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.
- We are NOT going to use Autolab in this assignment.
- 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)
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)

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)
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)

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)

"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)

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.

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 $1, 2, 3$. 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

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)

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)

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.

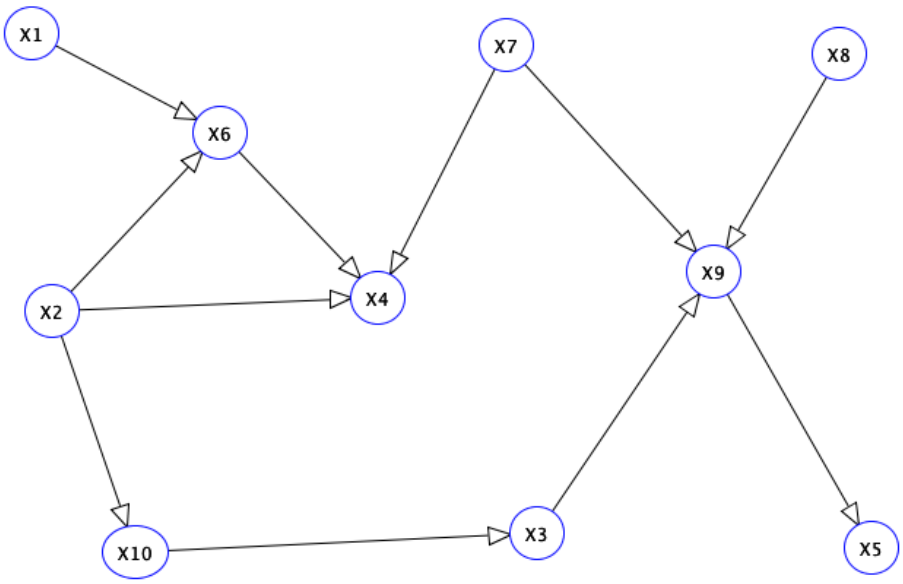
$$IG(x,y) \equiv KL(p(x,y)||p(x)p(y)) = -\sum_x \sum_y p(x,y) log_2(\frac{p(x)p(y)}{p(x,y)})$$

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)

Graphical Models (Jit)

D-Separation [10 points]

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

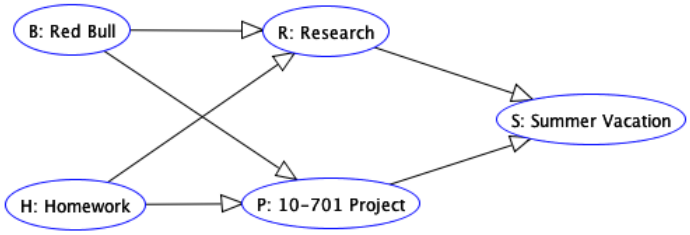


1. What is the largest set A, such that $X_1 \perp X_A | X_2, X_9$?
2. What is the largest set B, such that $X_{10} \perp X_B | X_3$?
3. What is the factorization of the joint probability distribution that creates the same independence relations as those specified by the Bayesian network?

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:

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