10-701 Machine Learning: Assignment 3

Due on April 1st, 2014 at 11:59am

Barnabas Poczos, Aarti Singh

Instructions: Failure to follow these directions may result in loss of points.

- Your solutions for this assignment need to be in a pdf format and should be submitted to the blackboard and the webpage `http://barnabas-cmu-10701.appspot.com` for peer-reviewing.
- For the programming question, your code should be well-documented, so a TA can understand what is happening.
- DO NOT include any identification (your name, andrew id, or email) in both the content and filename of your submission.

K-Means (Prashant)

K-Means (20 points)

In this problem we will look at the K-means clustering algorithm. Let $X = \{x_1, x_2, \ldots, x_n\}$ be our data and γ be an indicator matrix such that $\gamma_{ij} = 1$ if x_i belongs to the j^{th} cluster and 0 otherwise. Let $\mu_1, \ldots, \mu_k$ be the means of the clusters.
We can define the distortion J as follows

$$J(\gamma, \mu_1, \ldots, \mu_k) = \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} ||x_i - \mu_j||^2$$

Finally, we define $C = 1, \ldots, k$ be the set of clusters.
The most common form of the K-means algorithm proceeds as follows

- Initialize $\mu_1, \ldots, \mu_k$.
- While J is decreasing, repeat the following

  1. Determine γ breaking ties arbitrarily.

$$\gamma_{ij} = \begin{cases} 1, & ||x_i - \mu_j||^2 \leq ||x_i - \mu_{j'}||^2, \forall j' \\ 0, & \text{otherwise} \end{cases}$$

  2. Recompute μ_j using the updates γ Remove j from C if $\sum_{i=1}^{n} \gamma_{ij} = 0$. Otherwise,

$$\mu_j = \frac{\sum_{i=1}^{n} \gamma_{ij} x_i}{\sum_{i=1}^{n} \gamma_{ij}}$$

1. Show that this algorithm will always terminate in a finite number of steps. (How many different values can γ take?) (4 points)

   From class we know that J is a monotonic decreasing with every iteration of Kmeans. This means that we cannot revisit the same state of γ twice. Since γ has n^2 entries, each of which is either 0 or 1, it has a finitely many number of possible values. This implies Kmeans must eventually terminate
   Award 4 points to any answer that correctly details why it terminates in finitely many steps.

2. Let $\hat{x}$ be the sample mean. Consider the following quantities,

$$T(X) = \frac{\sum_{i=1}^{n} ||x_i - \hat{x}||^2}{n}$$

$$W_j(X) = \frac{\sum_{i=1}^{n} \gamma_{ij} ||x_i - \mu_j||^2}{\sum_{i=1}^{n} \gamma_{ij}}$$

$$B(X) = \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \gamma_{ij}}{n} ||\mu_j - \hat{x}||^2.$$

   Here, $T(X)$ is the total deviation, $W_j(X)$ is the intra-cluster deviation and $B(X)$ is the inter-cluster deviation. What is the relation between these quantities? Based on this, show that K-means can be views as minimizing a weighted average of intra-cluster deviation while approximately maximizing the inter-cluster deviation. Your relation may contain a term that was not mentioned above. (5 points)

We start from

$$\begin{aligned}
\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}W_j(X) + nB(X) &= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}||x_i-\mu_j||^2 + \gamma_{ij}||\mu_j-\hat{x}||^2 \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}(||x_i-\mu_j||^2 + ||\mu_j-\hat{x}||^2) \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}(x_i^2+\hat{x}^2-2x_i\hat{x}+2x_i\hat{x}+\mu_j^2-2x_i\mu_j-2\hat{x}\mu_j \\
&= \sum_{j=1}^{k}(\sum_{i=1}^{n}\gamma_{ij}(||x_i-\hat{x}||^2)) + K
\end{aligned}$$

Here K is the remainder of the summation clumped up into one term.

$$\begin{aligned}
\sum_{j=1}^{k}(\sum_{i=1}^{n}\gamma_{ij}(||x_i-\hat{x}||^2)) + K &= n\sum_{i=1}^{n}(||x_i-\hat{x}||^2) + K \\
&= n^2T(X) + K
\end{aligned}$$

Therefore, $\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}W_j(X) + nB(X) = n^2T(X) + K$.
Note that $\sum_{j=1}^{k}\sum_{i=1}^{n}\gamma_{ij}W_j(X)$ is the same as J and is therefore minimized during K-means. Since $n^2T(X)$ is constant, this means that $nB(X)$ is approximately maximized (approximately because of the other term.
Award 3 points to any justified equation relating the terms even if its not the one presented here. Subtract unto two points for an equation provided that is not justified.
Award 2 points to any solution that shows the terms contained W_j are monotonic decreasing and since T is a constant, the $B_j's$ must be approximately maximized (approximately because of the extra term). Subtract 1 point if it is not justified that the $W_j's$ are minimized.

