Johannes Gutenberg-Universität Mainz
Institut für Informatik
Data Mining
Prof. Dr. Stefan Kramer

Machine Learning - Sheet 7

14.06.2018
Deadline: 21.06.2018 - 16:00

Task 1: Lazy Learning *(5 Points)*
Read chapter 8 of Machine Learning book [1] and make yourself familiar with the concept of lazy and eager learning. Suggest a lazy version of the eager decision tree learning algorithm ID3. What are the advantages and disadvantages of your lazy algorithm compared to the original eager algorithm?

Task 2: Curse of Dimensionality *(5 Points)*
The nearest neighbor method breaks in high-dimensional spaces, because the "neighborhood" becomes very large for the Euclidean distance. This is called the curse of dimensionality. Suppose we have 5000 points uniformly distributed in the n-dimensional unit hypercube $C_n := [0, 1]^n$ and we want to apply the 5-nearest neighbor algorithm. Suppose our query point is at the origin $(0, \dots, 0)$, so, on average, we need to search 5/5000 of the hypercube's volume to capture the 5 nearest points.

(a) Assume $n = 2$, i.e. C_n is the unit square. What is the side length d of the square $[0, d]^2$ that on average captures the five nearest points to the origin?

(b) What is the side length d of the n-dimensional hypercube $[0, d]^n$ that on average captures the five nearest points to the origin?

(c) For which number of dimensions n do we need a hypercube $[0, d]^n$ whose side length is larger than half the side length of C_n (i.e. 0.5) to capture the five nearest points?

Task 3: Voronoi Diagram *(5 Points)*
Given are the following instances (attributes from $\mathbb{R}^2$, class label from $\{\oplus, \ominus\}$):

$$(3,1)\ominus, (9,2)\ominus, (5,3)\ominus, (8,5)\ominus, (7,7)\ominus, (1,4)\oplus, (3,4)\oplus, (1,8)\oplus, (4,9)\oplus, (5,6)\oplus$$

(1) Draw the points in the $[0, 10] \times [0, 10]$ square.

(2) Draw the Voronoi diagram associated with the set of points. Use Euclidean distance as your distance measure.

(3) Draw the decision boundary of the 1-NN classifier between the two classes using Euclidean distance.

(4) Why is it impractical to store the Voronoi diagram in order to speed up queries for k-nearest neighbor?

If you want, you can automate some of the steps.

References

[1] Tom M. Mitchell. *Machine learning.* McGraw Hill series in computer science. McGraw-Hill, 1997.