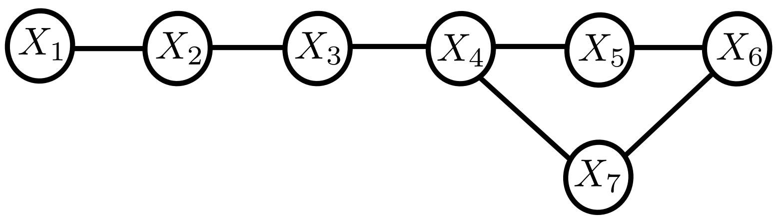
10-704 Information Processing & Learning

Quiz 1 Solutions *March 5th 2015*

Short Questions (50 Points)

Data Processing Inequality. Consider the Markov Random Field (Undirected Graphical Model) shown below.



Which of the statements are always true of the above graphical model?

1. **(5 Points)** $H(X_1|X_2) \leq H(X_1|X_3)$

2. **(5 Points)** $I(X_3; X_4) \geq I(X_3; X_5)$

3. **(5 Points)** $I(X_4; X_5) \geq I(X_4; X_6)$

Huffman Codes.

4. **(5 pts)** Construct a binary Huffman code for the following distribution on 5 symbols $p = (0.4, 0.25, 0.25, 0.05, 0.05)$. What is the expected length of this code ?

5. **(5 pts)** The above code does not meet the entropy lower bound $H(p) \approx 1.96$. Construct a distribution p' for which the above code meets the entropy lower bound $H(p')$.

Complexity Penalized ERM. Suppose you want to fit a Markov Chain (MC) distribution to your data. You don't know the order of the Markov chain to use and would like to set up a complexity penalized ERM approach using prefix codes to automatically do model selection for you. Let $\mathcal{F}_m$ be the class of Markov chain distributions of order m. In class, we saw how to encode any distribution in $\mathcal{F}_m$. You may assume that such a coding scheme is available – you *do not* need to derive it.

6. **(5 Points)** Propose a scheme to encode the *order* of the Markov Chain. You do not need to construct an optimal code for the order. (This is because the length of the code is dominated by the length to encode the Markov Chain given the order.)

7. **(5 Points)** Using your answer in part (a) state how you may construct a prefix code for a Markov chain distribution of any order – i.e. any distribution in the class $\mathcal{F} = \bigcup_{j=1}^{\infty} \mathcal{F}_j$.

8. **(5 Points)** Write down the corresponding complexity penalized ERM optimization problem in $\mathcal{F}$ using the minimum description length principle.

Redundancy and Mixture models. Suppose you would like to minimize redundancy with respect to a class of distributions $\{P_\theta\}_{\theta \in \Theta}$ where $P_\theta = \text{Bernoulli}(\theta)$ and $\Theta = \{0, 1\}$. You decide to use a mixture distribution $1/2P_0 + 1/2P_1$.

9. **(5 Points)** What is the worst-case redundancy $\sup_{P \in P_0, P_1} D(P \| 1/2P_0 + 1/2P_1)$?

10. **(5 Points)** Argue that coding using a mixture distribution $1/2P_0 + 1/2P_1$ is better than coding using either P_1 or P_0.

Solutions

1. True
2. True
3. False
4. A Huffman code is $\{0, 10, 110, 1110, 1111\}$ which has expected length of 2.05 bits.
5. The above code achieves the entropy lower bound on the distribution $p' = \{1/2, 1/4, 1/8, 1/16, 1/16\}$.
6. The order m can be encoded via a string of $m-1$ 0's and a 1 at the end. E.g: $m = 4, \implies$ 0001.
7. Let the order code me c_m and the code of the distribution f in the class $\mathcal{F}_m$ be c_f. We concatenate the two codes : $c_m c_f$.
8. The minimization problem is,
$$\min_{f \in \mathcal{F}} \widehat{R}_n(f) + \sqrt{\frac{c(f) + \log(1/\delta)}{2n}}$$
where $c(f) = m + length(c_f)$.
9. $\log(2)$
10. Since $\sup_{P_0,P_1} D(P\|P_0) = \sup_{P_0,P_1} D(P\|P_1) = \infty$.

