36-350: Data Mining

Lab 3
Date: September 13, 2002 **Due: end of lab**

---

1 Introduction

This lab teaches you the steps involved in finding discriminative variables and making some simple plots. You will compare the three methods taught in class: the chi-square measure, the raw odds-ratio measure, and the hedged odds-ratio measure.
There are 4 questions. For each one, submit your commands and a response from R demonstrating that they work. (Only hand in commands relevant to the question.) To submit a plot, click on the plot window and select

```
File -> Save as -> Postscript...
```

This saves the plot to a file which can be printed, incorporated into a Word document, or mailed to us as an attachment. All plots should be properly labeled from within R.

2 Starting R

Start R as in lab 1. On the class web page, go to "computer labs" and download the files for lab 3 into your work folder. Read the special functions into your running R application via the commands

```
source("lab1.r")
source("lab2.r")
source("lab3.r")
```

If this fails, check that the files were downloaded correctly.

3 The data

Load the document data from lab 1:

```
load("lab1_docs.rda")
```

Remove singletons and compute prototypes:

```
doc <- remove.singletons(doc)
xp <- as.matrix(prototypes(doc,doc.labels,sum))
```

Each prototype contains the word counts for the entire subgroup.

4 Computing relevance

All information about word relevance can be obtained from the prototypes. The function `score.features` will score the words in various ways. It returns a named vector of scores, higher means more relevant. For the chi-square measure:

```
s1 <- score.features(xp,type="chisq")
```

For the hedged odds ratio:

```
s2 <- score.features(xp,type="odds")
```

For the raw odds ratio:

```
s3 <- score.features(xp,type="odds",z=0)
```

The function `sort` will sort a vector, showing you the highest scores.

> **Question 1:** (a) What are the 5 most relevant words according to the chi-square measure? (b) What are the 5 most relevant words according to the hedged odds-ratio measure? (c) What are the 5 most relevant words according to the raw odds-ratio measure?

5 Making plots

To quickly see the difference between the measures, it helps to make plots. In general, the command

```
plot(x,y,xlab="x label",ylab="y label")
```

will make a scatterplot of the values in vector `x` versus vector `y`, with x axis label `xlab` and y axis label `ylab`.

> **Question 2:** Use the `plot` command above to make a properly-labeled scatterplot of the chi-square scores against the hedged odds-ratio scores. (See the beginning of the lab for instructions on how to submit plots.)

A more effective scatterplot would use the words as labels instead of just dots. There is no single command in R to do this. Instead, you first make an "empty" scatterplot, then draw labels on top of it. To make an empty scatterplot, call plot with the argument `type="n"`, like so:

```
plot(x,y,xlab="x label",ylab="y label",type="n")
```

Then you draw labels via

```
text(x,y,labels,cex=0.75)
```

Here `x`, `y`, and `labels` are vectors of the same length. `x` and `y` are numeric vectors, describing position, and `labels` is a vector of text strings. The value `0.75` specifies that the text should be 75% of its normal size (if the plot is still too cluttered, make this smaller).
You can obtain a vector of the word names in several ways. For example, the columns of `xp` are named, so you can say `colnames(xp)`. Alternatively, the elements of `s1` are named, so you can say `names(s1)`.

> **Question 3:** (a) Use `plot` and `text` to make a better scatterplot of the chi-square scores against the hedged odds-ratio scores, using the words themselves instead of dots. (b) Do the same to make a scatterplot of the hedged versus raw odds-ratio scores.

6 Examining word counts

The subtable of counts for a given word in a given class can be extracted from the prototypes as follows:

```
> subtable(xp,"politics","such")
             such not such
politics        7     1998
not politics    2     3617
```

You can use this to examine why certain words are judged relevant.

> **Question 4:** (a) Using the scatterplots, find a word which is considered highly relevant by the chi-square measure, but not by the hedged odds-ratio measure. Submit (and keep) its subtable of counts. (b) Find a word which is considered highly relevant by the hedged odds-ratio measure, but not by the chi-square measure. Submit (and keep) its subtable of counts. (c) Find a word which is considered more relevant by the raw odds-ratio measure than by the hedged odds-ratio measure. Submit (and keep) its subtable of counts.