Briefly define the following terms. Formulas are OK, but explain what the symbols in them mean.

(a) Complete method (of clustering)

(b) Entropy

(c) Kernel trick in SVM

(d) Cross-validation

(e) Nearest neighbor classifier

(f) Dendrogram

(g) Gini index

What is regularization, and what purpose does it serve?

A decision tree is said to be fully grown if each leaf only contains trainin g points with the same label. Is a decision tree more likely to overfit if it is fully grown or not?

You are given a test that can determine whether a person is a terrorist or not from the emails they’ve written and received in the past year. The test is correct 95% of the time. Assume the prevalence of terrorists in this population is 1 in 10,000. If 10,000 people have been labeled “positive” by this test, how many of them are not terrorists?

Table 1 gives the decision table for OR. Give a perceptron that predicts this function.

Table 1: Decision table for OR. x1 and x2 are the inputs, and y the output.

|  |  |  |
| --- | --- | --- |
| x1 | x2 | y |
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 1 |

1. Figure 1 shows a plot of Ozone against Temperature, from a database of daily measurements in New York over 154 summer days. The following fits are summarized in the code extract:

> data(airquality);attach(airquality)

> library(gam)

> out1 = gam(Ozone ~ Temp)

> out2 = gam(Ozone ~ as.factor(Temp))

> out3 = gam(Ozone ~ s(Temp)) # the degrees of freedom

# for s(.) will be selected by GCV

* 1. (a)  The first model is the smoothest possible model; the second is the roughest possible model, and the third is somewhere in between.1 Explain why this is the case.
  2. (b)  A summary of the output is presented below. Which model has the best fit judging by the residual deviance? Which model has the best fit judging by the AIC? Why might the choice of the best model differ depending on which measure of fit is used? Which model seems to be the most useful judging from Figure 2?







’