Stats 545: Homework 1

Due before midnight on Aug 31.
All plots should have labelled axes and titles.

Important: Rcode, tables and figures should be part of a single .pdf or .html files from R Markdown and knitr. See the class reading lists for a short tutorial.Any derivations can also be in Markdown, in Latex or neatly written on paper which you can give to me.

1 Problem 1: A little text-mining. [50 pts]

1. Install the `tm` package for Text-mining. Lecture 2 tells you how to install it, and load it into your `R` session. The package manual is here:
   http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf,
   though more useful for us is the following: http://onepager.togaware.com/TextMiningO.pdf. We will refer to this manual as [DataScR] [5pts]

Create at least two sufficiently long text-files containing anything (legal). E.g. I created a text-version of one of my papers here: http://www.stat.purdue.edu/~varao/STAT545/vinayak_papers.txt. Possible sources you can use are wikipedia, Project Gutenberg, the New York Times, blogspot etc. If you use a pdf file, convert it to text using Acrobat, the results won't be perfect (see my file), but that's ok. Store the text-files in a directory.

2. Read them into `R` using the `Corpus` function from `tm`. Page 3 of [DataScR] tells you how. Save it as a variable (we will call it `corpus`). [5pts]

3. Look at what type of variable `corpus` is, using the `typeof()`. Explain the difference between single and double square bracket indexing ([ ] vs [[ ]]). [5pts]

A corpus is a named-list of documents (the number of documents equals the number of textfiles you created). You can index the ith document with `corpus[[i]]`, (if the result scrolls past your screen too quickly, use the `page` command).

4. Sections 9 to 16 describe a set of transformations you can apply to the corpus: these include removing punctuation, coverting to lower case removing common words etc. You can play with this, but at the end, decide which transformations are suitable for your text files, and why (read the remaining points first). (Note: to convert to lower-case, you might have to run
   $\texttt{corpus} < -\texttt{tm\_map}(\texttt{corpus}, \texttt{content\_transformer}(\texttt{tolower}))$ [5pts]

5. Having processed the corpus, we use `DocumentTermMatrix` to convert it to a document-term matrix. This is a matrix where element $[i, j]$ tells you how often word j occurs in document i (you can also index the matrix with strings, e.g. `inspect(tdm[c("price", "texas"), c("127", "144")])`. See if the output makes sense to you. Also look at the help example from `?DocumentTermMatrix`. [10pts]

6. Calculate the relative frequency of the words in each document as well as across all documents (don't use for-loops). Sort the words in order of decreasing frequency, and plot a histogram of the frequencies of the top 15 words (for individual documents as well as the combination). [10pts]

7. Use the `wordcloud` package to plot a wordcloud for each of your text files as well as the combination. Section 26-29 of [DataScR] shows you how to do this. Extra points for colours, and interpretable word-clouds (e.g. top words in statistics papers vs top words in physics papers). Here is the result I produced from all my papers: `http://www.stat.purdue.edu/~varao/MISC/papers_wc_col.png`. [10pts]

2 Problem 2: Checking the central limit theorem. [20 pts]

1. Write a function that adds n `Unif(-1,1)` random variables and scales the results by $\sqrt{n}$. The resulting random variable will have a mean 0, but a variance that is not equal to 1. Thus include an additional scaling term so that the variance equals 1. Write all of this as an `R` function who's first input is n. The second input to the function should be m, with the function returning a collection of m such variables. Thus if your function is called `my_CLT`, then `my_CLT(n,m)` will return m variables, each of which is a scaled sum of n random variables. Do not use for loops. [10pts]
2. For a large value of m (say 10000), plot the density for $n = 1, 2, 3, 5, 10, 15$ and 20. [10pts]

3 Problem 3: Permutation test [20pts]

The dataset `reaction_times.txt` containts 200 people organized into two groups: the first 100 who have taken a drug, and the last 100 who have taken a placebo. For each individual, we have a measurement of how long they took to complete a task (in seconds). We want to test if there is any significant difference in the average response times of the two groups.

1. Calculate the different in the average response times of the two groups. [2pts]
2. Randomly assign the users to the two groups (while still maintaining group-sizes of 100). What is the average difference in reaction times now? [3pts]
3. Repeat this 1000 times, plotting the histogram of differences. Comment on whether the observed difference is plausible under the null hypothesis that there is no group difference. [7pts]

4 Problem 4: The power method [20pts]

1. Without using any `R` package, generate a random symmetric 5×5 matrix (the elements can be distributed however you like, for example Gaussian, or Uniform$[0, 1]$). Use the `eigen()` function to print its eigenvalues and eigenvectors. [6pts]
2. Write an `R` function that takes a matrix as input and implements the power method to generate its first eigenvector and eigenvalue. Apply this function to the matrix of the previous question and print your output. [7pts]
3. Now apply your function to generate the second eigenvector and eigenvalue of your matrix. This might help you: `http://goo.gl/rZ6Hgw`. Explain the intuition behind why this works. [7pts]

5 Problem 5: *Briefly* answer the following questions [10pts]

1. What is a positive-definite matrix? [2pts]
2. What is the Hessian of a function? What is a convex function? Plot a convex function. [3pts]
3. Write down the probability density of the multivariate Gaussian distribution? [2pts]
4. What is the law of large numbers? What is the central limit theorem? [3pts]

Feedback

Brief comments on each of these points would be greatly appreciated.

(a) Is the material presented at an adequate pace during lecture (too slow/too fast)?

(b) What general material would you like to spend more time on?

(c) Are the homework questions generally representative of material covered in lecture?

(d) Any further comments/questions/feedback?