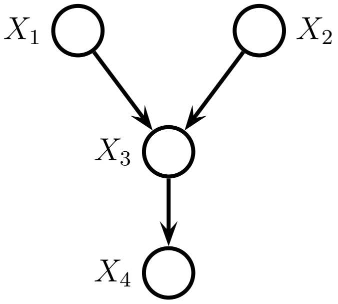
STA561/CS571 — Fall 2013 **Homework 2** Due: September 16, 2013

Instructions: Please put all answers in a single PDF with your name and NetID and upload to SAKAI before class on the due date (there is a LaTeX template on the course web site for you to use). Definitely consider working in a group; please include the names of the people in your group and write up your solutions separately. If you look at any references (even wikipedia), cite them. If you happen to track the number of hours you spent on the homework, it would be great if you could put that at the top of your homework to give us an indication of how difficult it was.

Problem 1

Conditional independence and Bayes Ball algorithm. Answer the questions below, and use either Bayes Ball algorithm or conditional probability to explain. If you use Bayes ball, please describe the path that the 'ball' takes and give an intuition using the canonical three-node graph structures about why (or not) the reachability argument holds, and thus conditional independence is not satisfied (or vice versa).
Considering the following graph (random variables that we are conditioning on are not shaded here because they change based on the problem):

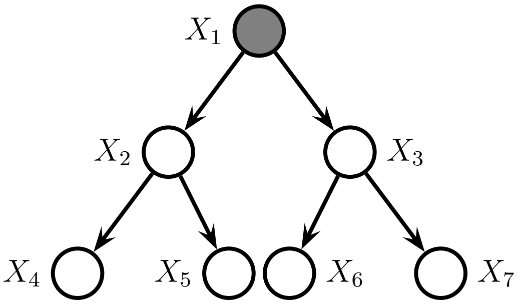


(a) Is $X_1 \perp X_2 | X_3$ (are X_1 and X_2 conditionally independent given X_3)?

(b) Is $X_1 \perp X_2 | X_4$?

(c) Is $X_1 \perp X_2$? (X_3 and X_4 are unobserved here)

Considering the following graph,

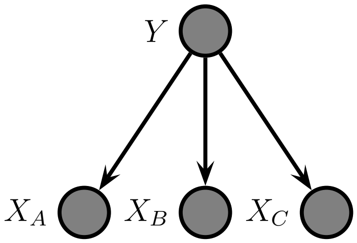


(d) Is $X_4 \perp X_7 | X_1$?

(d) Is $X_4 \perp X_5 | X_1$?

Problem 2

Naive Bayes classifier (NBC). Naive Bayes classifier is widely used to classify emails. It can be expressed in the graph below, where Y is the class label, $Y \in \{0, 1\}$, where class 1 means *spam email*, and class 0 means *non-spam email.* Random variables X represent the features of the emails; for instance, X_A can be the number of words in an email, X_B can be the time of day receiving the emails, X_C can be the number of words not found in dictionary. Let us assume that these three features are Gaussian distributed, although this assumption is not entirely accurate. Let us also assume that Y is Bernoulli with $p(Y = 1|\pi) = \pi$.



(a) Write down the joint probability of the random variables in the graph and factorize it according to the graph structure.

(b) For all pairs of features (X_i and X_j, $i \neq j$), is $X_i \perp X_j | Y$? Explain what this means with respect to the problem of classification and the implications for the features. Is this true of our real data in practice? Can you think of a situation when this assumption will be explicitly violated, and may impact our classification accuracy? (Two sentences)

(c) We have a training set $D_{training} = \big\{(Y, X_A, X_B, X_C)_1, ..., (Y, X_A, X_B, X_C)_n\big\}$ (with observed features and class labels, in other words, n emails classified as spam or not spam), and a test set $D_{test} = \big\{(X_A, X_B, X_C)_1, ..., (X_A, X_B, X_C)_m\big\}$ (with observed features, but unknown class labels).
With this classifier, we could predict the probability that a test email belongs to the spam class. Write down how to compute $P(Y_j = 1|(X_A, X_B, X_C)_j)$, $j = 1, ..., m$ given that we know the components of our factorized joint probability (*hint: ignore the training set, use Bayes' Rule).*

Problem 3

Maximum likelihood estimates for NBC.
Still using the graph in Problem 2 and the task of classifying emails, let's focus on the training data, $D_{training} = \big\{(Y, X_A, X_B, X_C)_1, ..., (Y, X_A, X_B, X_C)_n\big\}$, $Y \in \{0, 1\}$, i.e.,

the features and class labels are fully observed here. Suppose our features $X_{\{A,B,C\}}$ in each class of emails have Gaussian distributions, e.g., $X_A|Y = y \overset{iid}{\sim} \mathcal{N}(\mu_{A,y}, \sigma^2_{A,y})$. In other words, the Gaussian parameters for the features are class-specific, meaning there is one mean for feature A when the email is spam and another mean for feature A when the email is not spam.

(a) $Y \overset{iid}{\sim} Ber(\pi)$, $\pi \in [0, 1]$. How do we find the maximum likelihood estimate (MLE) of π using the training set D?

(b) If $\sigma^2_{A,y}$ is fixed, derive the MLE of $\mu_{A,1}$ and $\mu_{A,0}$ (in other words, the class mean of feature A for spam $\mu_{A,1}$ and not spam $\mu_{A,0}$) using training set D.

(c) If μ_{A_1} and μ_{A_0} are fixed, derive the MLE for $\sigma^2_{A,0}$ using training set D.