Internet Analytics (COM-308)

Homework Set 6

Exercise 1

One of the most important applications of Bayesian text classification is spam filtering. In this exercise we use naive Bayesian classification to filter spam emails based on their subjects.

Suppose you have the following training corpus of ham (G) and spam (B) subjects:

Ham (messages that your filter should forward to the inbox):

1. "World money crisis"
2. "Expect extra economic crisis"
3. "Online world war"

Spam (messages that your filter should drop):

1. "Extra income opportunity"
2. "Make money online"
3. "Earn money"
4. "Expect money income"

(a) Compute the prior $P(B), P(G)$, and the model $P(W|G), P(W|B)$, where W is a word.

(b) Compute the posterior probabilities $P(G|M), P(B|M)$ for the following messages M (with the naive Bayes classifier) and classify them as ham or spam:

1. M_1="Online money crisis"
2. M_2="Expect online income"
3. M_3="Earn extra cash"

Answer the same questions as above, but using Laplace smoothing (with $k = 1$).

Exercise 2

In class we defined the cosine similarity:

$$CosSim(x, y) = \frac{\langle x, y\rangle}{||x||\ ||y||},$$

which has an interpretation as the cosine of the angle between the two vectors. We also saw the Pearson correlation $Corr(x, y)$. Let $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$, and $\bar{y}$ analogously. Then

$$Corr(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{1}$$

Can you see a relationship between the two similarity metrics?

Exercise 3

(a) Find the partitioning which maximizes the modularity of the graph G in Figure 1. What is the maximum modularity value Q?

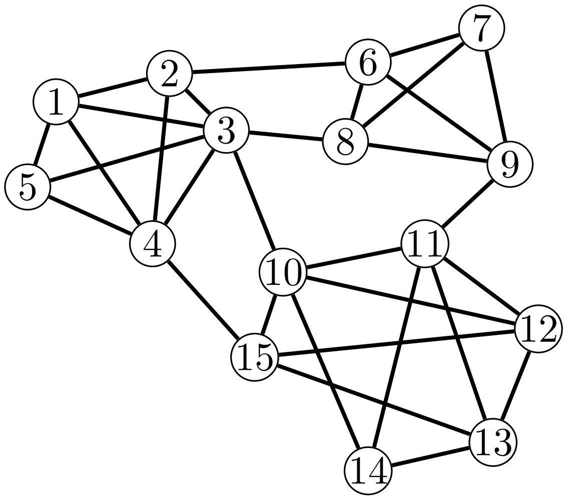

Figure 1: Graph G

(b) We want to increase the value of modularity by removing one edge from graph G. Guess the edge whose deletion results in the largest increase, and compute the new Q.