10-701 Machine Learning, Fall 2012: Homework 4

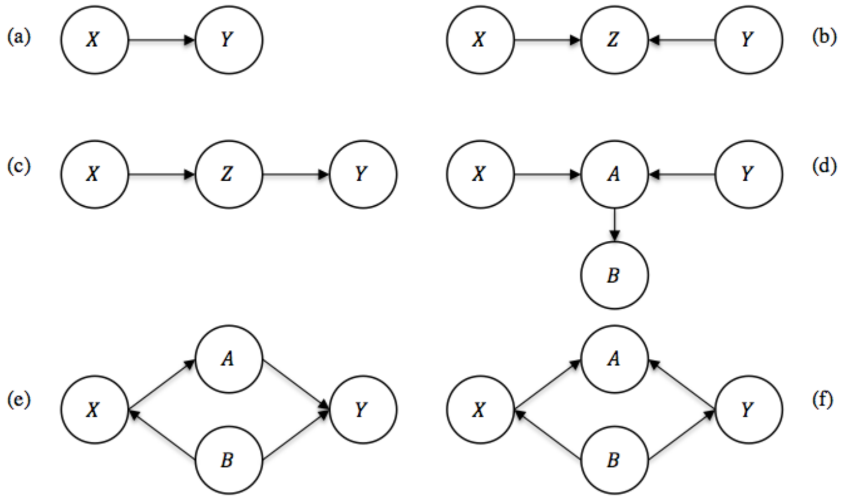
Due Wednesday 11/28 at the beginning of class.

Instructions (important): Remember to submit your solution to **each problem separately** in **different piles** on the table at the front of the classroom at the beginning of class on the day the homework is due. At the *top* of the first page of each solution you hand in, clearly write the class number (10701), the **number of the problem** (i.e. "Problem 1", "Problem 2", etc.), your **first and last name** (assuming you have both), and your Andrew ID.

1 Graphical models 2 (Zeyu, 30 points)

1.1 Short questions (14 points)

1. [**2 points**] Show that $a \perp (b, c)|d$ implies $a \perp b|d$
2. [**4 points**] Using the d-separation criterion, show that the conditional distribution for a node x in a directed graph, conditioned on all of the nodes in the Markov blanket, is independent of the remaining variables in the graph.
3. [**8 points**] Reversing the direction of all the arrows in a GM might give us the same GM, but sometimes it does not. Consider the following simple cases and identify what is the difference (about the independence/conditional independence among all these points) between the reversed GM and original one? Fill in the form below. Do not put duplicated independency statement in the same row (you may refer to the fact $a \perp (b, c)|d \Rightarrow a \perp b|d$ as in question 1)



No.	the same?	original model	new model
(a)	Yes		
(b)	No	$X \perp Y$	$X \perp Y \| Z$
(c)			
(d)			
(e)			
(f)			

1.2 Exact Inference (7 points)

In this question we are going to apply variable elimination over a tree structure. Consider the Bayesian Tree in figure 1. Assume each node takes the m values $\{v_1, ..., v_n\}$, and all the local probability are known, i.e. $p(g), p(c|g,h), p(a|b,c,d), etc$

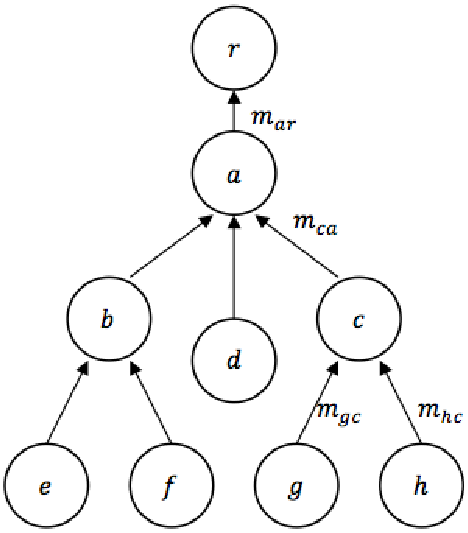


Figure 1: Bayesian Tree for question 2

1. **[3 points]** write the the following message function in terms of local probabilities and its precedence (e.g. m_{gc} and m_{hc} precedes m_{ca} in message passing)
   (a) $m_{hc}(v_i) = p(h) = ?$
   (b) $m_{ca}(v_i) = p(c) = ?$
   (c) $m_{ar}(v_i) = p(a) = ?$
2. **[4 points]** write the following conditional inference problem in terms of local probability and message function. Make your formula as compact as possible.
   (a) $P(r = v_i)$
   (b) $P(g = v_i | r = v_j)$

1.3 Stochastic Inference (9 points)

Recall Gibbs sampling algorithm and consider the butterfly-shaped Bayesian network in Figure 2. Assume all the nodes take binary values.

