Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Eirini Ntoutsi
PD Dr. Matthias Schubert

Knowledge Discovery in Databases II

WS 2015/2016

Übungsblatt 8: Streams

Aufgabe 8-1 CF-Tree

Given the following one dimensional data set:
1, 2, 5, 10, 4, 7, 15, 9, 3

In the following, use Euclidian distance L_2-Norm).

(a) Construct a CF-Tree with $B = 2$ and $L = 3$ by inserting each value in the given order. As threshold T use 2. Draw the resulting CF-tree after each insertion.

(b) Why does the resulting tree depend on the insertion order?

Aufgabe 8-2 Cohen's Kappa

Gegeben seien die folgenden Konfusionsmatrizen zu den Zeitpunkten $t = 1, 2, 3$:

$t = 1$

	positiv	negativ
positiv	37	14
negativ	17	32

$t = 2$

	positiv	negativ
positiv	65	8
negativ	7	20

$t = 3$

	positiv	negativ
positiv	90	4
negativ	5	1

Berechnen Sie Accuracy und Cohen's Kappa, und vergleichen Sie die Ergebnisse.

Aufgabe 8-3 Naive Bayes on Streams

In this exercise, we will examine the effect of aging older instances when learning from a stream. Therefore, we will implement two naive Bayes classifiers. The first, will incrementally train without aging old training instances. The second one will give a new training instance as a fixed weight in the training set (e.g. 1 %). Thus, older instances will contribute less and less to the given class models.

(a) Download the templates python_Stream_NB.py and Random_stream.py. Study and visualize the data stream.

(b) Implement a class Naive_Bayes with methods for updating the model by on instance and predicting the class of an instance. Though the classifier can be trained in an incremental way, all instances are equally weighted when computing the model. For the model, assume that each class is represented by an axis parallel Gaussian distribution having a mean and a variance. The class prior is estimated by the relative portion of training instances for each class.

(c) Implement a class Stream_Naive_Bayes which also implements a naive Bayes classifier based on axis parallel Gaussians. For this class, the update method weighs each new instance as if it would represent a fixed percentage of the training set. Thus, which each update the impact of older instances is decreased.

(d) Compare both classifier on the first 5000 instances of the data stream.

(e) What does the fixed impact factor in the stream classifier regulate?