Assignment 01: Solutions

Problem 1: Random variables review

A random variable X is a real number with a value that depends on a random event. For example, X may be the measurement from a sensor, or the outcome of a die roll, or a property of an item chosen at random from a population of items. The probability that a random variable X is less than or equal to a real number x is denoted $\Pr(X \leqslant x)$. The function that maps x to $\Pr(X \leqslant x)$ is known as the *cumulative distribution function* (cdf) of X (this function may be denoted f_X). The cdf has many properties, for example $\lim_{x\to\infty} \Pr(X \leqslant x) = 1$. In some situations, and under some assumptions, random variables often have a cdf that is of a particularly well studied form. These forms are called laws, and we may say for example 'X is distributed according to the law $\text{Exp}(\lambda)$' or '$X \sim \text{Exp}(\lambda)$' to specify that a random variable is exponentially distributed with rate $\lambda > 0$.

Given the cdf of a random variable, we may ask how the value of the random variable concentrates around a particular real number x. The expression $\Pr(X > x \text{ and } X \leqslant x + \mathrm{d}x)$ gives the probability that X lies in the interval $(x, x + \mathrm{d}x]$. By the laws of probability, this expression is equal to $f_X(x + \mathrm{d}x) - f_X(x)$. By taking the limit as $\mathrm{d}x$ goes to zero, and scaling by $1/\mathrm{d}x$, we find the concentration of probability around x: $\lim_{\mathrm{d}x\to 0} (f_X(x + \mathrm{d}x) - f_X(x))/\mathrm{d}x = \dfrac{\mathrm{d}}{\mathrm{d}x} f_X(x)$. This derivative (if it exists) is known as the *probability density function* (pdf) of the random variable X.

a) Suppose X is a random variable with pdf proportional to $e^{-\lambda x}$ if x is positive and 0 otherwise. What is the pdf of X? (i.e., what is the constant of proportionality?) What is the cdf of X?

(4 points)

Let p be the pdf of X. We are given that $p = ae^{-\lambda x}$ for $x > 0$ and 0

c) Suppose $x \in \mathbb{R}$. For the random variable given in part a), what is $\Pr(X = x)$?

(4 points)

Let $\varepsilon > 0$ and $A_\varepsilon = (-\infty, x-\varepsilon] \cup (x+\varepsilon, \infty])$. By definition of integration and the cdf, $\Pr(X \in A_\varepsilon) = f_X(x-\varepsilon) + (1 - f_X(x+\varepsilon)) = 1 - e^{-\lambda(x-\varepsilon)} + e^{-\lambda(x+\varepsilon)}$. Thus $\lim_{\varepsilon \to 0} \Pr(X \in A_\varepsilon) = 1$. Suppose $\Pr(X = x) = \delta > 0$. Due to the limit, there exists a ε such that $1 - \Pr(A_\varepsilon) < \delta \Rightarrow \Pr(A_\varepsilon) > 1 - \delta$. The sets $\{x\}$ and A_ε are disjoint, so for this ε, $\Pr(A_\varepsilon \cup \{x\}) = \Pr(A_\varepsilon) + \delta > 1 - \delta + \delta > 1$. Probabilities must always be less than or equal to 1, so δ cannot be greater than 0. Probabilities must always be greater than or equal to 0, so δ must be 0. Thus, $\Pr(X = x) = 0$.

d) Prove that the pdf of a random variable X is non-negative (provided that the pdf exists).

(4 points)

If $b > a$, then set $(-\infty, a]$ is contained in the set $(-\infty, b]$. Therefore, when $b > a$, $f_X(b) \geqslant f_X(a)$ (if $A \subseteq B$, then $\Pr(A) \leqslant \Pr(B)$). Thus, $(f_X(x + \mathrm{d}x) - f_X(x))/\mathrm{d}x$ is non-negative for all x and $\mathrm{d}x > 0$. Thus, $p(x) \geqslant 0$ (if it exists), as it is the limit of a non-negative function.

Problem 2: Review of R

Consider the following for loops in R. For each for loop, list the values (in order) that the variable `i` takes on in the body of the loop. Briefly (in no more than a few sentences) explain why.

a)

```
for(i in 1+2:3.4*5) { }
```

(2 points)

The variable `i` *takes on the values 11, 16. This is due to order of operations, and how* `:` *interacts with non-integers. From highest to lowest, order of operations here is* `*`, `+`, `+`. *First,* `2:3.4` *is interpreted as the vector 2, 3 (the upper limit is rounded down when it's non-integer). Next, through broadcasting operations we multiply that vector by 5, then add 1 yielding 11, 16.*

b)

```
for(i in dim(matrix(0, nr = 7, nc = 8))) { }
```

(2 points)

The variable `i` *takes on the values 7, 8. The range of the for loop variable is the function* `dim` *applied to a matrix. The matrix is a zero matrix with 7 rows and 8 columns. The function* `dim` *returns the dimensions of a matrix (starting with number of rows) as a vector, yielding 7, 8.*

c)

```
for(i in rnorm(3)) { }
```

(2 points)