3. Show that the minimum of J is a non increasing function of k the number of clusters. Argue that this means it is meaningless to choose the number of clusters by minimizing J. (4 points)

This can be seen by a simple induction argument. Assume that till some k, J has been non decreasing in k. Let us now add another cluster center to some arbitrary location. Since this new run of Kmeans has not yet converged, we know that we are not at the minimum possible J for $k+1$ clusters. If we show that J is still non-decreasing, then we have completed our induction step. Observe that at least for the point that is now a cluster center, the term in J will be 0. This means that J has decreased when we have added a new cluster.
If we were to pick the k that minimized J, we would end up picking $k = n$ since this makes J 0. since we do not want to vastly overfit, we cannot simply arg max for k.
Award 2 points to any correct reasoning for why J is non increasing.
Award 2 points to the correct reasoning for why we cannot pick k from minimizing J.

4. Assume that now we use the ℓ_1 norm in J as opposed to the squared Euclidean distance

$$J'(\gamma, \mu_1, \ldots, \mu_k) = \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_{ij} ||x_i - \mu_j||_1$$

- Derive the steps to this new K-means formulation. Note that the answers may not be unique. (5 points)
- If your data contains outliers, which version of K-means would you use - the ℓ_1 norm one or the original Euclidean norm one? Justify. (2 points)

– Initialize $\mu_1, \ldots, \mu_k$.

– While J is decreasing, repeat the following

(a) Determine γ breaking ties arbitrarily.

$$\gamma_{ij} = \begin{cases} 1, & ||x_i - \mu_j||_1 \leq ||x_i - \mu_{j'}||_1, \forall j' \\ 0, & \text{otherwise} \end{cases}$$

(b) Recompute μ_j using the updates γ Remove j from C if $\sum_{i=1}^{n} \gamma_{ij} = 0$. Otherwise,

$$\mu_j = median(x_j | \gamma_{ij} = 1)$$

The goal is to maximize $\sum_{i=1}^{m} ||x_i - \mu_j||_1$ for the j^{th} cluster. Picking the μ that maximizes this will give us our centroid. Let us consider the derivate of this w.r.t μ. Each of the L1 norm terms will be either 1 or -1 depending on whether or not the L1 norm of the point is greater than μ. This derivative is 0 when the sum of the 1s and the -1s is 0. This happens when there are an equal number of points on either side of the centroid μ. The choice of centroid that achieves this is the median of the L1 norms of the points in the cluster. This solution may not be unique for a given cluster.
We might want the L1 norm for noisy data because the geometric median is robust to outliers. It only cares about picking the middle point in the ordering. It does not care about the distance of the furthest points to the center, only the number of points on each side.
Award 3 points to this answer or any other answer that picks the median
Award 2 points to an answer that states that L1 loss is robust to outliers.

Expectation Maximization (Dani)

In Naive Bayes, the joint likelihood of the data is:

$$\begin{aligned} p(\mathcal{D}) &= \prod_{i=1}^{N} p(X_i, Y_i) \\ &= \prod_{i=1}^{N} p(Y_i) \prod_{j=1}^{M} p(X_i^j | Y_i) \end{aligned}$$

Let us assume that $Y_i \in \{0, 1\} \forall i$ and $X_i^j \in \{1, 2, \ldots, V\} \forall i, j$. Denote the parameters of $p(Y)$ by θ and the parameters of $p(X|Y)$ by β. In the presence of labeled data (i.e., Y_i is observed for all i), we can get estimates of $\boldsymbol{\theta} = \{\theta_0, \theta_1\}, \boldsymbol{\beta} = \{\beta_{1|0}, \beta_{2|0}, \ldots, \beta_{V|0}, \beta_{1|1}, \beta_{2|1}, \ldots, \beta_{V|1}\}$ by counting and normalizing (you did

this in Homework 1). We denote random variables by capital letters and their values by lowercase letters (e.g., Y_i denotes a random variable, y_i denotes its assignment from the set of possible values of random variable Y_i).

Learning With Missing Data (13 points)

Suppose that we do not have labeled data. We can still estimate these parameters using the Expectation Maximization (EM) algorithm.

- Specify the (log) likelihood function that needs to be maximized (1 point; hint: the likelihood function will now be a summation over all possible assignments to all latent variables).
- Derive the E-step, i.e., compute the probability of all class assignments for each data point, given current parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. (5 points)
- Derive the M-step, i.e., compute the parameter updates for each class, given the class assignment distributions for each point from E-step. (5 points)
- Specify a good initialization technique and describe your rationale. (2 points)

$$\begin{aligned}
\log p(\mathcal{D}) &= \sum_{i=1}^{N} \log \sum_{y_i \in \{0,1\}} p(y_i) \prod_{j=1}^{M} p(x_i^j | y_i) \\
&= \sum_{i=1}^{N} \log \sum_{y_i \in \{0,1\}} q(y_i) \frac{p(y_i) \prod_{j=1}^{M} p(x_i^j | y_i)}{q(y_i)} \\
&\geq \sum_{i=1}^{N} \sum_{y_i \in \{0,1\}} q(y_i) \log \frac{p(y_i) \prod_{j=1}^{M} p(x_i^j | y_i)}{q(y_i)} \\
&= \sum_{i=1}^{N} \sum_{y_i \in \{0,1\}} q(y_i) \left[\log p(y_i) + \sum_{j=1}^{M} \log p(x_i^j | y_i) - \log q(y_i) \right]
\end{aligned}$$

