# 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.*

$P(B) = 1 - P(G) = 11/21$.

*(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"*

The first thing to do is to compute the conditional probabilities for each word $W$:

| $W$ | $10 \times P(W\|G)$ | $11 \times P(W\|B)$ |
|---|---|---|
| online | 1 | 1 |
| money | 1 | 3 |
| crisis | 2 | 0 |
| expect | 1 | 1 |
| income | 0 | 2 |
| earn | 0 | 1 |
| extra | 1 | 1 |
| cash | 0 | 0 |

To compute the posterior of a message $M = (W_1, W_2, W_3)$, we apply the conditional i.i.d. assumption and Bayes' rule to compute

$$P(G|M) = \frac{P(W_1|G)P(W_2|G)P(W_3|G)P(G)}{P(W_1|G)P(W_2|G)P(W_3|G)P(G) + P(W_1|B)P(W_2|B)P(W_3|B)P(B)}. \quad (1)$$

$P(B|M)$ is defined similarly.

For $G$ we find:

1. $P(G|\text{Online money crisis}) = \frac{1/10 \times 1/10 \times 2/10 \times 10/21}{1/10 \times 1/10 \times 2/10 \times 10/21 + 0} = 1.$

2. $P(G|\text{Expect online income})$ is undefined, because 'income" does not appear in a ham sentence in the training set.

3. $P(G|\text{Earn extra cash})$ is undefined, because because 'earn" and "cash" do not appear in a ham sentence in the training set.

For $B$ we find:

1. $P(B|\text{Online money crisis})$ is undefined, because 'crisis" does not appear in a spam sentence in the training set.

2. $P(B|\text{Expect online income}) = \frac{1/11 \times 1/11 \times 2/11 \times 11/21}{0 + 1/11 \times 1/11 \times 2/11 \times 11/2} = 1$.

3. $P(B|\text{Earn extra cash})$ is undefined because "cash" does not appear in a spam sentence in the training set.

1. $M_1$ is a ham message as $P(G|M_1) = 1$.

2. $M_2$ is a spam message as $P(B|M_2) = 1$.

3. We can not classify $M_3$ as both $P(G|M_3)$ and $P(B|M_3)$ are undefined.

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

The smoothed class prior is $P(B) = 1 - P(G) = 19/37$, and the smoothed conditional word probabilities are:

| $W$ | $18 \times P(W\|G)$ | $19 \times P(W\|B)$ |
|---|---|---|
| online | 2 | 2 |
| money | 2 | 4 |
| crisis | 3 | 1 |
| expect | 2 | 2 |
| income | 1 | 3 |
| earn | 1 | 2 |
| extra | 2 | 2 |
| cash | 1 | 1 |

For $G$ we find:

1. $P(G|\text{Online money crisis}) \approx 0.63$.

2. $P(G|\text{Expect online income}) \approx 0.27$.

3. $P(G|\text{Earn extra cash}) \approx 0.36$.

For $B$ we find:

1. $P(B|\text{Online money crisis}) \approx 0.37$.

2. $P(B|\text{Expect online income}) \approx 0.73$.

3. $P(B|\text{Earn extra cash}) \approx 0.64$.

Thanks to smoothing all probabilities involved are now $> 0$, giving us a well-defined posterior. The "trend to the middle" effect is quite pronounced because the training set is small.

For classification we have

1. $M_1$ is a ham message as $P(G|M_1) > P(B|M_1)$.

2. $M_2$ is a spam message.

3. $M_3$ is a spam message.

## 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*

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

*Can you see a relationship between the two similarity metrics?*

$$\begin{aligned} Corr(x, y) &= \frac{\langle x - \bar{x},\ y - \bar{y} \rangle}{||x - \bar{x}||\ ||y - \bar{y}||} \\ &= CosSim(x - \bar{x}, y - \bar{y}) \end{aligned}$$

## Exercise 3

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

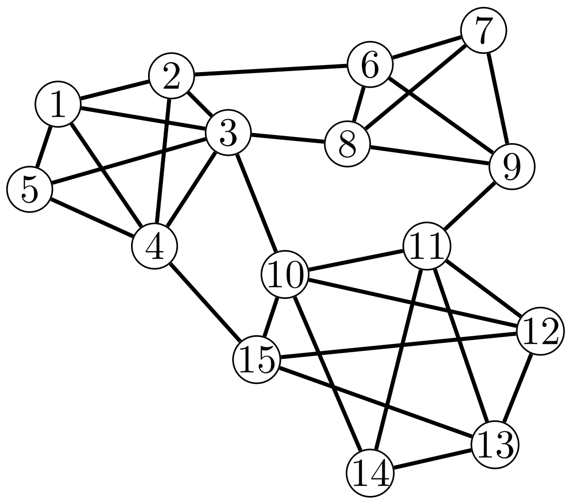


Figure 1: Graph $G$

Three partitions $c_1, c_2$ and $c_3$ which maximize the modularity are shown in Figure 2. The corresponding modularity is

$$Q = 0.4917.$$

*(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.*

Removing an inter–partition edge between partitions $c_1$ and $c_3$, for example edge $(3, 10)$, results in the highest increase of modularity. The modularity after this edge deletion is

$$Q = 0.5217.$$

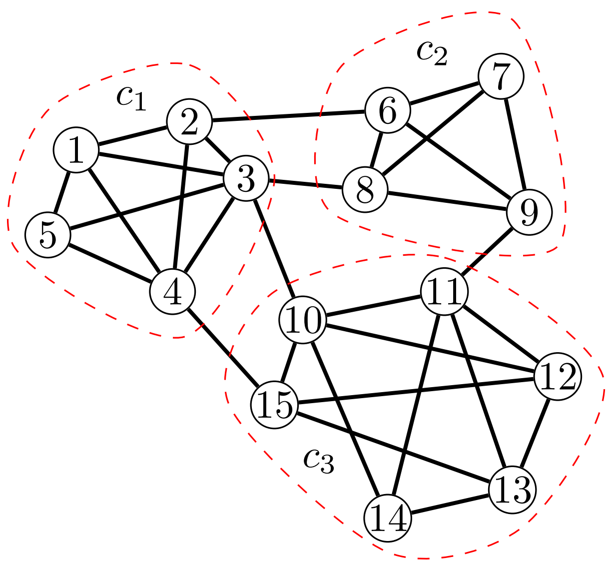


Figure 2: Graph $G$