K-nearest neighbours

Advanced Topics in High-Performance Computing

Faisal Qureshi
Classification

Given some data with corresponding labels, learn a function to predict a previously unseen data point.

Given data \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\), learn a function \(y = f(x)\) that predicts the label \(y\) for a previously unseen example \(x\). For the following example, \(y_i \in \{0, 1\}\).
K-nearest neighbour

Prediction function: label of the nearest training example.

Training examples from class 1

Training examples from class 2
K-nearest neighbour

- No training needed
- Need to store the entire training set
- Need a distance function to compute the similarity between the new data point (the query) and those in training set

Voronoi partitioning of feature space for two-category 2D and 3D data

(Duda et al.)
K-nearest neighbour

1-nearest Neighbour

3-nearest Neighbour

5-nearest Neighbour
K-nearest neighbour discussion

- K-nearest neighbour is a *non-parametric* learning algorithm that can be used for both classification and regression.
- K-nearest neighbour can achieve very high capacity
  - This implies that K-nearest neighbour can achieve very low training error, albeit at a very high computational cost
  - This also means that K-nearest neighbour may generalize very poorly
- K-nearest neighbour is also unable to learn if one feature is more discriminative than another feature.
- Reliance on local constancy
  - We shall see that many machine learning algorithms suffer from smoothness prior, and thus these fail on AI-level tasks, such as image recognition. Deep learning is in part motivated to relax local constancy and smoothness prior assumptions.
Curse of dimensionality

- As the dimensions of the data increases, the number of configurations of interest often grows exponentially.

Curse of dimensionality (Goodfellow et al., 2017)
Summary

We looked at K-nearest neighbour for classification and regression. K-nearest neighbour is a non-parametric learning algorithm with very high capacity, which leads to very low (zero, really) training error, but very high test errors. K-nearest neighbour suffers from the curse of dimensionality.

Implementation

Check out the Python Sklearn module, which implements k-nearest neighbour classification and regression:


Readings

▶ Sec. 5.7.3, Goodfellow, et. al., 2017