Information Mining - winter semester 2016

Exercise sheet 9

Exercise 1: Support-Vector-Machines

Support-Vector-Machines (SVMs) have proven to be a successful method fr classifying and are well established.

maximum margin hyperplane
support vectors

(a) Make a sketch of the geometric principle behind the SVMs.

(b) Explain the role of the parameter C. Indicate the behaviour of SVM for large and small C values.

C : regularization parameter

The larger the C value, the less likely the classifier is to allow classification errors ("outliers"). If the C value is too large, it will cause overfitting.
If the C value is too small, the classifier will "don't care" too much about the classification error, so the classification performance will be poor.

Exercise 2: Combination of multiple models

In the lecture methods to combine multiple models are shown. Explain briefly the underlying idea behind the following principles: (i) Bagging, (ii) Boosting and (iii) stacking.

bagging : voting
boosting : weighted voting , let expert models have stronger power.
stacking : learn a model of models ,predict from level-0

Exercise 3: Clustering: k-Means

Within information retrieval the vector space model is very important. All documents will be represented as vectors of terms and the parts of the vector represent the weight of the term inside the document.

This model can be used for clustering. Assuming we have the terms T = {`house`, `car`}; the corresponding vector space has 2 dimensions `house` and `car`. We have the following 5 documents, who can be described as vectors $\vec{d}$.

$$\vec{d_1} = \begin{pmatrix} 0,8 \\ 0,9 \end{pmatrix}, \vec{d_2} = \begin{pmatrix} 0,9 \\ 0,65 \end{pmatrix}, \vec{d_3} = \begin{pmatrix} 0,5 \\ 0,5 \end{pmatrix}, \vec{d_4} = \begin{pmatrix} 0,2 \\ 0,25 \end{pmatrix}, \vec{d_5} = \begin{pmatrix} 0,25 \\ 0,1 \end{pmatrix}$$

The document d_1 has the weight 0,8 for `house` and the weight 0,9 for `car`.

(a) Cluster the documents with k-Means-Clustering. Two clusters should be created and as initial seeds d_4 and d_5 should be used. {d1d2d3d4} , {d5}

(b) Graphically sketch the document space and the movement of the centroids. What can be seen?

(c) Calculate for your resulting cluster the *Purity*. Assume the following manual clustering: $C_1 = \{d_1, d_2, d_3\}$ and $C_2 = \{d_4, d_5\}$. 100%

Exercise 4: Clustering: k-Means on Text

Now from above you know how to treat text. Assume you have the following 10 sentences:

- The weather is cold.
- The weather is horrible.
- The temperature is very high.
- Duisburg has always horrible weather.
- The weather conditions and the daily temperature change four times a day in Duisburg.
- Football is the most known sport activity in the world.
- Although football is a game for fun there are always violances in it too.
- There is a saying that sport is mord and indeed performed on bad weather conditions it might harm someone.
- Many sport activities such as football, handball, etc. are performed with a ball.
- Also swimming, running, etc. are well knowing sport disciplines.

Do a k-means clustering on these sentences. Select $k = 2$. Note you must determine your term/vocabulary list. According to this you can create your term vectors for each sentence. Each vector dimension representing a term should contain the count how many times that term appears in the sentence. Make sure you delete punctuations. Use automatic clustering, e.g. on R, and visualize your results.

R 软件，不用做