The function `rnorm(3)` *returns three random draws from a Gaussian distribution, and* `i` *varies over those values. We cannot know exactly what those values are, as we do not know the random number seed.*

d)

```
for(i in iris[1:3,3]) { }
```

(2 points)

The variable `i` *varies over the first 3 rows of the third column of the* `iris` *dataset. This dataset is built into R, and those specific values are 1.4, 1.4, 1.3.*

e)

```
for(j in c(1, 2, 3, 4, 5)) {
```

(2 points)

The variable `i` *is not a for loop variable (note the for loop is over* `j`*). In addition, the body is not syntactically correct (missing a closed curly brace).* `i` *does not vary over any values here.*

f)

```
for(i in (function(x) x*x)(c(1, 2, 3))) { }
```

(2 points)

The variable `i` *takes on the values 1, 4, 9. The for loop expression defines a function that returns the square of its input, and then broadcasts that function to the vector 1, 2, 3, yielding the squares.*

g)

```
for(i in NULL) { }
```

(2 points)

The variable `i` *takes on no values.* `NULL` *is interpreted as the empty list.*

h)

```
for(i in strsplit(as.character(4*atan(1)),'')
   [[1]][1:10]) {  }
```

(2 points)

The variable `i` *takes on the string values '3', '.', '1', '4', '1', '5', '9', '2', '6', '5'. Briefly,* `4*atan(1)` *is a mathematical expression that evaluates (in R's 'calculator' to 3.14159265.... This real number is converted to a string, and then the string is chopped into a vector of strings, with one character per vector coordinate. The first 10 coordinates of the vector are extracted to form the for loop expression.*

Problem 3: Using *knitr*

There are several ways to interleave R code and the output of R code (including plots) into a pdf. The R package *knitr* is one such way, and you may find it useful for creating reports and doing subsequent assignments. The way *knitr* works is by using a style of coding and creating documents called 'markdown'. In RStudio, install *knitr* using the command `install.packages('knitr')`. Then, create a new R markdown document by selecting 'File — New File — R Markdown' from the menus. You'll be prompted to give a name to the new R markdown file and select if you want it to output to pdf or html. Choose pdf. The basic format of an R markdown file is as follows (and your document may be populated with an example text such as this):

````
---
title: "Untitled"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## R Markdown
````

````
This is an R Markdown document. Markdown is a simple
   formatting syntax for authoring HTML, PDF, and MS
   Word documents. For more details on using R
   Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be
   generated that includes both content as well as the
    output of any embedded R code chunks within the
   document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```
## Including Plots

You can also embed plots, for example:

```{r pressure, echo=FALSE}
plot(pressure)
```
````

Click the 'knit' on the toolbar of RStudio's editor to render the markdown as a pdf. The document should pop up in a preview window. The pdf will also be saved in the same directory that your new R markdown file is saved in. The element '`##`' specifies a section title, the element '`title: "Untitled"`' specifies the document title, the elements '` ```{r ...} ... ``` `' specify R code that is to be executed. The element `echo=FALSE` indicates that the R code should not be emitted in the file (and instead, only the results of the code should be emitted). Markdown allows you to specify bolding, hyperlinks, bullets and other text aspects through annotations such as `**Knit**` for a bold 'Knit' (the asterisks indicate the bolding). An overview of the options for formatting and running code in R markdown is available here: https://www.rstudio.com/wp-content/uploads/

2015/02/rmarkdown-cheatsheet.pdf.

a) The University of California at Irvine provides a repository of datasets that are popular for demonstrations of machine learning and statistical methodology. Choose one of their datasets from this site: https://archive.ics.uci.edu/ml/datasets.php, and download it. The downloads usually include two files: one ending in `.data` which can be loaded by R using the command `read.table` and one ending in `.names` containing a detailed description of the dataset.

   In an R markdown, provide a short summary of the dataset What is the dataset about? When was it collected? How many items are in the dataset? How many variables are provided? Broadly, what types of variables are there, and broadly, what are their units? (For example, if there are thousands of variables all with the same units indicating measurements at different times, you can just say what the measurement is, what the units are and what the times are: you don't have to list each individual variable.) This summary should be no more than half a page.

   Choose one of the variables and plot a histogram of that variable. Ensure that the x-axis is labelled correctly, with units. Make the histogram so that its y-axis is 'proportion' and not 'count' (i.e., the sum of the areas of the histogram rectangles should equal 1). Superimpose on top of the histogram a plot of the pdf of a normal distribution (a.k.a. Gaussian distribution, or bell-curve) with mean and variance given by the empirical mean and variance of the variable. (For example, if you've chosen the 5th variable of the dataset and the dataset is loaded into R as the variable `df`, then the empirical mean is `mean(df[, 5])` and the empirical variance is `var(df[, 5])`.) Provide a single pdf including the rendered R markdown followed by a listing of the text of the R markdown file.

   (10 points)