Classification with clustering algorithms

Posted Sep 27, 2022

By Davide Evangelista

2 min read

What is classification?

Consider, as an example, the MNIST dataset we used to test PCA and LDA clustering algorithm. Here, we have a dataset $X \in \mathbb{R}^{d \times N}$ representing handwritten digits, and a vector $Y \in \mathbb{R}^N$ such that the i-th element of Y is the digit represented in the i-th column of X. When this is the case, we say that the numbers collected in Y are the *classes* associated with the elements in X.

An important task in Machine Learning is to understand the semantic relationship between a datapoint x and the corresponding class y. This task is called **classification**.

> ***Classification:*** A classification algorithm is a model mapping a datapoint $x \in \mathbb{R}^d$ to the corresponding class y, chosen among a **finite** set $\mathcal{C} = \{C_1, \dots, C_K\}$.

In practice, when developing a ML classification algorithm, we are required to implement a rule $f_\theta(x)$, associating x to a *potential* class $\hat{y} = f_\theta(x)$. The classification guess is correct if the predicted class $\hat{y}$ equals the right class y.

Decision boundaries

Most of the classification algorithm is based on the concept of **decision boundaries**. The idea is to *learn* a curve (either a straight line or a smooth, non-linear curve), called a *boundary*, from the training set and then, given a test datapoint, it will be classified to the first class if it is below the curve, to the second class if it is above the curve. This idea is summarized in the following Figure.

Measuring the accuracy

The accuracy of a classification algorithm is easy to compute. We can just consider the percentage of correct guesses of the model over the test set. Mathematically, this is implemented as

$$Acc(f_\theta) = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \mathbb{1}(f_\theta(x_i) = y_i)$$

where

$$\mathbb{1}(f_\theta(x_i) = y_i) = \begin{cases} 1 & \text{if } f_\theta(x_i) = y_i \\ 0 & \text{if } f_\theta(x_i) \neq y_i \end{cases}$$

the higher the accuracy of a classification algorithm, the better it is in practice.

Classification with clustering

Clustering can be used to define a classification algorithm. In particular, here, the idea is that since by definition the datapoints that are semantically similar are close together in the projected space of a clustering algorithm, we can hope that points living in the same class, will be mapped inside of the corresponding cluster. Consequently, we can define a classification algorithm in the following way:

- Compute the projection matrix $P \in \mathbb{R}^{k \times d}$ associated with a clustering algorithm;

- Given a new datapoint $x \in \mathbb{R}^d$, project it into the cluster space by $z = Px$;
- Compute the distance between z and each cluster centroid $c_i, i = 1, \dots, K$ as $d_i = ||z - c_i||_2^2$;
- Classify x to be the class of the centroid such that d_i is the smallest.

Note that the algorithm above naturally defines decision boundaries, as descripted in the following Figure.

Lab2 , teaching

LDA PCA clustering dimensionality reduction SVD unsupervised learning classification

Share:

Further Reading

Sep 25, 2022

Dimensionality Reduction with PCA

Dimensionality Reduction While working with data, it is common to have access to…

Sep 26, 2022

Clustering with Linear Discriminant Analysis (LDA)

Why PCA is not sufficient? Since PCA is an unsupervised learning technique, the lack…

Sep 24, 2022

A (very short) introduction to Machine Learning

Definition: Machine Learning (ML) is the set of all the techniques and algorithms able t…

OLDER
Clustering with Linear Discriminant Analysis (LDA)

NEWER
Gradient descent