CS 429 - Information Retrieval
Homework 5
Shiva Sankar Modala(A20517528)

1 Recitation Exercises

These exercises are to be found in: Introduction to Information Retrieval 1 st Edition (Online) by Christopher Manning, Prabhakar Raghavan, Hinrich Sch¨utze.

1.1 Chapter 11

Exercises: 11.1

Work through the derivation of Equation (11.20) from Equations (11.18) and (11.19)

Ans:

11.18: $c_t = log\ \frac{p_t(1-u_t)}{u_t(1-p_t)}$

11.19: $p_t = \frac{s}{S}$ and $u_t = \frac{(df_t - s)}{(N-S)}$

Using 2nd, 3rd equations and solving 1st:

$$c_t = log\ \frac{\frac{s}{S}(1-\frac{(df_t-s)}{(N-S)})}{\frac{(df_t-s)}{(N-S)}(1-\frac{s}{S})} = log\ \frac{s(N-S)(1-\frac{(df_t-s)}{(N-S)})}{S(df_t-s)(1-\frac{s}{S})} = log\ \frac{s((N-S)-(df_t-s))}{(df_t-s)(S-s)}$$

$$= log\ \frac{s/(S-s)}{(df_t-s)/((N-S)-(df_t-s))} = \text{equation 11.20}$$

11.20: $c_t = log\ \frac{s/(S-s)}{(df_t-s)/((N-S)-(df_t-s))}$

1.2 Chapter 12

Exercises: 12.3,12.4,12.6

Exercises: 12.3

What is the likelihood ratio of the document according to M1 and M2 in Example 12.2?

Ans: P(s|M1)/P(s|M2) = 0.00000000000048 / 0.000000000000000384 = 1250

Exercises: 12.4

No explicit STOP probability appeared in Example 12.2. Assuming that the STOP probability of each model is 0.1, does this change the likelihood ratio of a document according to the two models?

Ans: No since it's the same probability for both models the ratio remains the same.

Exercises: 12.6

Consider making a language model from the following training text:

the martian has landed on the latin pop sensation ricky martin

a. Under a MLE-estimated unigram probability model, what are P(the) and P(martian)?

Ans: P(the) = 2/11

P(martian) = 1/11

b. Under a MLE-estimated bigram model, what are P(sensation|pop) and P(pop|the)?

Ans: P(sensation|pop) = 1
P(pop|the) = 0

1.3 Chapter 13
Exercises: 13.1,13.2,13.4,13.5,13.6,13.7,13.13
Exercise 13.1
Why is |C||V| < |D|Lave in Table 13.2 expected to hold for most text collections?
Ans: For most text collections, the number of vocab terms multiplied by the number of classes is less than the number of documents multiplied by the average length of a document. Heap's law uses that second value, the number of tokens in its calculations. If there's a reasonable number for tokens, parameter values in Heap's, and classes then it would make sense for this rule |C||V| < |D|Lave to work.

Exercise 13.2
Which of the documents in Table 13.5 have identical and different bag of words representations for (i) the Bernoulli model (ii) the multinomial model? If there are differences, describe them.
Ans: (i) For Bernoulli model, bag of words is same for all three
(ii) For multinomial model, first two are identical but third is different (count of london is one less)

Exercise 13.4
Table 13.3 gives Bernoulli and multinomial estimates for the word the. Explain the difference.
Ans: Multinomial:
$P(X = the|c) \approx 0.05$
X = t iff t occurs at given pos. So the probability is calculated by the percentage of "the" there is compared to other tokens, which is 5%

Bernoulli:
$P(Uthe = 1|c) \approx 1.0$
Ut = 1 iff t occurs in doc. So this means that the docs contains it, since the value is 1 (this makes sense, it's the most common word so most every doc will)

Exercise 13.5
Consider the following frequencies for the class coffee for four terms in the first 100,000 documents of Reuters-RCV1:

term	N00	N01	N10	N11
brazil	98,012	102	1835	51
council	96,322	133	3525	20
producers	98,524	119	1118	34
roasted	99,824	143	23	10

Select two of these four terms based on (i) $\chi 2$, (ii) mutual information, (iii) frequency.
Ans: (i) Using equation 13.19:

$$X^2(\mathbb{D}, t, c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) \times (N_{10} + N_{00}) \times (N_{01} + N_{00})}$$

