Statistical Thinking: Week 7 Lab

Due 12noon Wednesday 7 October 2020

Introduction

The purpose of this Lab is to learn more about the application of Bayes' theorem. We'll practice using Bayes' theorem when the unknown variable takes on only a finite number of values.

The Lab consists of two applications, one in Part A and the other in Part B. Each part contains several questions for you to answer. Once you have completed these questions, you can attempt the Lab submission.

For this lab, you may prefer to write your answers out on paper. However, you can also use this as an opportunity to practice including mathematical symbols with your text in an **R Markdown** file, and combine this with doing the calculations using **R**.

To produce mathematical notation in your rendered **R Markdown** file, you will need to introduce **LaTeX** equations into the text. As noted on Homework Assignment 1, it is suggested that you refer to the **R Markdown Quick Reference** available under the **RStudio** *Help* menu (located at the very top of the **RStudio** environment). In particular, note the section on *LaTeX Equations*. Also, a useful resource for the LaTeX math symbols can be found at https://www.caam.rice.edu/~heinken/latex/symbols.pdf.

Lab Submission

To obtain credit for this lab, you are required to complete a Lab 7 Submission Moodle quiz, which is due in Week 8 on Wednesday 7 October at 12noon. You are not required to submit the .Rmd or .pdf for this Lab.

Good luck and have fun!

Part A: Is that a typo?

Have you ever wondered how an internet search engine decides when to correct the spelling of a word?[1] The answer is . . . by using Bayes' theorem![2] Let's see how it works.

Q1. Do a search on the word 'radom' using Google. What information do you find?

Now imagine that you are the computer search engine. . .

When you type a word like 'radom' into a Google search, how does Google know whether you really intended to type 'radom' or whether instead you made a spelling error? It seems to consider that you may have intended to type a different word, such as 'random' or 'radon'.

In statistics and machine learning, a *language model* is a probability distribution defined over sequences of words. Simple language models are defined for individual words found in a nominated 'universe of words' - a

[1] Much of this question has been taken from *Bayesian Data Analysis*, Gelman, A, Carlin, JB, Stern, HS, Dunson, DB, Vehtari, A and Rubin, DB. 3rd edn, Chapman & Hall/CRC texts in statistical science, 2014.

[2] Truthfully, the rules for spell checking are much more complex than described here. But the basic principle is right. For more information, see the references mentioned in the Peter Norvig post. Peter Norvig is the current research director at Google. See his information on Wikipedia.

collection usually obtained from scans of historical texts encompassing at least a million words. The model then defines the probability[3] of each word (meaning the chance it will occur again in a new sentence) as a proportion of times that word appears in the nominated 'universe of words'. Some language models cover a wide variety of topics, while others may be more context specific. For example, the 'universe of words' file may relate specifically to science, so words that occur more often in scientific discussions would be associated with a higher probability than they might be in a more general language model. More complex language models construct probabilities for sequences of words, and again can be context specific.

In this Lab, we will work with a simple language model, and focus on only three words. While a simplification, it provides a nice introduction to the use of Bayes' theorem in a machine learning application.

The language model Here we construct the language model probabilities from (past) relative frequencies of the three words available from an available universe of words.

Let the variable W represent the *word that was intended to be typed* on a given occasion. The three possible values of W in this setting are given by

- W_1 = 'random'
- W_2 = 'radon'
- W_3 = 'radom'

Q2. How common do you think each of these three words in the English language? (If you are not familiar with the word 'radon', you should ask Google for a definition!)

Now consider the available relative frequencies, denoted by RF_i, for each W_i, as shown in Table 1.

W_i	RF_i = Relative frequency of W_i
random	7.6×10^{-5}
radon	6.05×10^{-6}
radom	3.12×10^{-8}

Table 1: Relative frequency (RF_i) for each intended word.

As we can see from Table 1, the word 'random' occurs much more frequently than does the word 'radon', which in turn occurs much more frequently than the word 'radom'. (Does this seem sensible to you?)

Q3. With the words 'random', 'radon' and 'radom' being the only words under consideration, construct marginal probabilities for each word as implied by the relative frequencies shown in Table 1.

An observation

Now suppose Google's search engine has been given the search request, 'radom', which is a word it recognizes. This is an *observation*, which we will denote as y, i.e. y = 'radom'.

Google's search engine is programmed to consider whether the typed word is the word that was intended or if a different word was intended and a typographical error occurred. Bayes' theorem is used to update the marginal probabilities implied by the information in Table 1, to incorporate the observation of y = 'radom'.

[3] Some would argue this is really an *estimate* of the relevant probability. Alternatively, it may be taken is as an assessment of *prior belief* about the next likely intended word to appear in the search engine.