Note: question 1 only asks about the original log likelihood (the first line). Give 1 point if the first line is correct, even though the following lines aren't (we will deduct points for this in the E-step and M-step since the updates must be wrong if the lower bound is wrong).
In the E-step, for all i, we update $q(y_i)$ given the current values of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. Since we know the current values of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, we can compute $p_{\boldsymbol{\theta}}(y_i)$ and $p_{\boldsymbol{\beta}}(x_i^j | y_i)$. Taking the derivative with respect to $q_(y_i)$ and set it to zero, we get the following update rule:

$$\begin{aligned}
q(Y_i = 0) &\propto p_{\boldsymbol{\theta}}(Y_i = 0) \prod_{j=1}^{M} p_{\boldsymbol{\beta}}(x_i^j | Y_i = 0) \\
q(Y_i = 1) &\propto p_{\boldsymbol{\theta}}(Y_i = 1) \prod_{j=1}^{M} p_{\boldsymbol{\beta}}(x_i^j | Y_i = 1)
\end{aligned}$$

If the lower bound (line 4 above) is correct but the E-step update is not correct, you can give 2 points. Also, an answer that simply says $q(Y_i = 1) = 1 - q(Y_i = 0)$ is fine as long as the update for $q(Y_i = 0)$ is correct.
In the M-step, we know the values of $q(y_i)$, so this step is very similar to supervised Naive Bayes updates. Maximizing the objective function with respect to the parameters gives the following update:

$$\begin{aligned}
\theta_0 &\propto \sum_{i=1}^{N} q(Y_i = 0) \\
\theta_1 &\propto \sum_{i=1}^{N} q(Y_i = 1) \\
\beta_{v|0} &\propto \sum_{i=1}^{N} q(Y_i = 0) \text{count}(x_i^j == v) \\
\beta_{v|1} &\propto \sum_{i=1}^{N} q(Y_i = 1) \text{count}(x_i^j == v)
\end{aligned}$$

Note: 1.25 points each. Also, an answer that simply says $\theta_1 = 1 - \theta_0$ is fine as long as the update for θ_0 is correct.
A good initialization technique should incorporate our knowledge about the possible solution. If we think one class is more likely than the other, use this in our initialization of $\boldsymbol{\theta}$. For the $\boldsymbol{\beta}$ parameters, if we think some words are indicative of a class, we should also set the corresponding β for that word to be higher. In the absence of prior knowledge, we can initialize these parameters randomly. **Note: you can give 2 points if the rationale makes sense. For example, an answer that says "random initialization" because we do not know anything else about the problem is fine. However, if the rationale does not make sense, give only 1 point.**

"Hard" EM [6 points]

Instead of summing over all possible assignments to all latent variables, we can instead set the values of the latent variables to their most likely values under current parameter estimates. This is often called "hard" EM, which sometimes work well in practice (for example, K-means clustering is learned using "hard" EM). For the unsupervised Naive Bayes problem above, we can also use "hard" EM to estimate the parameters.

- Show that if we replace the summation of possible assignments to latent variables with maximization, we are still optimizing a lower bound on the (log) likelihood function (2 points).
- What is the E-step? (2 points)
- What the M-step? (2points)

There are many ways to do this. The easiest way is to note that:

$$\log p(\mathcal{D}) = \sum_{i=1}^{N} \log \sum_{y_i \in \{0,1\}} p(y_i) \prod_{j=1}^{M} p(x_i^j | y_i)$$
$$\geq \sum_{i=1}^{N} \log \max_{y_i \in \{0,1\}} p(y_i) \prod_{j=1}^{M} p(x_i^j | y_i),$$

since the term inside the sum is always positive. Another way is to note that the soft EM bound holds for any q distribution, including when we choose $q(Y_i = 0) \in 0, 1$ and $q(Y_i = 1) = 1 - q(Y_i = 0$. **Note: deduct 1 point if the answer does not say that the term inside the sum is always positive if using the first way.**

In "Hard" EM, our objective function can be rewritten as:

$$\log p(\mathcal{D}) = \sum_{i=1}^{N} \log \max_{y_i \in \{0,1\}} p(y_i) \prod_{j=1}^{M} p(x_i^j | y_i)$$
$$= \sum_{i=1}^{N} \max_{y_i \in \{0,1\}} \log p(y_i) \prod_{j=1}^{M} p(x_i^j | y_i),$$

since log is a monotonic function.

In the "E-step", we find the max under current parameter estimates (i.e., we do MAP inference on the hidden variables y). We compute $\hat{y}_i = \arg\max_{y_i \in \{0,1\}} \log p(y_i) \prod_{j=1}^{M} p(x_i^j | y_i)$.

