Pattern recognition - 9th lab
Naive Bayes

Viktor Kocur
viktor.kocur@fmph.uniba.sk

DAI FMFI UK

21.4.2020

Bayes equation

Bayes equation

We will use the Bayes equation:

$$P(\omega_i|\vec{x}) = \frac{P(\vec{x}|\omega_i)P(\omega_i)}{P(\vec{x})} \tag{1}$$

Naive Bayes

Our classifier is naive and assumes that features are independent:

$$P(\vec{x}|\omega_i) = \prod_k P(x_k|\omega_i) \tag{2}$$

Classifier

Classifier

We classify by finding the class with the highest probability:

$$pred_i = \arg\max_i \left(\frac{P(\vec{x}|\omega_i)P(\omega_i)}{P(\vec{x})} \right) \tag{3}$$

$$= \arg\max_i \left(P(\vec{x}|\omega_i)P(\omega_i) \right) \tag{4}$$

$$= \arg\max_i \left(P(\omega_i) \prod_k P(x_k|\omega_i) \right) \tag{5}$$

Classifier

Calculating values

We assume categorical features. Therefore for each k the x_k can only have finitely many values. Let us have a training set of size N. We denote the amount of elements in the category ω_i as N_i and the amount of elements which are in the category ω_i for the k-th feature which has value of v as $N_{i,k,v}$. We can then define:

$$P(\omega_i) = \frac{N_i}{N} \tag{6}$$

$$P(x_k = v|\omega_i) = \frac{N_{i,k,v}}{N_i} \tag{7}$$

Classifier

age	income	student	credit rating	buys computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Exercise

Calculate which class would a client with random features belong to according to naive bayes.

Classifier

Numerical features

In case that the feature is numeric we cannot apply the calculation from the previous slide. We will therefore need to estimate the probability $P(x_k|\omega_i)$ by some distribution function.

Parametric methods

If we use a parametric method we select a distribution and fit its parameters to the data.

Non-parametric methods

We can also create a distribution function which is calculated based on the points in the training set in the neigborhood of the point we are interested in.

Matlab

fitcnb

Mdl = fitcnb(T,'column_name') - retursn a naive Bayes classifier for the table T and classification target in the column column_name.

Malab - Table type

Working with tables:
https://www.mathworks.com/help/matlab/tables.html

This is the most important part for us:
https://www.mathworks.com/help/matlab/matlab_prog/access-data-in-a-table.html

Naive Bayes on table data

Na dátach

```
load census1994
Mdl = fitcnb(aduldata, 'salary');
```

Exercise

Determine the accuracy of the classifier by using Mdl.predict on the table adulttest and compare the results.

Matlab

fitcnb

Mdl = fitcnb(X,y) - returns naive Bayes classifier for non-table data.

Exercise

Test the naive Bayes classifier on the fisheriris database.

Exercise

Use the data from the 6th lab and display the classification boundary using your modified version of showSVM from the same lab.