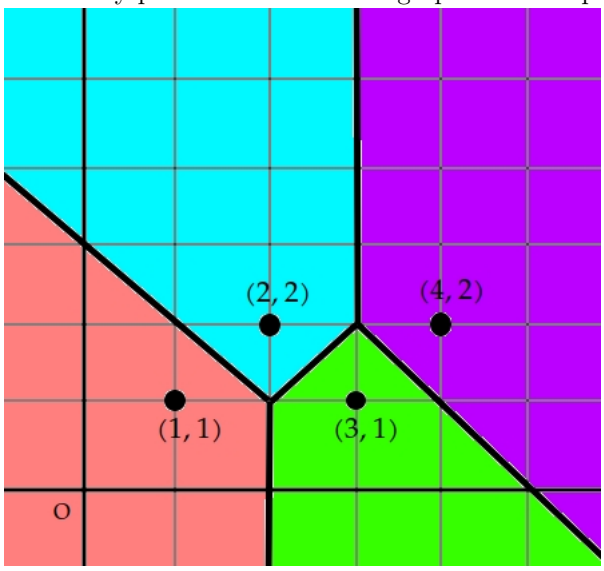
B365 Homework 5

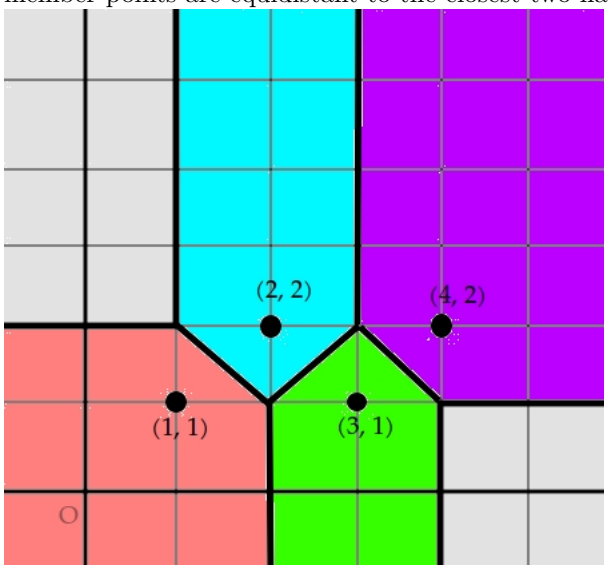
Zac Monroe

October 2018

1. (a) Here is my partitioned Euclidean graph for these points.



(b) Here is my partitioned Manhattan graph for these points. The light grey regions are regions whose member points are equidistant to the closest two named points.



2. (a) See prob2.r.

   (b) Source prob2.r in R to get a printout of my error rate for this part.

   (c) Source prob2.r in R to get a printout of my error rate for this part. It tends to be very similar to that of the last part, around 0.05.

3. (a) Look at prob3.r to see how I did this. Sourcing prob3.r into R will give a printout of the error rate I got for this classifier.

   (b) I would estimate an error rate of $\sim$0.5125101, my error rate from performing leave-one-out cross-validation, for outside data. Again, this is the number that prints out for part (a) when sourcing prob3.r into R.

   (c) Prior distributions over the four classes/brands are shown in a printout when sourcing prob3.r into R.

   (d) It appears that the price data and the consumer's purchase are indeed independent. This is odd but not unexpected.