Given $\hat{y}_i$, in the M-step, we update the objective function by taking the derivative of the objective function with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. We can rewrite the objective function as:

$$\log p(\mathcal{D}) = \sum_{i \in \mathcal{N}_0} \log p(y_i) \prod_{j=1}^{M} p(x_i^j | y_i) + \sum_{i \in \mathcal{N}_1} \log p(y_i) \prod_{j=1}^{M} p(x_i^j | y_i),$$

where $\mathcal{N}_0$ and $\mathcal{N}_1$ are disjoint sets of indices $i \in \{1, \ldots, N\}$, where $\hat{y}_i = 0$ and $\hat{y}_i = 1$ respectively. This is again similar to the parameter updates for supervised Naive Bayes, so we get the following M-step:

$$\theta_0 \propto \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i == 0)$$
$$\theta_1 \propto \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i == 1)$$
$$\beta_{v|0} \propto \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i == 0) \text{count}(x_i^j == v)$$
$$\beta_{v|1} \propto \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i == 1) \text{count}(x_i^j == v)$$

where $\mathbb{I}$ is an indicator function.

EM in Practice [1 point]

EM converges only to a local optimum. Give a high-level description of a strategy you would use to get reasonably good parameter estimates when using EM in practice.

Inspect the log likelihood and choose the model that has the highest likelihood from multiple (possibly random) initializations. If a development data is available, we can use a development data and choose a model that has high likelihood on development data to avoid overfitting to the training data. **Note: use your best judgement, you can give 1 point if the strategy makes sense even though it is not what I described above.**

Hidden Markov Models (Pengtao)

HMMs (20 points)

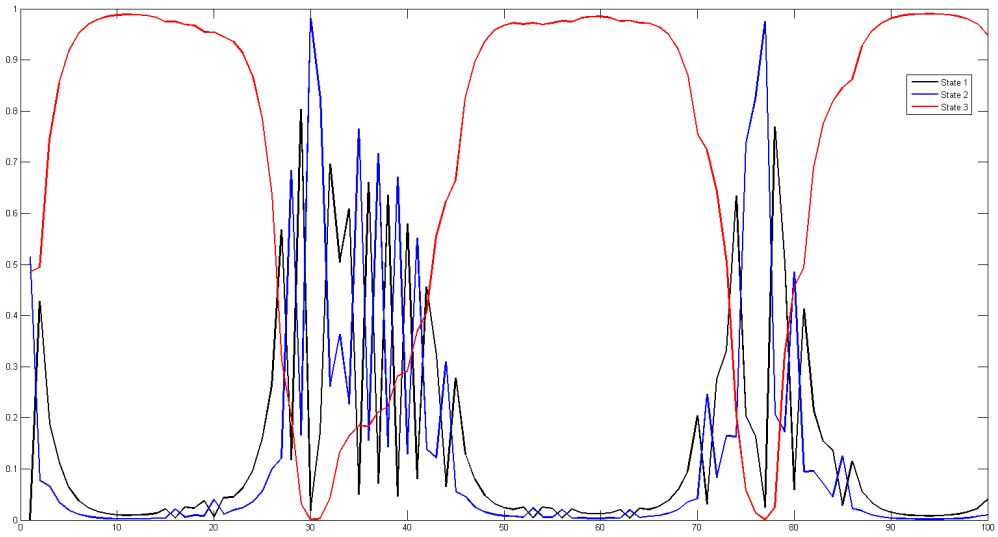
In this problem, we will use Hidden Markov Model (HMM) to detect latent topics from documents. We assume the documents are written with 5 words (you can think of them as Obama, basketball, congress, GDP, NBA) and contains three latent topics (Politics, Sports, Economics). Each topic has a multinomial distribution over words. For example, the politics topic might have such a multinomial distribution $(0.4, 0.05, 0.4, 0.1, 0.05)$ over the vocabulary (Obama, basketball, congress, GDP, NBA). This topic puts a high probability mass 0.4 over Obama because politics is highly corrected with Obama while puts a low probability mass 0.05 over basketball since politics has little to do with basketball. Moreover, think about the transition between topics. When writing an article, the author is more likely to change the topic from politics to economics, than to make a transition from politics to sports. Given a sequence of tokens, we are interested in: which topic is each token generated from? HMM can be utilized to answer this question. We model topics as latent states and use a transition probability matrix A to describe the transition between topics. $A_{ij} = p(z_t = j|z_{t-1} = i)$, where z_t and z_{t-1} are topic assignments of tokens at position t and $t-1$ respectively. Topics' multinomial distribution over words are put into the emission probability matrix O. $O_{ik} = p(x = k|z = i)$, where x denotes a word and z denotes a topic. (Note: be aware of the difference between words and tokens. In natural language processing, word refers to each item in a vocabulary. For example, given a vocabulary containing three items {apple, car, dog}, these three items are called words. Documents are composed of these items. In a document, strings separated with blanks are called tokens. For example, given a sentence "I love dog because dog is lovely", it contains seven tokens "I", "love", "dog", "because", "dog", "is", "lovely".)
We provide you the learned parameters in the folder "hmm_paras" in the handout.

