Exercise 19: Instance based learning: k-nearest-neighbour

(a) Briefly sketch how a classification with \( k \)-nearest-neighbour is done.

(b) \( k \)-NN belongs to the so called lazy learning methods. What does this mean.

(c) When using \( k \)-NN a distance value is needed, which calculates the distance between. Often the so called cosine similarity is used. The cosine similarity is defined as:

\[
\cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| \cdot |\vec{b}|},
\]

where \( \vec{a} \) and \( \vec{b} \) are the two instances in vector space.

Given are the following four training instances:

\[
\vec{a} = \begin{pmatrix} 0 \\ 3 \\ 5 \end{pmatrix}, \quad \vec{b} = \begin{pmatrix} 3 \\ 3 \\ 8 \end{pmatrix}, \quad \vec{c} = \begin{pmatrix} 2 \\ 6 \\ 1 \end{pmatrix}, \quad \vec{d} = \begin{pmatrix} 4 \\ 3 \\ 0 \end{pmatrix}
\]

To which class would the instance \( \vec{x} = (5, 8, 0)^T \) be assigned, if \( \vec{a} \) and \( \vec{b} \) belong to class X and \( \vec{c} \) and \( \vec{d} \) belong to class Y (2-NN-classification with cosine similarity as distance calculation)?

Exercise 20: Multiway splits

(a) Split the example data \(^1\) by the attribute temperature. Use the information gain criteria to determine the best splitting points.

Hint: One possibility to speed up the process to determine the best split point can be to define functions in \( R \). Perhaps not all calculations are necessary. If you skip some calculations reason why.

(b) Do a supervised discretization in \textit{Rapidminer} and compare the results with your calculated splits. What is the difference? If the results are different try to explain how \textit{Rapidminer} gets these results.

\(^1\)http://www.is.inf.uni-due.de/courses/im_ws15/uebung/data_a20.csv