Matt Dickenson `mcd31`
STA561/CS571 — Fall 2013 **Homework 2** Due: 16 September, 2013

Homework Notes: I did not work with anyone else on this homework or refer to resources other than the course notes, textbook, and course Piazza page.

Problem 1

A No, $X_1 \not\perp X_2 | X_3$:

$$\begin{aligned} p(x_1, x_2, x_3) &= p(x_1)p(x_2)p(x_3|x_1, x_2) \\ p(x_1, x_2|x_3) &= \frac{p(x_1)p(x_2)p(x_3|x_1, x_2)}{p(x_3)} \end{aligned}$$

Also, by the Bayes Ball approach there is an active path from X_1 to X_2 when conditioning on X_3.

B Yes, $X_1 \perp X_2 | X_4$. By the Bayes Ball approach, X_1 and X_2 "bounce back" when they hit X_3 conditioning only on X_4.

C Yes, $X_1 \perp X_2$ when X_3 and X_4 are unobserved. This is because the balls stop when they hit the unobserved X_3, so there are no active paths between them.

D Yes, $X_4 \perp X_7 | X_1$. The X_4 ball can pass through X_2 but stops when it hits X_1. Similarly, X_7 can pass through X_3 but stops when it hits X_1.

E No, $X_4 \not\perp X_5 | X_1$. Both X_4 and X_5 stop when they hit X_2.

Problem 2

A The joint probability represented by the graph is

$$\begin{aligned} p(X_A, X_B, X_C, Y) &= p(Y)p(X_A|Y)p(X_B|Y)p(X_C|Y) \\ &= p(Y)\frac{p(Y|X_A)p(X_A)p(Y|X_B)p(X_B)p(Y|X_C)p(X_C)}{p(Y)} \\ &= p(Y|X_A)p(X_A)p(Y|X_B)p(X_B))p(Y|X_C)p(X_C) \end{aligned}$$

because the graph structure indicates that

B In the model given, all pairs of features (X_i, X_j) are conditionally independent given Y. This means that each feature influences π independently of the others. In practice this is not realistic–the number of words, time of day, and number of non-dictionary words are not independent of one another. For example, the number of non-dictionary words likely increases with the total number of words in the email, and time of day is likely associated with the number of words in the email.

C We can compute the probability that a new email is spam using Bayes Rule:

$$\begin{aligned} P(Y_j = 1|(X_A, X_B, X_C)_j) &= \frac{P((X_A, X_B, X_C)_j|Y_j = 1)P(Y_j = 1)}{P((X_A, X_B, X_C)_j)} \\ &\propto P((X_A, X_B, X_C)_j|Y_j = 1)P(Y_j = 1) \end{aligned}$$

We can simplify this using conditional independence (part B):

$$\begin{aligned} P(Y_j = 1|(X_A, X_B, X_C)_j) &\propto P(X_{Aj}|Y_j = 1)P(X_{Bj}|Y_j = 1)P(X_{Cj}|Y_j = 1)P(Y_j = 1) \\ &\propto P(X_{Aj}|Y_j = 1)P(X_{Bj}|Y_j = 1)P(X_{Cj}|Y_j = 1)\hat{\pi} \end{aligned}$$

where $\hat{\pi}$ is our prior for $P(Y_j = 1)$ given the training set.
In other words, we can compare the likelihood of the features $(X_A, X_B, X_C)_j$ under the hypotheses $Y_j = 1$ and $Y_j = 0$.

Problem 3

A To compute MLE(π) from the training set D, we can simply take the proportion of emails Y_D that were classified as spam ($Y_i = 1$):

$$\hat{\pi}_{MLE} = \frac{\sum_D Y}{N_D}$$

B With fixed $\sigma^2_{A,y}$, we can just partition the data based on whether or not it is spam and compute each class mean MLE estimate:

$$\begin{aligned} \hat{\mu}_{A,1} &= \frac{1}{N_1} \sum_{i:y_i=1} X_{A,i} \\ \hat{\mu}_{A,0} &= \frac{1}{N_0} \sum_{i:y_i=0} X_{A,i} \end{aligned}$$

where N_i is the number of cases in D for which $Y = i$.

C With fixed class means, we can derive the MLE of the variance for feature A in the training set:

$$\hat{\Sigma}_0 = \frac{1}{N_0} \sum_{i:y_i=0} (x_i - \hat{\mu}_{A,0})(x_i - \hat{\mu}_{A,0})^T$$