- transition.txt: the 3×3 transition probability matrix A, where $A(i, j) = p(z_t = j|z_{t-1} = i)$
- prior.txt: the prior distribution over z_1, where $prior(i) = p(z_1 = i)$
- emission.txt: the 3×5 emission probability matrix O, where $O(i, k) = p(x = k|z = i)$
- tokens.txt: a sequence of 128 tokens $X = \{x_t\}_{t=1}^{128}$ in one document.

Here are the details of your tasks:

- 1. [10 pts] Implement the Forward-Backward algorithm and infer the posterior distribution of hidden states (topics) given the observed tokens. Report the inferred distributions over the hidden states by plotting the probabilities $p(z_t = i|X)$ for $i = 1, 2, 3$ over $t = 1, \cdots, 128$. Make sure you label the 3 topics (one for each hidden state) in your plot. (Hint: in the plot, the x-axis is t, the y-axis is probability $p(z_t = i|X)$. For $i = 1$, you get a sequence of numbers $p(z_1 = 1|X), p(z_2 = 1|X), \cdots, p(z_{128} = 1|X)$, plot them over $t = 1, \cdots, 128$. Do the same thing for $i = 2$ and $i = 3$. By plotting the probability change for each topic, you can see the trend of each topic.)
- 2. [10 pts] Implement the Viterbi algorithm and find the most likely sequence of hidden states. Report the most likely hidden states $\{\hat{z}_t\}_{t=1}^{128}$ by plotting their values over $t = 1, \cdots, 128$. (Hint: in the plot, the x-axis is t, the y-axis is the most probable hidden state $\hat{z}_t$, which can take values of $1, 2, 3$. Again

Figure 1: Plot of problem 1



make sure you label which topic (Politics, Sports, Economics) is associated with which numeric value (1,2,3) in your plot. By plotting this curve, you can visualize which topic is been discussed in different segments of the document.)

Decision Trees (Pulkit)

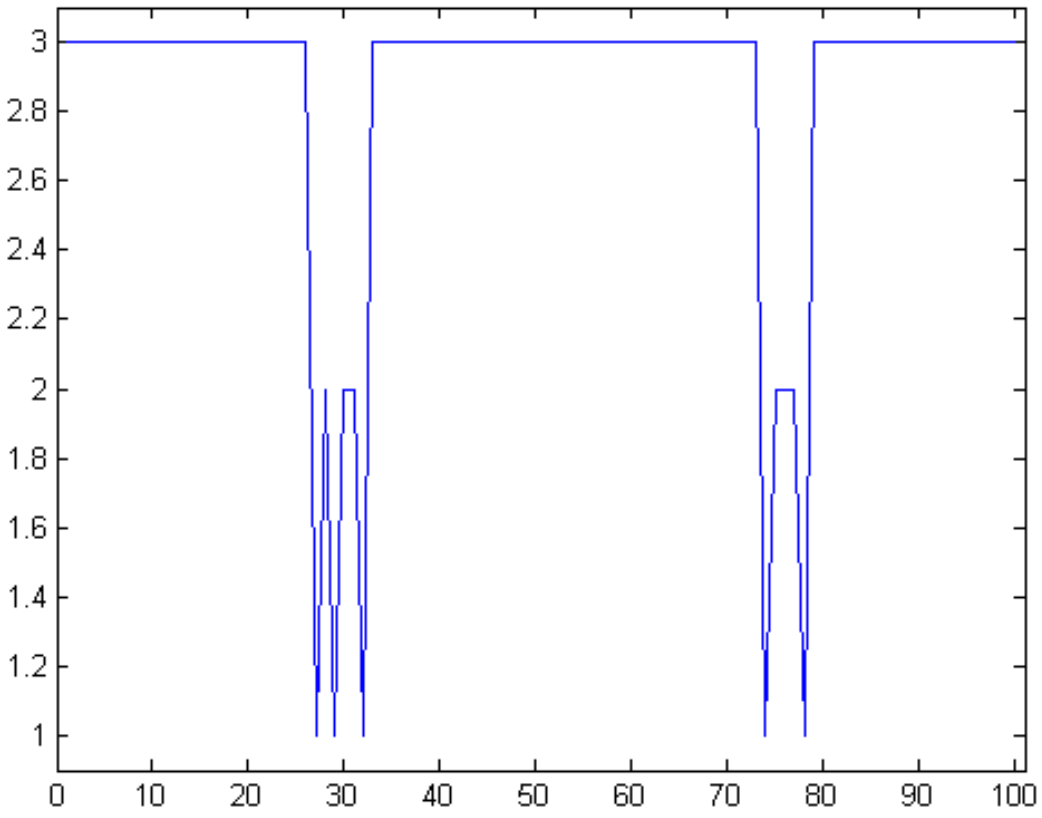
1. The following is a small synthetic data set about the conditions of ill patients with tumors. We are going to try and use decision trees to predict the malginancy of a tumor.
   You may assume the following about the tree building algorithm :

   (i) The decision tree uses the ID3 algorithm for building the tree - each attribute is used only once as an internal node.

   (ii) You can treat age as a continuous variable and split on a range of age values.

