

# Introduction to Machine Learning

## Assignment 4

Group 31

Stijn Kammer (s4986296) & Ramon Kits (s5440769)

October 12, 2022

### 1. INTRODUCTION

### 2. METHODS

Vector Quantization makes use of prototype vectors to represent the data and is often used for identification and grouping of clusters in similar data. To use Vector Quantization, we first need to define a set of prototype vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ . The prototype vectors are chosen by selecting  $k$  data points from the data set. After which present a single example  $\mathbf{x}$ , we can find the closest prototype vector  $\mathbf{v}_i$  by minimizing the distance between  $\mathbf{x}$  and  $\mathbf{v}_i$ . The distance between  $\mathbf{x}$  and  $\mathbf{v}_i$  is defined as the Euclidean distance between the two vectors. When the closest prototype vector  $\mathbf{v}_i$  is found, we move the prototype vector  $\mathbf{v}_i$  towards  $\mathbf{x}$  by a fraction  $\eta$  of the distance between  $\mathbf{x}$  and  $\mathbf{v}_i$ . Repeat this process for all the examples in the training set and the prototype vectors will converge to a set of vectors that represent the data well. The prototype vectors are then used to classify new examples.

Quantization error is the error that is made when using the prototype vectors to classify new examples. The quantization error is defined as the average squared distance between the prototype vectors and the examples in the training set. The quantization error is minimized by moving the prototype vectors towards the examples in the training set. The quantization error is also minimized by choosing a good set of prototype vectors. The prototype vectors should be chosen in a way that they are spread out over the data set. This is done by choosing the prototype vectors to be the  $k$  data points that are furthest apart from each other. The prototype vectors should also be chosen in a way that they are representative of the data set. This is done by choosing the prototype vectors to be the  $k$  data points that are closest to the mean of the data set.

### 3. LEARNING CURVES

### 4. TRAJECTORIES OF PROTOTYPES

### 5. DISCUSSION

### 6. BONUS