CIS 635 Data Mining

Homework 5

Description

This homework is similar to the previous one, except that you will be going through the process of building a naive Bayes model. Like last time, the process is different for nominal vs numeric data so there will be tasks for each.

Instructions

Part 1 – manual calculations

The table to the right (same as previous hw) contains information about patients in a study for a particular virus. The attributes consist of id, gender, age group, pulse/heart rate, test and the class, virus. The test is the results of a medical test that can range from 1 to 100. It is thought that gender, ageGroup, pulse and/or test may be predictive of the presence of the virus. Do the analysis to answer the questions below.

id	gender	ageGroup	pulse	test	virus
101	M	1	67	45	Y
102	M	3	68	58	N
103	F	2	73	72	N
104	M	2	77	67	N
105	F	1	62	39	Y
106	F	3	81	61	N
107	M	1	70	73	N
108	F	2	81	52	Y
109	M	3	105	66	N
110	F	2	63	47	Y

1. calculate the prior probabilities for the class values
2. calculate the likelihoods for
    a) gender
    b) ageGroup
3. calculate the likelihoods (mean and std dev.) for
    a) pulse
    b) test

Part 2 – classifying instances

In this part, we are using a different data set from the one above with more instances (thousands) but the same attributes. It is also not the same as the data set used in the next exercise below. You are not allowed to see the data but you can see the model (prior probabilities and likelihoods) that was built from the data set to the right.

naive Bayes model

	virus = n	virus = y
prior	0.47	0.53
p(gender=f | v)	0.65	0.39
p(gender=m | v)	0.35	0.61
p(ageGroup=1 | v)	0.62	0.20
p(ageGroup=2 | v)	0.25	0.32
p(ageGroup=3 | v)	0.13	0.48
p(pulse | v)	μ = 62 σ = 7	μ = 75 σ = 8
p(test | v)	μ = 48 σ = 11	μ = 52 σ = 12

Using just the model, do the following:

1. state whether a patient is more or less likely to have the virus without knowing anything about the patient.
2. state whether a patient who has the virus is more likely to be male or female.
3. calculate the numbers (numerator of posterior) to categorize the following instances (for pulse and test, use the R function dnorm to calculate the probability) :
   a) {2031, f, 2, 66, 50 ,?}
   b) {2032, m, 1, 90 ,81 ,?}
   c) {2034, m, 3, 90 ,55 ,?}
4. exercise can often lower someone's pulse rate. What advise would you give a friend to help them to avoid getting the virus?

Part 3 – building a model in R

You have been provided three data sets with many instances (same attributes as above but different data). The data sets are hw05a.txt, hw05b.txt and hw05c.txt.

The directions below describe how to process the data, create models and examine the results. Write a function to do these steps and then call the function three times, each time with a different data set.

1. for each data set, perform the following actions
   a) read the data from the file into a data matrix
   b) make gender, ageGroup and virus factors
   c) use naiveBayes to build a model using virus as the class and all of the attributes (except id)
   d) display the model (e.g. the priors and likelihoods)
   e) identify the one (if only one) or two most predictive attributes

2. answer the following general questions

a) what kind of data distribution does naive Bayes assume?

b) would it work for other distributions (observe the ones below for examples)?