   (iii) Attribute selection happens through information gain

Figure 2: Plot of problem 2



Age	Vaccination	Tumor Size	Tumor Site	Malignant
5	1	Small	Shoulder	0
9	1	Small	Knee	0
6	0	Small	Marrow	0
6	1	Medium	Chest	0
7	0	Medium	Shoulder	0
8	1	Large	Shoulder	0
5	1	Large	Liver	0
9	0	Small	Liver	1
8	0	Medium	Shoulder	1
8	0	Medium	Shoulder	1
6	0	Small	Marrow	1
7	0	Small	Chest	1

(a) What is the initial entropy of *Malignant*? (2 points)

(b) Which attribute would the decision-tree building algorithm choose at the root of the tree? Choose one through inspection and explain your reasoning in a sentence. (2 points)

(c) Calculate and specify the information gain of the attribute you chose to split on in the previous question. (3 points)

(d) Draw the full decision tree for the data. (Note: You do not need to calculate the information gain for each attribute. The choices can be made through inspection) (3 points)

(a)

$$-\frac{5}{12}log_2\frac{5}{12} - \frac{7}{12}log_2\frac{7}{12} = 0.9798$$

Rubric: 2 points for correct calculation.

(b) Vaccination. It splits the training data with maximum determinability. Those with vaccination are certain to not be malignant.

Rubric: 1 point for attribute. 1 point for reasoning. Any valid reasoning can get full points.

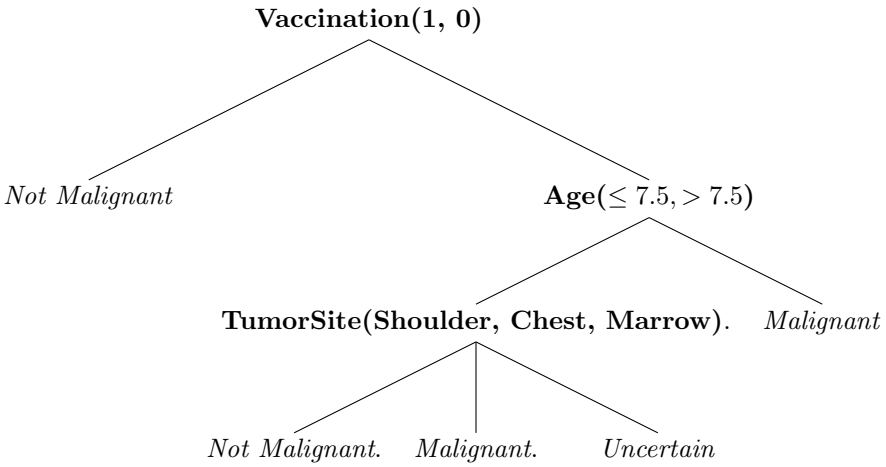
(c)

$$0.9798 - (\frac{5}{12}(-\frac{0}{5}log_2\frac{0}{5} - \frac{5}{5}log_2\frac{5}{5}) - \frac{7}{12}(-\frac{2}{7}log_2\frac{2}{7} - \frac{5}{7}log_2\frac{5}{7})) = 0.4763$$

(Note: we define $0log0 = 0$)

Rubric: 3 points for correct calculation. Partial grade - 1 point for writing down the values correctly in terms of information gain. Deduct 1 point if there is a minor error in calculation.

(d)



Rubric: 3 points for correct decision tree. Deduct 1 point for each of the attributes/labels that are wrong.

2. Consider a decision tree built from an arbitrary set of data. If the output is discrete-valued and can take on k possible values, what is the maximum *training* set error (expressed as a fraction) that a data set could possibly have? (Please note that this is the error on the same dataset the tree was trained on. A new test set could have arbitrary errors.)
Write a small data set which attains the maximum error for a tree in the case when the output can take k = 2 possible values (Please limit this to 1-2 input variables, and 4-5 training examples.) (5 points)

$\frac{k-1}{k}$. The maximum confusion for the decision tree happens when the data is distributed evenly amongst all the k discrete outputs. In this case all the training samples will be labeled as a single value. Still $\frac{1}{k}$ results would be right.

Rubric: 2 points for the correct answer. 1 point for reasoning.

X	Y
0	0
0	1
1	0
1	1

Rubric: 2 points for corret answer. There may be other acceptable solutions.

3. We will explore the link between KL-Divergence and Information Gain in this question. The KL-divergence from a distribution $p(x)$ to a distribution $q(x)$ can be thought of as a distance measure from p to q:

$$KL(p||q) = -\sum p(x) log_2 \frac{q(x)}{p(x)}$$

From an information theory perspective, the KL-divergence specifies the number of additional bits required on average to transmit values of x if the values are distributed with respect to p(x) but we encode them assuming the distribution q(x). If $p(x) = q(x)$, then $KL(p||q) = 0$. Otherwise, $KL(p||q) > 0$. The smaller the KL-divergence, the more similar the two distributions. We can define information gain as the KL-divergence from the observed joint distribution of X and Y to the product of their observed marginals.

$$IG(x,y) \equiv KL(p(x,y)||p(x)p(y)) = -\sum_x \sum_y p(x,y) log_2 (\frac{p(x)p(y)}{p(x,y)})$$

