CS909: 2013-14

Week 5: Naïve Bayes

1. View the HairEyeColor dataset. What is the predicted class here?

What are the parameters of a Naïve Bayes classifier model that would classify this data?

2. Convert the HairEyeColor dataset to an expanded data frame:

```
> HairEyeColorDF <- as.data.frame(HairEyeColor)
# expanded data frame
> HEC <- HairEyeColorDF[rep(row.names(HairEyeColorDF),
HairEyeColorDF$Freq), 1:3]
```

Calculate the parameters in 1, using R. [Hint: Use the table() function]

3. Write a function NB that, given a data frame with discrete values and a class, returns the above parameters.

Verify your answers for the HairEyeColor dataset by using the Naïve Bayes classifier in R package e1071.

4. What would happen if one of the features were zero, given your Naïve Bayes function? E.g. if there were no red haired men in the dataset?

How could you remedy this?

5. Recall Exercise 3 and the iris dataset with missing values. This time, rather than filling in the missing values using a mean or median, use Naive Bayes to help you with the task of finding missing values. (We recommend you use the R package klaR and the function NaiveBayes.).

What steps do you need to take?

What assumptions have you made here and how do they influence the final classification into species?

Submission deadline: Midday, Thursday 13th February.