An error model for $y =$ 'radom'

Now to actually use Bayes' theorem, we need to be able to combine the new information about the observation y with the marginal probabilities we have for each W_i. In this language context, an *error model* is used as it identifies the chance that y will occur (as a *mis-typed* word), when in fact the true intended word is W_i.

Google has a model of spelling and typing errors[4], which provides the conditional probabilities shown in Table 2. (Do you think these error model probabilities seem reasonable?)

$\Pr(y = \text{'radom'} \mid W_i)$	W_i
0.00193	random
0.000143	radon
0.975	radom

Table 2: From Google's model of spelling and typing errors, in the case when $y =$ 'radom'.

As the observation $y =$ 'radom' is fixed, we are viewing $\Pr(y = \text{'radom'} \mid W_i)$ as a *function* of W_i with the data fixed at the observed value $y =$ 'radom', then we should refer to $\Pr(y = \text{'radom'} \mid W_i),\ i = 1, 2$ and 3, as the *likelihood function.*

Q4. Should the values in Table 2 sum to one? Why or why not?

Updated probabilities

Q5. Use Bayes' theorem to complete the values for Table 3. State your answers to at least four (4) decimal places.

W_i	$\Pr(W_i \mid y = \text{'radom'})$
random	
radon	
radom	

Table 3: Conditional distribution of W_i given y.

Note that your updated probabilites should combine *both* of the following facts:

- That 'radom' occurs relatively infrequently, according to the given language model, compared to the word 'random', and
- Depending on the intended word, the chance of actually typing 'radom' can happen with different probabilites.

Q6. Should the values in Table 3 sum to one? Why or why not?

Q7. If you had used the relative frequencies RF_i (from Table 1) in place of the normalised probabilities $\Pr(W_i)$ in Bayes' theorem, would your resulting conditional probabilities have the same values? Why or why not?

[4] Such a model may have been developed from a study where people were followed up after writing emails, so that the intended meaning of words that might have seemed incorrect in those emails could be identified.

Part B: Auto insurance claims

Auto insurance claims come from policies that are classified into three groups, corresponding to different deductible amounts and the expected size of an eventual claim. The classifications are: Low (L), Medium (M) and High (H). Historically, 80% of all claims are classified L, 15% are classified M and the remaining are classified H.

In addition to the large expected differences in observed in claim sizes *between* each of these three groups, differences are also expected to be observed in claim amounts *within* each classification group. To reflect this, claims within each group are described as arising from a *shifted Pareto* probability distribution, with the random claim amounts having probability density function (pdf) given by

$$f_Y(y \mid \theta_j) = \frac{2\theta_j^2}{y^3}, \quad y > \theta_j, \quad \text{for } y \text{ in group } j, \text{ and } j = L,\ M \text{ and } H$$

and where where $Y = y$ represents the size of an individual claim, in \$1000 dollars. Note that if $y \leq \theta_j$, then $f_Y(y \mid \theta_j)$ is defined to equal zero.[5] The parameter for each of the classes L, M and H, are given by $\theta_L = 1$, $\theta_M = 3$ and $\theta_H = 7$.

Note that although there is a different lower bound on the size of a claim for each classification, there is no upper bound.

Q8. Given you have received a single claim for \$7500, what is the probability that the claim came from group L? (i.e. what is $\Pr(\theta = \theta_L \mid y = 7.5)$?)

Q9. Again given the single claim \$7500, what is the probability that the claim came from group M?

Q10. Again given the single claim \$7500, what is the probability that the claim came from group H?

Notice the denominator

As you may have noticed, all of the denominators for the three conditional probabilities are the same, with the denominator is the sum of all of the numerator values. So you could have just calculated each of the three numerators (one each for θ_L, θ_M and θ_H, given the same $y = 7.5$ value) and then "normalise" them so that they add up to one. (i.e. divide each numerator by the sum of all three numerators).

Q11. Given you have received a claim for \$4500, what are the probabilities corresponding to the claim belonging to each of the groups L, M and H?

Q12. If you know that a given claim size is greater than \$1000 but less than \$3000, to which group must it belong? Provide support for your answer.

Q13. Show the form of the denominator of Bayes' theorem in this situation, and provide its value for each of values below. (Hint: In this situation, Bayes' theorem becomes $f(\theta_j \mid y) = \frac{f_Y(y|\theta_j) \times p(\theta_j)}{f_Y(y)}$, so we are looking for the denominator.)

i. $y = 7.5$,

ii. $y = 4.5$, and

iii. $y = 2$.

[5] The Pareto pdf here corresponds to a $Pareto(\alpha = 2, \lambda = \theta_j)$ as defined in the Week 6 slides, but with a shift in the random variable with $Y = X + \lambda = X + \theta_j$, for each $j = 1, 2, 3$. For example, $E[Y \mid \theta_j] = 2 * \theta_j$.