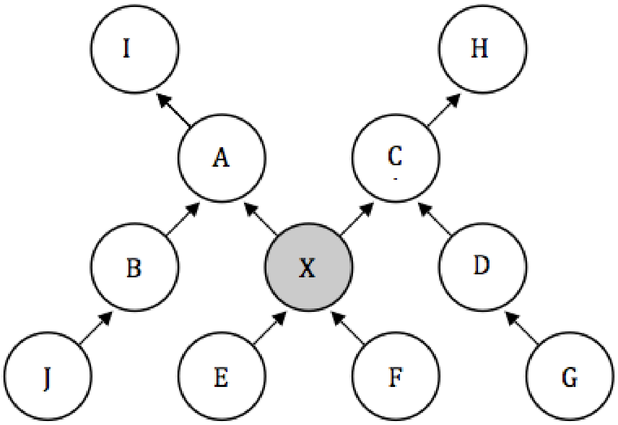


Figure 2: Find the Markov blanket of X

1. [**3 points**] What's the Markov Blanket, MB(X), of node X?
2. [**4 points**] Show how do you calculate the MCMC update probability $p(X|\text{MB}(X))$ using given local probability where MB(X) is the Markov blanket of node X
3. [**2 points, open question**] Suppose $(G = 1)$ is a very rare event $p(G = 1) < 0.0001$. but we have to deal with inference query that involve G=1. How can you modify Gibbs sampling method to avoid sampling too many samples? Show your idea for the following two scenarios.
   (a) G = 1 is in the condition (e.g. $P(A = 0, E = 1|G = 1)$)
   (b) G = 1 is the main event (e.g. $P(G = 1|E = 1)$).

2 Learning Theory [20 points, Martin]

2.1 VC dimension [10 points]

Recall that, given a hypothesis space H defined over an instance space X, the Vapnik–Chervonenkis dimension VC(H) is defined to be the largest integer such that there exists a subset $x_1, ..., x_{\text{VC}(H)} \in X$ that is shattered by H (i.e. for any binary labeling of $x_1, ..., x_{\text{VC}(H)}$ there exists a hypothesis $h \in H$ that is consistent with that labeling). Thus, in order to prove that the VC dimension of a hypothesis space H is some integer d, we must prove both that there exists a subset of X of size d that is shattered by H, *and* that for any $D > d$ there exist **no** subsets of X of size D that are shattered by H. The latter task can seem daunting – not only do we have to prove that **all** subsets of X of size D are **not** shattered by H, but we must also prove this for all $D > d$. The following result can make this much easier:

(a) **[5 points]** Prove that if there exists a subset of X of size d (for some integer d) that is shattered by H, then for any $1 \leq k < d$ there also exists a subset of size k of X that is shattered by H.

The above claim implies that if no subset of size d is shattered, then no subset of size $D > d$ is shattered either. Hence, to prove that the VC dimension of H is d, it is sufficient to find *some* subset of X of size d that is shattered, and to prove that *no* subset of size $d+1$ is shattered.

(b) **[5 points]** Let $X = \mathbb{R}$ (one dimension). Let H be the set of all classifiers h that, for some set of non-intersecting intervals $R_1, ..., R_p$, classify a point x as $h(x) = 1$ if $x \in \bigcup_{i=1}^{p} R_i$, and $h(x) = 0$ otherwise (p is fixed and given). Find $\mathrm{VC}(H)$ (prove your answer is correct).

