RMIT University
School of Science
COSC2110/COSC2111 Data Mining
Tutorial Problems Week 12

1. The following heights(cm) of a number of sportspeople were measured:

Height (cm)	Sport
214	Basketball
200	Basketball
210	Football
213	Basketball
180	Football
200	Football
175	Cricket
170	Cricket
210	Basketball
185	Cricket
195	Basketball
175	Football
190	Cricket
198	Football
184	Cricket

   (a) What are the prior probabilities for each sport, ie $P(Basketball)$, $P(Football)$ and $P(Cricket)$?

   (b) What are $P(Height|Basketball)$, $P(Height|Football)$ and $P(Height|Cricket)$?

   (c) Using Bayes rule, find the posterior probabilities $P(Basketball|height)$, $P(Football|height)$ and $P(Cricket|height)$?

   (d) Sketch the posterior probabilities [You can ignore the denominator]

   (e) [Optional] Use a program like gnuplot, to draw the distributions. (See `/KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/code-and-scripts/gnuplot-normal` for an example of using gnuplot)

   (f) How would a person who is 190 cm tall be classified?

   (g) Work out an approximate error rate.

   (h) Is this a good classifier?

   (i) Could the classifier be improved by getting the heights of 1000 sportspeople?

2. Suppose we want a classifier for 'Flu' and 'Well'. We find 100 people who are well and 100 who have the flu and build a classifier based on the univariate normal distribution.

   (a) What is wrong with this procedure?

   (b) What if we built a decision tree?

3. Consider the following data from a factory:

Run	Operator	Machine	Length	Overtime	output
1	Joe	a	51	no	high
2	Sam	b	85	yes	low
3	Jim	b	63	no	low
4	Jim	b	39	no	high
5	Joe	c	32	no	high
6	Sam	c	47	no	low
7	Joe	c	63	no	low
8	Jim	a	70	yes	low
9	Jim	a	51	yes	??

   (a) Using only the nominal attributes of runs 1 to 8, how will case 9 be classified by a naive Bayes classifier.

   (b) How would case 9 be classified by naive Bayes if all attributes are used?

   (c) How does the result of the previous question change if the Laplace correction is used?