After calculating using the calculator we get,
Brazil: ~ 819
Council: ~ 41
Producers ~ 597
Roasted: ~ 1965
Two highest X^2 values -> roasted, brazil

Using equation 13.17:

$$\begin{aligned} I(U;C) \quad = \quad & \frac{N_{11}}{N} \log_2 \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} \log_2 \frac{NN_{01}}{N_{0.}N_{.1}} \\ & + \frac{N_{10}}{N} \log_2 \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} \log_2 \frac{NN_{00}}{N_{0.}N_{.0}} \end{aligned}$$

Took the help of python to solve the equations:
Brazil: 0.00155
Council: 0.00017
Producers: 0.00104
Roasted:0.00064
Two highest: brazil, producers

(iii) Frequency (the number of documents in the class c that contain the term t):
Brazil: 51
Council: 20
Producers: 34
Roasted: 10
Two highest: brazil, producers

Exercise 13.6
Assume a situation where every document in the test collection has been assigned exactly one class, and that a classifier also assigns exactly one class to each document. This setup is called one-of classification (Section 14.5, page 306). Show that in one-of classification (i) the total number of false positive decisions equals the total number of false negative decisions and (ii) microaveraged F1 and accuracy are identical.
Ans: (i) If every document is assigned one class and the classifier tries to guess which, we can say that the total number of false positives equals the false negatives. If the classifier makes the wrong guess, it is both a false positive (it has been falsely assigned to this class) and a false negative (it has been falsely not assigned to the right class).

(ii) In one-of classification, microaveraged F1 is the same as accuracy. We're looking at the amount of correctly classified out of the total amount, so this is just the same as accuracy. (Count all tp, fn, and fp and then plug into f1 equation tp / (tp + ½*(fp+fn)). You can also just microaverage the precision and recall before plugging in, which will result in the same thing.

Exercise 13.7

The class priors in Figure 13.2 are computed as the fraction of documents in the class as opposed to the fraction of tokens in the class. Why?

Ans: Because we are classifying the documents into the classes, we need to learn about the prior probabilities of this rather than something else like the fraction of tokens in the class.

Exercise 13.13

Features can also be selected according to information gain (IG), which is defined as: IG(D, t, c) = H(pD) − Σ x∈{Dt+,Dt− } |x|/|D| H(px) where H is entropy, D is the training set, and Dt + , and Dt − are the subset of D with term t, and the subset of D without term t, respectively. pA is the class distribution in (sub)collection A, e.g., pA(c) = 0.25, pA(c) = 0.75 if a quarter of the documents in A are in class c. Show that mutual information and information gain are equivalent

Ans: Mutual Information: measures how much information the presence / absence of a term contributes to making the correction decision

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)}, \tag{13.16}$$

$$I(U;C) = \frac{N_{11}}{N} \log_2 \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} \log_2 \frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N} \log_2 \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} \log_2 \frac{NN_{00}}{N_{0.}N_{.0}} \tag{13.17}$$

Information Gain: IG(D, t, c) = H(pD) − Σ x∈{Dt+,Dt− } |x|/|D| H(x)

H is entropy (https://en.wikipedia.org/wiki/Entropy_(information_theory)), which is:

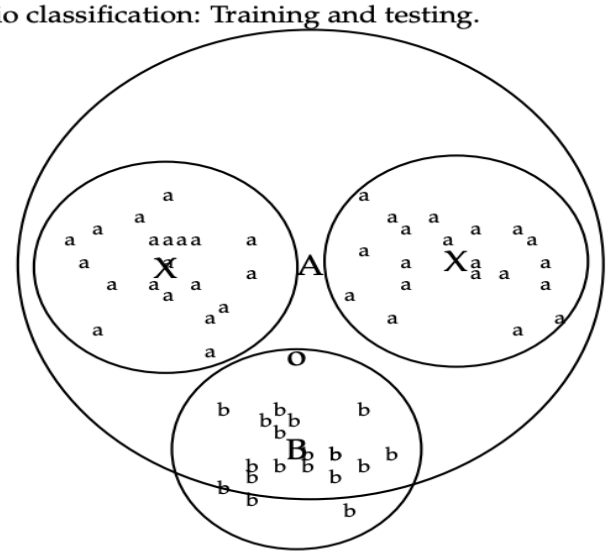
$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log \mathrm{P}(x_i)$$