2.2 Structural risk minimization [10 points]

Recall the PAC bound using VC dimension: given a hypothesis class H and m training samples, with probability $\geq 1 - \delta$, for all $h \in H$

$$|\mathrm{error}_{\mathrm{true}}(h) - \mathrm{error}_{\mathrm{train}}(h)| \leq \varepsilon(H, m, \delta)$$

where

$$\varepsilon(H, m, \delta) = 8\sqrt{\frac{\mathrm{VC}(H)\left(\ln \frac{m}{\mathrm{VC}(H)} + 1\right) + \ln \frac{8}{\delta}}{2m}}.$$

In class you saw how we can use this to bound the true error of the empirical risk minimizer $\widehat{h} = \mathrm{argmin}_{h \in H} \mathrm{error}_{\mathrm{train}}(h)$; with probability $\geq 1 - \delta$,

$$\begin{aligned}
\mathrm{error}_{\mathrm{true}}(\widehat{h}) &\leq \mathrm{error}_{\mathrm{train}}(\widehat{h}) + \varepsilon(H, m, \delta) \\
&\leq \mathrm{error}_{\mathrm{train}}(h^*) + \varepsilon(H, m, \delta) \\
&\leq \mathrm{error}_{\mathrm{true}}(h^*) + 2\varepsilon(H, m, \delta)
\end{aligned}$$

where $h^* = \mathrm{argmin}_{h \in H} \mathrm{error}_{\mathrm{true}}(h)$ is the true risk minimizer.

Given a set of hypothesis classes $H_1, H_2, ..., H_K$ (K possibly infinite) with $\mathrm{VC}(H_1) \leq \mathrm{VC}(H_2) \leq ... \leq \mathrm{VC}(H_K)$, *structural risk minimization* is the following procedure. First for each $k = 1, ..., K$ we find the empirical risk minimizer $\widehat{h}_k = \mathrm{argmin}_{h \in H_k} \mathrm{error}_{\mathrm{train}}(h)$ within H_k. Then we find

$$\widehat{k} = \underset{k=1,...,K}{\mathrm{argmin}} \left(\mathrm{error}_{\mathrm{train}}(\widehat{h}_k) + \varepsilon(H_k, m, \delta_k)\right),$$

(for some $\delta_1, ..., \delta_K$), and the structural risk minimizer is $\widehat{h} = \widehat{h}_{\widehat{k}}$.

Another possible procedure would be to simply use the empirical risk minimizer in the union of $H_1, ..., H_K$. In the next two problems, we'll try to see why structural risk minimization might be a better idea.

(a) **[5 points]** Let

$$\widehat{h}_{\mathrm{union}} = \underset{h \in \bigcup_{k=1}^{K} H_k}{\mathrm{argmin}} \mathrm{error}_{\mathrm{train}}(h)$$

and

$$h^*_{\mathrm{union}} = \underset{h \in \bigcup_{k=1}^{K} H_k}{\mathrm{argmin}} \mathrm{error}_{\mathrm{true}}(h).$$

Show an upper bound on $\text{error}_{\text{true}}(\widehat{h}_{\text{union}})$ in terms of $\text{error}_{\text{true}}(h^*_{\text{union}})$ (and some other terms) that holds with probability $\geq 1 - \delta$ for given δ. How does your bound simplify in the case that $H_1 \subseteq H_2 \subseteq ... \subseteq H_K$?

Let H_k be the set of "interval classifiers" (as defined in Problem 2.1(b)) with *up to* k intervals, and let $K = 100$. Give a lower bound on the number of samples m needed that is sufficient to *guarantee* that $\text{error}_{\text{true}}(\widehat{h}_{\text{union}}) - \text{error}_{\text{true}}(h^*_{\text{union}}) \leq 0.25$ with probability at least 0.95.

(b) **[5 points]** Give a lower bound on the number of samples m so that the structural risk minimizer $\widehat{h_{\widehat{k}}}$, computed using $\delta_1 = ... = \delta_K = 0.05/100$ on the sequence of hypothesis spaces defined in part (a), satisfies $\text{error}_{\text{true}}(\widehat{h_{\widehat{k}}}) - \text{error}_{\text{true}}(h^*_{\text{union}}) \leq 0.25$ with probability at least 0.95, assuming that $h^*_{\text{union}} \in H_5$.

