programming) [15 points]

You are provided with a CSV file ushape.csv and an IPython notebook knn-exercise7.ipynb. The existing code in the IPython notebook already gives you the code to read the CSV file and to plot its contents.
Your tasks are as follows :

a) For all the points on the 2D plane bounded by the points in the file ushape.csv, plot their class using K-NN classification for values of K=1, 2 and 3. This will show you the decision surface for the points in the 2-D plane. Based on this, outline the behavior of choosing different values of k. How the choice of k, affect the smoothness of the decision surface?

b) Implement K-NN classification yourself without using scikit-learn or any helper library which comes with its own implementation of K-NN classification.

Task 3 – Levenshtein Distance [10 points]

Use the dynamic programming algorithm ("Wagner-Fischer-Algorithm") presented in the lecture to compute the Levenshtein distance between the sequences $(i, n, t, e, n, t, i, o, n)$ and $(e, x, e, c, u, t, i, o, n)$. In your solution, please give the full matrix as computed by the algorithm. Also use the matrix to read off the Levenshtein distances between (i, n, t, e, n, t) and (e, x, e) and the distance between (i, n, t) and (e, x, e, c, u, t).