Sample Exam Questions
Introduction to Statistical Learning
MATH96067(M3S20)/MATH97287(M4S20)

6th April 2020

1. The iris data set gives the measurements in centimetres of the four variables sepal length, sepal width, petal length and petal width for three species of iris plant. The data contains 50 sets of those four measurements for each species, giving 150 sets in total. The data is organised in a data frame and the first three observations are:

```
> iris[1:3,]
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
...
```

The three species are `setosa`, `versicolor` and `virginica`.

(a) The `tree` function in R was used to produce a classification tree with the iris data and is shown in Figure 1.

i. How many terminal nodes does the tree shown in Figure 1 have? How many variables are used in the tree? **[2 Marks]**

ii. Suppose somebody buys an iris plant and does not know what species it is from. They measure the plant and it has sepal length of 6.9cm, sepal width of 2.9cm, petal length of 5.2cm and petal width of 1.6cm. Using the classification tree shown in Figure 1 predict what species the plant is from. **[2 Marks]**

(b) Suppose we have a regression problem with a set of n univariate observations Y_i to be modelled in terms of a set of p-dimensional explanatory variables, X_i. We decide to use a regression tree with binary splits on the variables. Let the variable space be partitioned into M regions $R_1, \ldots, R_M$ and the response is modelled in each region by a constant c_m so that our model $f(x)$ is given by

$$f(x) = \sum_{m=1}^{M} c_m \mathbb{I}(x \in R_m). \tag{1}$$

Suppose we wish to minimise the quantity

$$\text{SSQ} = \sum_{i=1}^{n} \{Y_i - f(X_i)\}^2. \tag{2}$$

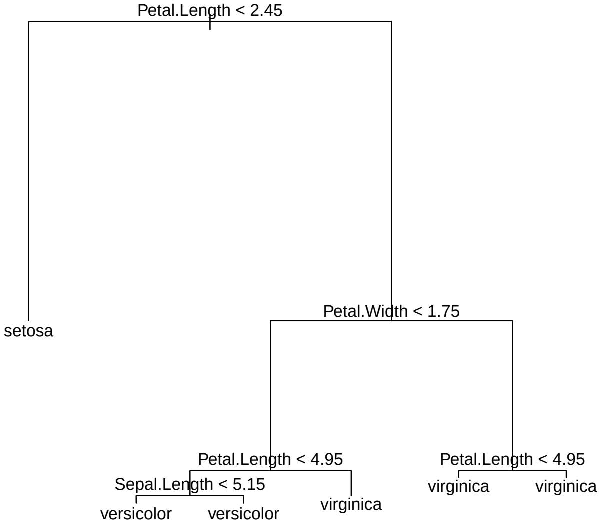

Figure 1: Classification Tree resulting from Iris data.

Explain why c_m is typically estimated by

$$\hat{c}_m = n_m^{-1} \sum_{i:X_i \in R_m} Y_i. \tag{3}$$

[2 Marks]

(c) To construct the tree, we have to decide which variables to split on and whereabouts to split on any given variable. Suppose we have decided to split on variable j, where j is some value in $1, \ldots, p$, split point s and define the two regions $R_1(j, s) = \{x | X_j < s\}$ and $R_2(j, s) = \{x | X_j > s\}$. Write down the appropriate residual sum of squares quantity that encapsulates the error for splitting variable j at split point s and show how it can be minimised efficiently. **[5 Marks]**

(d) Let T_0 be the large tree obtained by continually growing a tree, only stopping when the leaves all contain five observations or fewer. Explain why a large tree, T_0, is not usually useful for predicting new observations.

Define a subtree $T \subset T_0$ to be any tree that can be obtained by pruning T_0, that is collapsing any number of its internal nodes. Terminal nodes are indexed by m representing each region R_m. Let $|T|$ be the number of terminal nodes of T.

Let $\hat{c}_m$ be defined as in (3) and define

$$Q_m(T) = n_m^{-1} \sum_{i:X_i \in R_m} (Y_i - \hat{c}_m)^2. \tag{4}$$

The cost-complexity criterion is given by

$$C_\alpha(T) = \sum_{m=1}^{|T|} n_m Q_m(T) + \alpha|T|. \tag{5}$$

Explain how $C_\alpha(T)$ can be used to find a good tree. **[3 Marks]**

(e) Given two reasons why classification trees are good statistical methods and two reasons why not. **[2 Marks]**

(f) Describe by using sketches or otherwise, two data configurations for which classification and regression tree might not efficiently model the data, but other methods could. **[4 Marks]**

2. Let Y_i and X_i, $i = 1, \ldots, n$, be a set of univariate response and explanatory variables respectively, where X_i are independent and identically distributed with density function f.

