Rohan Tilva

1 Analytical (70 points)

1. (16 points) Consider the Bayesian Network. Are the sets **A** and **B** d-separated given set **C** for each of the following definitions of **A**, **B** and **C**? Justify each answer.

1. (a) Sets **A** and **B** are d-separated given set **C**. We know that x_8 blocks the path because it is head-to-tail and observed. We also know that x_{16} is head-to-head, but since it is observed, it doesn't block the path. However, the only path connecting x_1 to x_9 is through node x_{15}, which blocks the path because it is head-to-head and unobserved. Thus **A** and **B** are d-separated given set **C**.

2. (b) In this case we only care about x_{15}. Since x_{15} is observed and head-to-head, it doesn't block the path between x_{11} and x_{13}. Thus sets A and B are not d-separated.

3. (c) Since x_{15} is unobserved and head-to-head, it blocks the only path between A and B. Thus they are d-separated.

4. (d) x_{15} is observed and head-to-head, so it does not block the path from sets A and B. x_7 also does not block any paths, and x_{16} is head-to-head and observed, so it does not block the path either. Thus A and B are not d-separated.

Now consider the Markov Random Field, which has the same structure as the previous Bayesian network. Re-answer each of the above questions with justifications for your answers.

1. (a) All paths from A to B pass through x_8 and x_{16}, which are both in set C, so A and B are d-separated.

2. (b) All paths from A to B pass through x_{15}, and set C contains x_{15}, so A and B are d-separated.

3. (c) All paths from A to B pass through x_{10}, and since x_{10} is in set C, A and B are d-separated.

4. (d) All paths from A to B pass through x_{15}, and since x_{15} is in set C, A and B are d-separated.

2. (6 points) Let $X = (X_1, ..., X_{16})^T$ be a random vector with distribution given by the graphical model in Figure **??**. Consider variable X_2.

1. (a) The minimal subset is just X_5. This is because X_5 is head-to-tail, and so if it is observed it blocks all of the paths from X_2 to every other node. So if we put X_5 in the subset, X_2 is independent of the rest of the variables given X_5.

2. (b) The minimal subset is X_5. This is because every path from X_2 to every other node passes through X_5. So the subset is X_5, so X_2 and every other node are d-separated given X_5.

3. (16 points) First we have to redefine P(X) because we do not know the labels for U. For each of the n examples, we do not necessarily know the corresponding labels, so we have to sum over all possible labels. Thus we can redefine this without Y as follows:

$$P(X) = \prod_{i=1}^{n} \sum_{y=1}^{k} f(y) \prod_{j=1}^{m} r(X_{ij}|y)$$

where we have:

$$f(y) = \frac{\sum_{i=1}^{n} p(y|x_i, \theta_t)}{n}$$

$$r(x|y) = \frac{\sum_{i=1}^{n} p(y|x_i, \theta_t) x_i}{\sum_{i=1}^{n} \sum_{j=1}^{J} p(y|x_i, \theta_t) x_{ij}}$$

$f(y)$ represents the probability distribution for labels y, and $r(x|y)$ represents the probability of having a certain example given a label y. If we encounter a labeled example from L, then in the E step defined below, we will simply let $p(y|x_i, \theta_t) = 1$ when y is equal to the true label, and let it equal 0 otherwise.
First, we want to use L to get a baseline for our model parameters, so we estimate θ_0 using the labeled examples from the training set. We are now ready to begin the EM step. E: in this step, we compute the expected labels given the current model's parameters. We need to consider every single possible label for each example, since we do not know the labels of all examples. So for each example x_i, and each label y for a given example, we calculate the value of $p(y|x_i, \theta_t)$. M: now, we want to maximize our model's parameters. We use the two functions defined above to calculate our new model's parameters for $t + 1$.

4. (16 points) The probability density function of most Markov Random Fields cannot be factorized as the product of a few conditional probabilities. This question explores some MRFs which can be factorized in this way.

1. (a) A factorization of the joint is:
$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7) = P(X_7|X_4)P(X_6|X_3)P(X_5|X_3)P(X_4|X_2)P(X_3|X_2)P(X_2|X_1)P(X_1)$.

2. (b) No this factorization is not unique because it depends on the way we create the directed edges. Thus we could have written other factorizations that correspond to the model.

3. (c) An alternate factorization would be:
$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7) = P(X_5|X_3)P(X_6|X_3)P(X_3|X_2)P(X_1|X_2)P(X_2|X_4)P(X_4|X_7)P(X_7)$

4. (d) These examples can be factored in such a way since they do not have any cycles greater than the size of 3. With cycles greater than size 3, all of the conditional independencies implied by the original MRF cannot be captured by a single directed representation. An example of an MRF that could not be factored is one with nodes $\{x_1, x_2, x_3, x_4\}$ arranged in a diamond, such that there are edges (x_1, x_2), (x_1, x_3), (x_2, x_4), (x_3, x_4). Any directed representation of this MRF will not be able to capture BOTH of the conditional independences implied by the original MRF.

5. Overfitting in Clustering (16 points)

1. (a). As the number of clusters (k) increases, the overall distribution of examples will always be closer and closer to where the cluster centers are located, resulting in a lower total within-cluster sum of squares. When we have fewer clusters (ie. when k is smaller), there is a higher probabaility of having an example that is "far" away belong to a given cluster. This results in a higher γ_k. However, when we increase k, examples that were further from the cluster centers with smaller k will be, on average, closer to the new cluster centers. This results in a smaller within-cluster sum of squares, and thus a smaller γ_k. One intuitive example is the following: let's say we have $k = 1$ and some γ_k. If we add a new cluster j anywhere, either some examples will be assigned to the new cluster j or still be assigned to the old cluster. For an example to be assigned to cluster j, it must be closer to j than the old cluster. Thus, we can see that the overall distances between examples and clusters will either stay the same or decrease with increased k, meaning that γ_k is non-increasing in k.

2. (b). $\gamma_k \leq \gamma'_k$ since the max() function will either choose $\|x_j - \mu_j\|_2^2$ (in which case $\gamma_k = \gamma'_k$), or it will choose τ, in which case $\gamma_k \leq \gamma'_k$. So γ_k is always less than or equal to γ'_k.

3. (c). For K-means, examples that are very far away are considered when placing the cluster centers, whereas for K-medoids, the cluster centers are not affected by these outliers. Thus, we know that any examples that are farther away from the cluster centers in K-medoids will result in a much higher squared distance than in K-means (since K-means actually takes these farther away examples into account). This means that K-medoids will have a greater than or equal to optimal solution than the K-means optimal solution.

4. (d). To minimize the objective function, I would use 20. This is because a given example's cost (sum of squares cost) can be minimized all the way down to 0, whereas for 21, there is a threshold (which is the max of those two terms in the function). Thus we are bounded by τ, and thus cannot minimize the sum of squares cost for a given example past τ. 21 will also result in smaller number of clusters since τ will lead to

multiple points having the same cost (of cost τ). We already proved in (a) that more clusters minimizes the objective, meaning that 21 will have a greater than or equal cost than 20. Thus 20 is optimal.