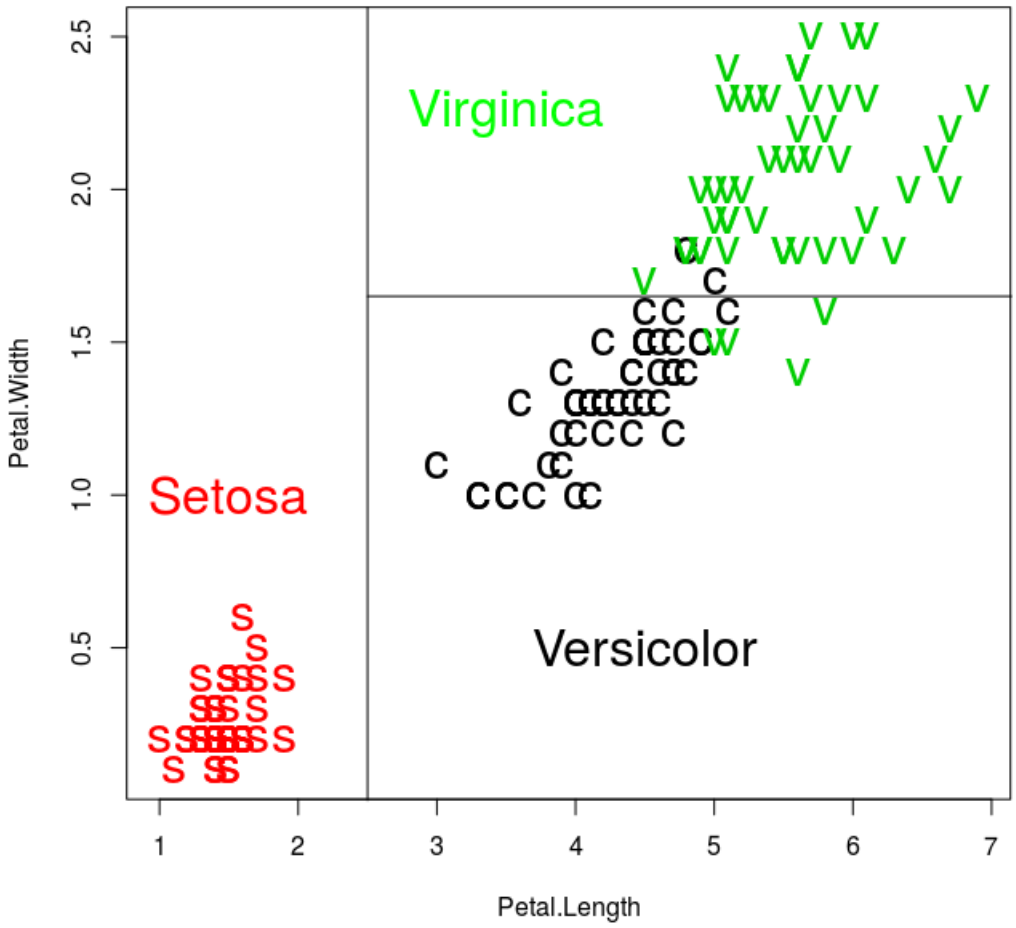
   (e) Consumers' purchase choices (at least in this case) are likely influenced by past experience and positive brand association, i.e. it is likely that many consumers made their choice (e.g. Heinz) based solely on the brand's name, disregarding price.

   (f) In terms of error rate, the best classifier for this data would probably be one that looks only at prior distributions of brand purchases, i.e. always choosing the previously most-chosen class, Heinz. This would give an error rate of 1 - (prior distribution of Heinz chosen) = 1 - 0.5096852 = 0.4903148, which is a (slight) improvement on what the error rate for the nearest-neighbor classifier gave.

4. (a) Source prob4.r in R to get a view of the pairs plot for the iris data. I would choose Petal.length as my split variable, and Petal.Length = 2.5 as my split point. When Petal.Length $\leq$ 2.5, all data points are of class Setosa, while all others are only of the other two classes (Versicolor and Virginica).

   (b) For the mixed region after the first part, I would choose Petal.Width as my next split variable and Petal.Width = 1.65 as my split point. In this region, when Petal.Width $\leq$ 1.65, I would classify points as Versicolor, and when Petal.Width ¿ 1.65, I would classify points as Virginica. This appears to give minimal error.

   (c) Here is my sketch/plot for this part. It's on the next page.


Virginica
Setosa
Versicolor
Petal.Width
Petal.Length
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
6
7