Show that $IG(x,y) = H[x] - H[x|y] = H[y] - H[y|x]$ by starting from the equation specified above. (Note that showing it in one direction is enough.) (5 points)

$$KL(p(x,y)||p(x)p(y)) = -\sum_x \sum_y p(x,y) log_2 (\frac{p(x)p(y)}{p(x,y)})$$

$$= -\sum_x \sum_y p(x,y) log_2 (\frac{p(x)}{p(x|y)})$$

$$= -\sum_x \sum_y p(x,y) log_2 (p(x)) + \sum_x \sum_y p(x,y) log_2 (p(x|y))$$

$$= -\sum_x p(x) log_2 (p(x)) + \sum_y p(y) \sum_x p(x|y) log_2 (p(x|y))$$

$$= H(X) - H(X|Y)$$

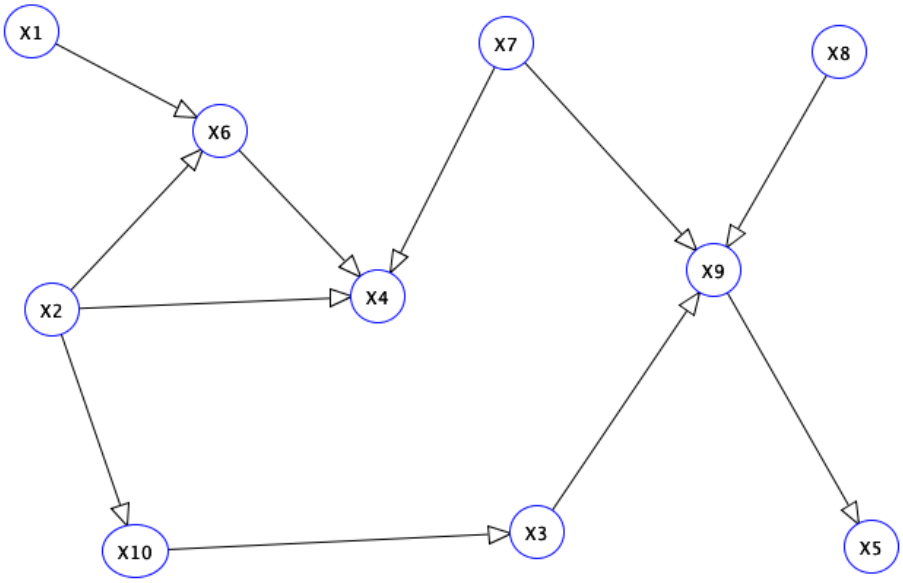
$$= IG(X,Y)$$

Rubric: 5 points for correct derivation. 1 point per step can be given for partial grade.

Graphical Models (Jit)

D-Separation [10 points]

For the following three questions, please refer to the Bayesian Network provided below.



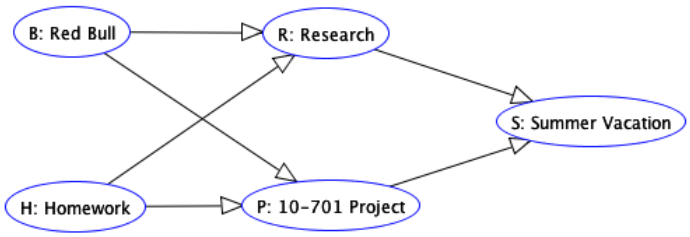
1. What is the largest set A, such that $X_1 \perp X_A | X_2, X_9$?
2. What is the largest set B, such that $X_{10} \perp X_B | X_3$?
3. What is the factorization of the joint probability distribution that creates the same independence relations as those specified by the Bayesian network?

1. A = $\{X_7, X_8, X_5, X_3, X_{10}\}$
2. B = $\{X_1, X_7, X_9, X_5, X_8\}$
   Grading: $\frac{1}{2}$ point for each correct node in the set. Each part is worth 2.5 points. If the student's answer contains all the correct nodes plus some incorrect ones, subtract $\frac{1}{2}$ point for each incorrect one. To avoid double-counting mistakes, subtract points for incorrect nodes only when the student's solution contains all of the correct nodes.
3. $P(X_1, ..., X_{10}) = P(X_5|X_9)P(X_9|X_3, X_7, X_8)P(X_3|X_{10})$
   $P(X_{10}|X_2)P(X_4|X_2, X_6, X_7)P(X_6|X_1, X_2)P(X_1)P(X_2)P(X_7)P(X_8)$
   Grading: $\frac{1}{2}$ point for each conditional probability in the factorization. Because there should be 10 factors, this question is worth 5 points.

Inference on Bayesian Network (10 points)

Now, let's consider the following binary Bayesian Network reflecting the life of a typical 10-701 student. The variables are as described below:

B: Red Bull	Indicator that there is a constant supply of Red Bull available.
H: Homework	Indicator that the last 10-701 assignment needs to be completed.
R: Research	Indicator that Research needs to be completed.
P: 10-701 Project	Indicator that the 10-701 Project still needs to be finished.
S: Summer Vacation	Indicator that the student will go on a relaxing summer vacation.