(a) Let $f(x, y)$ be the joint density function of (X_i, Y_i). Write down the kernel density estimators of $f(x)$ and $f(x, y)$ with kernel K and bandwidth h. **[2 Marks]**

(b) Using the estimators, derive the Nadaraya-Watson estimator for $\mathbb{E}(Y|X = x)$ given by

$$\hat{\mathbb{E}}(Y|X = x) = \frac{\sum_{i=1}^n Y_i K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)}. \tag{6}$$

[3 Marks]

(c) Figure 2 shows the results of applying the Nadaraya-Watson, the local linear and local quadratic regression methods to a set of simulated data using a reasonably chosen bandwidth. The true unknown conditional mean, $\mathbb{E}(Y|X = x)$, is shown as dashed black line and the data points $\{(X_i, Y_i)\}_{i=1}^n$ as grey circles. Compare and contrast the three estimation methods. **[3 Marks]**

(d) An additive signal plus noise model can be written as $y_i = f_i + \epsilon_i$, where y_i are the observations, f_i the true unknown signal and ϵ_i the noise, for $i = 1, \ldots, n$, where $n = 2^J$, for some integer $J > 1$. Suppose that $\epsilon_i \sim N(0, \sigma^2)$ for some variance σ^2 independently, again for $i = 1, \ldots, n$. Let W be a discrete wavelet transform matrix.

i. Show that $w = d + e$, where $w = Wy, d = Wf$ and $e = W\epsilon$.**[1 Mark]**

ii. Show that $e_i \sim N(0, \sigma^2)$ and that $\{e_i\}$ forms a set of mutually independent observations.**[4 Marks]**

(e) Let $g(d)$ be a mixture of two normal densities such that

$$g(d) = (1 - p)\, \phi_{0,\tau^2}(d) + p\, \phi_{0,\gamma^2}(d), \tag{7}$$

where $\gamma \gg \tau$, $p \in (0, 1)$ and $\phi_{\mu,\nu^2}(d)$ is the probability distribution function of a normal random variable with mean μ and variance of ν^2. Let $g(d)$ be a prior probability density for the wavelet coefficients d of a function. Assume the wavelet coefficient priors are independent.

i. Explain the roles of hyperparameters p, τ and γ in the prior distribution and how they pertain to wavelet coefficients of a function. **[3 Marks]**

ii. Compute the posterior distribution and posterior mean of $d|w$. **[4 Marks]**

assume w is a single data

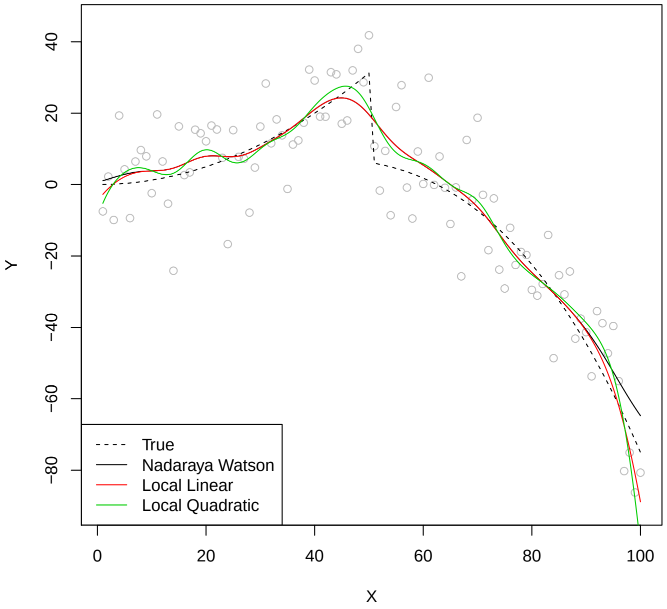

Figure 2: Comparison of local regression methods on simulated data. Grey circles are the data points (X_i, Y_i).

.