Where the xi's are possible outcomes. So with $H(pD) - \Sigma\, x \in \{Dt+, Dt-\}\ |x|/|D|\ H(x)$, we will multiply $x/|D|$ by this H(x) equation for all subsets. This ends up being the same as 13.16 as shown above. Putting the first part of IG with the second ($H(pD)$ with $\Sigma\, x \in \{Dt+, Dt-\}\ |x|/|D|\ H(x)$), you can simplify to get 13.16. It's hard to show not written out on paper with all of the steps, but essentially the $|x|/|D|$ is just going to be a proportion / probability of subset from the whole. So when you look at the second half of the IG equation, it will just be 13.16 without the second part of the denominator. To get the second half of that denominator, factor in the first part of IG which is the H.

1.4 Chapter 14
Exercises: 14.1,14.2,14.3,14.6
Exercises: 14.1
For small areas, distances on the surface of the hypersphere are approximated well by distances on its projection (Figure 14.2) because $\alpha \approx \sin\alpha$ for small angles. For what size angle is the distortion $\alpha / \sin(\alpha)$ (i) 1.01, (ii) 1.05 and (iii) 1.1?
Ans: (i) a = sin(a) for small angles
I put the equations into wolframalpha in order to get the angle size:
$x/\sin(x) = 1.01 \rightarrow x \sim= 0.244096696$
(ii) $x/\sin(x) = 1.05 \rightarrow x \sim= 0.53841167\ldots$
(iii) $x/\sin(x) = 1.1 \rightarrow x \sim= 0.74898664\ldots$

Exercises: 14.2
Show that Rocchio classification can assign a label to a document that is different from its training set label.
Ans:

► **Figure 14.4** Rocchio classification: Training and testing.



► **Figure 14.5** The multimodal class "a" consists of two different clusters (small upper circles centered on X's). Rocchio classification will misclassify "o" as "a" because it is closer to the centroid A of the "a" class than to the centroid B of the "b" class.

This example from the textbook shows a Rocchio classification in which a label is assigned incorrectly to that of its training label set. Rocchio uses centroids in order to determine which class a document (or whatever the vector here corresponds to) belongs to. The “o” will be classified as an “a” since it is closest to the centroid “A” that belongs to a class. It is actually in the “b” circle though and it is closest to the centroid of B if you consider each “a” circle to have their own centroids (denoted by the “X”s in the picture).

Exercises: 14.3
Explain why kNN handles multimodal classes better than Rocchio.
Ans: kNN uses k nearest neighbor classification. It can handle non-spherical and other complex classes better than Rocchio. If you use Rocchio, which divides the vector space into regions centered on centroids for each class, then it could be inaccurate if classes are not approximately spheres with similar radii. The continguity hypothesis is that documents in the same class form a contiguous region and regions of different classes do not overlap. Since this may not hold in multimodal classes, we should use knn instead.

Exercises: 14.6
In Figure 14.14, which of the three vectors~a,~b, and~c is (i) most similar to ~x
according to dot product similarity, (ii) most similar to ~x according to cosine similarity,
(iii) closest to ~x according to Euclidean distance?
Ans: (i) dot product similarity
a = (0.5 1.5) b = (4 4) c = (8 6) x = (2 2)
a and x -> (0.5*2) + (1.5*2) = 4
b and x -> (4*2) + (4*2) = 16
c and x -> (8*2) + (6*2) = 28
Highest similarity is vector c

(ii) cosine similarity
a and x -> 4 / ((sqrt((0.15)^2+(1.5)^2)) * (sqrt((2)^2+(2)^2))) = 0.94
b and x -> 16 / ((sqrt((4)^2+(4)^2)) * (sqrt((2)^2+(2)^2))) = 1
c and x -> 28 / ((sqrt((8)^2+(6)^2)) * (sqrt((2)^2+(2)^2))) = 0.99
Highest similarity is vector b

(iii) euclidean distance:
a and x -> sqrt((0.15 - 2)^2 + (1.5 - 2)^2) = 1.92
b and x -> sqrt((4 - 2)^2 + (4 - 2)^2) = 2.83
c and x -> sqrt((8 - 2)^2 + (5 - 2)^2) = 7.21
Lowest distance is vector a

1.5 Chapter 15
Exercises: N/A