Find the probability that a student goes on a relaxing summer vacation, given that (s)he has a constant supply of Red Bull and the last 10-701 assignment needs to be completed. Find the probability that a student does not go on a relaxing summer vacation, given (s)he has a constant supply of Red Bull and the last 10-701 assignment is completed. Use the following probabilities:

$$Pr(B = T) = 0.3$$
$$Pr(H = T) = 0.85$$
$$Pr(R = T|B = T, H = T) = 0.6$$
$$Pr(R = T|B = T, H = F) = 0.9$$
$$Pr(R = T|B = F, H = T) = 0.05$$
$$Pr(R = T|B = F, H = F) = 0.35$$
$$Pr(P = T|B = T, H = T) = 0.45$$
$$Pr(P = T|B = T, H = F) = 0.75$$
$$Pr(P = T|B = F, H = T) = 0.25$$
$$Pr(P = T|B = F, H = F) = 0.55$$
$$Pr(S = T|P = T, R = T) = 0.10$$
$$Pr(S = T|P = T, R = F) = 0.65$$
$$Pr(S = T|P = F, R = T) = 0.15$$
$$Pr(S = T|P = F, R = F) = 0.80$$

$Pr(S|b,h) = \sum_{B,H,R,P} Pr(S|P,R)Pr(R|B,H)Pr(P|B,H)\delta(B=b)\delta(H=h)$
$Pr(S|b,h) = \sum_{R,P} Pr(S|P,R) \sum_{B,H} Pr(R|B,H)Pr(P|B,H)\delta(B=b)\delta(H=h)$
$Pr(S|b,h) = \sum_{R,P} Pr(S|P,R)Pr(R|B=b,H=h)Pr(P|B=b,H=h)$

P	R	Pr(S = T, P = p, R =r | P = p, R = r)	Pr(S = F , P = p, R = r| P = p, R = r)
T	T	0.10*0.45*0.60 = 0.027	0.90*0.45*0.60 = 0.243
T	F	0.65*0.45*0.40 = 0.117	0.35*0.45*0.40 = 0.063
F	T	0.15*0.55*0.60 = 0.0495	0.85*0.55*0.60 = 0.2805
F	F	0.80*0.55*0.40 = 0.176	0.20*0.55*0.40 = 0.044
		Total: 0.3695	Total: 0.6305

Therefore

$Pr(S=T|H=T,B=T)$
$= Pr(S=T|P=T,R=T)\ Pr(P=T|H=T,B=T)$
$Pr(R=T|H=T,B=T) + Pr(S=T|P=T,R=T)$
$Pr(P=T|H=T,B=T)\ Pr(R=F|H=T,B=T) + Pr(S=T|P=T,R=T)\ Pr(P=F|H=T,B=T)$
$Pr(R=T|H=T,B=T) + Pr(S=T|P=T,R=T)\ Pr(P=F|H=T,B=T)\ Pr(R=F|H=T,B=T)$
$= 0.027 + 0.117 + 0.0495 + 0.176 = 0.3695$

Now, we want to find $Pr(S=T|H=F,B=T)$. Using $Pr(S|b,h) = \sum_{R,P} Pr(S|P,R)Pr(R|B=b,H=h)Pr(P|B=b,H=h)$, we get:

P	R	Pr(S = T, P = p,R = r | P = p, R = r)	Pr(S = F, P = p,R = r | P = p, R = r)
T	T	0.10*0.90*0.75 = 0.0675	0.90*0.90*0.75 = 0.6075
T	F	0.65*0.10*0.75 = 0.04875	0.35*0.10*0.75 = 0.02625
F	T	0.15*0.90*0.25 = 0.03375	0.85*0.90*0.25 = 0.19125
F	F	0.80*0.10*0.25 = 0.02	0.20*0.10*0.25 = 0.005
		Total: 0.17	Total: 0.83

Therefore, $Pr(S=F|H=F,B=T) = Pr(S=F|P=T,R=T)\ Pr(P=T|H=T,B=T)\ Pr(R=T|H=T,B=T) + Pr(S=F|P=T,R=T)\ Pr(P=T|H=T,B=T)\ Pr(R=F|H=T,B=T) + Pr(S=F|P=T,R=T)\ Pr(P=F|H=T,B=T)\ Pr(R=T|H=T,B=T) + Pr(S=F|P=T,R=T)\ Pr(P=F|H=T,B=T)\ Pr(R=F|H=T,B=T) = 0.6075 + 0.02625 + 0.19125 + 0.005 = 0.83$

Grading: Each part is worth 5 points total. The students did not have to include a full table of probabilities like I did. This is for the reviewer's benefit. I'm hoping you could use the table to provide more concise feedback (Eg. "Based on your answer, you calculated X when you're supposed to calculate Y"). Using variable elimination, each marginal should be calculated as the sum of four sub-factors (i.e $Pr(S=T|P=T,R=T)\ Pr(P=F|H=T,B=T)\ Pr(R=F|H=T,B=T)$ is one sub-factor for $P(S=T|H=T,B=T)$). Give 1 point for each correct sub-factor, and give 1 point for a correct answer for the marginal probability.