# CS909: 2013-14

**Week 4: Decision trees**

1. Given the Loan dataset below, manually apply the ChiMerge algorithm to discretise the Income variable. Use 0.05 as your significance threshold.

| Income | Loan |
|---|---|
| 12 | Y |
| 13 | Y |
| 14 | Y |
| 12 | N |
| 14 | Y |
| 16 | Y |
| 18 | N |
| 33 | Y |
| 22 | N |
| 24 | N |
| 46 | N |
| 53 | N |
| 24 | N |
| 19 | N |
| 25 | N |
| 32 | Y |
| 33 | Y |
| 37 | N |
| 21 | N |
| 25 | Y |

2. Write an R function `disc()` to discretise a dataset using equal width binning. It should take a data frame `dataset` and the number of bins as arguments and return `dataset` with non ordinal attributes discretised.

   Load the Loan dataset into R and use your function to discretise it. Compare the results with those you obtained in 1. Explain the differences you observe.

3. Manually generate the decision tree for the Play dataset below. Use Information Gain as your split measure.

| Outlook | Temp | Humidity | Wind | Play |
|---|---|---|---|---|
| Sunny | Hot | High | Weak | No |
| Sunny | Hot | High | Strong | No |
| Overcast | Hot | High | Weak | Yes |
| Rain | Mild | Normal | Weak | Yes |
| Rain | Cool | Normal | Weak | Yes |
| Rain | Cool | Normal | Strong | No |
| Overcast | Cool | High | Strong | Yes |
| Sunny | Mild | Normal | Weak | No |
| Sunny | Hot | Normal | Weak | Yes |
| Rain | Mild | Normal | Strong | Yes |
| Sunny | Cool | Normal | Strong | Yes |
| Overcast | Mild | High | Strong | Yes |
| Overcast | Hot | Normal | Weak | Yes |
| Rain | Mild | High | Strong | No |

   Sketch out the resulting decision tree and write out the equivalent rule set.

4. Install the rpart and rpart.plot R packages and read the documentation on the rpart() and rpart.plot () functions. Explain the significance of the argument control=rpart.control(minsplit=x) on the behaviour of rpart().

   Use the rpart() function to generate the decision tree for the Play dataset. Justify your choice of minsplit.

   ```
   > playtree.ig<-rpart(Play ~ Outlook + Temp + Humidity + Wind,
   data=play, control=rpart.control(minsplit=2),
   parms=list(split="information"))
   ```

   minsplit=2 will make rpart continue expanding each branch to its full extent.

   Use the rpart.plot() function to visualise the decision tree.

   Account for the difference between decision tree generated by rpart() and the decision tree you generated manually.

**Submission deadline**: Midday, Thursday 6th February.