3 Boosting [25pt, Derry]

Consider a stepwise algorithm **A**:

```
Input parameters: $T$, $\mathcal{H}$, $\phi$
Initialize the classifier $f_0(x) = 0$
for $t = 1$ to $T$ do:
    1. Compute
        $(h_t, \alpha_t) = argmin_{\alpha \in \mathbb{R}, h \in \mathcal{H}} \sum_{i=1}^{m} \phi(y_i, f_{t-1}(x_i) + \alpha h(x_i))$
    2. Update the classifier
        $f_t(x) = f_{t-1}(x) + \alpha_t h_t(x)$
end for
return the classifier $sign(f_T(x))$
```

The intuition is that, at each step, the algorithm greedily adds a hypothesis $h \in \mathcal{H}$ to the current hypothesis to minimize the ϕ-risk.

1. **[5 pts]** What would be the form of $\phi(y, y')$ that will make algorithm **A** equivalent to AdaBoost?

2. **[10 pts]** Using the risk function you have defined above, prove that AdaBoost is equivalent to algorithm **A**.

   Hint: Work out the value of h_t that will minimize the risk function $\phi(y, y')$ for any fixed value of $\alpha > 0$ (further hints: h_t is not a function on α). Then, given this h_t find the α_t that will minimize the risk function $\phi(y, y')$. Think also how the weights $D_t(i)$ in AdaBoost is related to algorithm **A**.

3. **[10 pts]** Now consider a more general algorithm where $h_t \in \mathcal{H}$ from $t = 1$ to T be any arbitrary sequence of classifiers. Let $\{x_i, y_i\}_{i=1}^m$ be a training set of m observations. Starting with $f_0 = 0$, f_t is recursively defined as $f_t = \sum_{i=1}^t \alpha_i h_i$ and $\alpha_t = \beta \; log \; \frac{1-\epsilon_t}{\epsilon_t}$ where

$$\epsilon_t = \sum_{i=1}^m D_{t-1}(i)\mathbf{1}\{y_i \neq h_t(x_i)\}$$

which is the weighted training error of the classifier h_t. Prove that for all T:

$$\sum_{i=1}^m \frac{1}{m} \; exp\;(-\frac{1}{\beta} y_i f_T(x_i)) = 1$$

which implies that any sequence of classifiers can be combined linearly to form a good combination while maintaining a constant exponential loss on the data.

4 PCA [Avi 25 pts]

In this question we will try to understand PCA by showing two cool ways of interpreting the first principal component. One is the direction of maximum variance after projection and the second is the direction that minimizes reconstruction error. Note that the first principal component is the first eigenvector of the sample covariance matrix.

Consider n points $X_1, ..., X_n$ in p-dimensional space, and let X be the $n \times p$ matrix representing these points. Assume that the data points are centered, ie, $\vec{1}^\top X = \vec{0}$. Consider a unit vector $v \in \mathbb{R}^p$ and project all the points onto this vector (hence every point becomes a one-dimensional point on the direction of unit vector v).

1 [1 pt] Argue that the projection is given by Xv.

2 [2 pt] What is the sample mean of all the points after the projection?

3 [2 pt] What is the sample variance of all the points after the projection?

4 [2 pt] Setup the problem of maximizing the sample variance of the projection onto v subject to a constraint on the L2-norm of v.

5 [4 pt] Solve the minimization problem to show that the solution is the first PC. (Hint: take the Lagrangian of the above problem, differentiate and substitute to zero, to get to the optimum solution)

So we have now proved that the direction of maximum covariance is the first PC. Now we show that the direction that minimizes reconstruction error is also the first PC.

6 [1 pt] Argue that the reconstruction of X_i using v is $(X_i^\top v)v$.

7 [2 pt] You projected X_i to $X_i^\top v$ and then reconstructed it using $(X_i^\top v)v$. What is the reconstruction error of X_i, when measured in L2-norm?

8 [2 pt] What is the total squared reconstruction error over all points?

9 [2 pt] Show that minimizing total squared reconstruction error is equivalent to minimizing $-\|Xv\|_2^2$.

10 [4 pt] Solve the minimization problem to show that the solution is the first PC. (Hint: take the Lagrangian of the above problem, differentiate and substitute to zero, to get to the optimum solution)

4.1 [3 pts] SVD and PCA

Let us define a new variable Y as

$$Y = X^T \tag{1}$$

where X is a $n \times p$ matrix containing the data points as defined before. If the SVD of Y is given by $Y = U\Sigma V^T$ then show that the columns of V are the PCA of X.