Long Questions (50 Points)

1. **(15 pts)** Let $X \sim p$ where p is a distribution on the set $\mathcal{X} = \{1, 2, \ldots, m\}$. We are given a subset $S \subset \mathcal{X}$. We ask whether $X \in S$ and receive the answer Y where $Y = 1$ if $X \in S$ and $Y = 0$ if $X \notin S$. Prove that $I(X;Y) = H(\mathrm{Bern}\,(\mathbb{P}(X \in S)))$. Here $I(X;Y)$ is the mutual information between X and Y and $H(p)$ is the entropy of the distribution p.

   Solution

   Let $\mathbb{P}(X \in S) = \alpha$. Then,

$$\begin{aligned} I(X;Y) = H(X) - H(Y|X) &= H(X) - \left(-\sum_{x,y} p(x,y) \log p(x,y)/p(y)\right) \\ &= H(X) - \left(-\sum_{x \in S} p(x) \log(p(x)/\alpha) - \sum_{x \notin S} p(x) \log(p(x)/(1-\alpha))\right) \\ &= \sum_{x \in S} p(x) \log(1/\alpha) + \sum_{x \notin S} p(x) \log(1/(1-\alpha)) \\ &= \alpha \log(1/\alpha) + (1-\alpha) \log(1/(1-\alpha)) = H(\alpha) \end{aligned}$$

2. Consider three dependent binary random variables (X_1, X_2, X_3) where $X_i \in \{0, 1\}\ \forall i$. You have the following 4 data points from their joint distribution: $(0,0,1)$, $(0,1,1)$, $(1,0,0)$ and $(1,1,1)$.

   (a) **(15 Points)** Obtain the plug-in estimates $\widehat{I}_{12}$, $\widehat{I}_{13}$ and $\widehat{I}_{23}$ for the Mutual Inforations $I(X_1;X_2)$, $I(X_1;X_3)$ and $I(X_2;X_3)$. Recall that $I(X,Y) = \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$.
   You do not need to simplify the expressions completely – you can leave fraction/ log terms.

   (b) **(5 Points)** Depict the Chow-Liu tree for (X_1, X_2, X_3) obtained from the above data.

   Solution

   (a) We have the following empirical marginal probability distributions,

X_1	0	1
	1/2	1/2

X_2	0	1
	1/2	1/2

X_3	0	1
	1/4	3/4

$X_1 \backslash X_2$	0	1
0	1/4	1/4
1	1/4	1/4

$X_1 \backslash X_3$	0	1
0	0	1/2
1	1/4	1/4

$X_2 \backslash X_3$	0	1
0	1/4	1/4
1	0	1/2

Therefore, the plug-in estimates are

$$\widehat{I_{12}} = 4\left(\frac{1}{4}\log\left(\frac{1/4}{1/2\cdot 1/2}\right)\right) = 0$$
$$\widehat{I_{13}} = \frac{1}{4}\log\left(\frac{1/4}{1/2\cdot 1/4}\right) + \frac{1}{4}\log\left(\frac{1/4}{1/2\cdot 3/4}\right) + \frac{1}{2}\log\left(\frac{1/2}{1/2\cdot 3/4}\right) > 0$$
$$\widehat{I_{23}} = \frac{1}{4}\log\left(\frac{1/4}{1/2\cdot 1/4}\right) + \frac{1}{4}\log\left(\frac{1/4}{1/2\cdot 3/4}\right) + \frac{1}{2}\log\left(\frac{1/2}{1/2\cdot 3/4}\right) > 0$$

(b) Since $\widehat{I}_{12} = 0$ and the other two are positive, the Chow-Liu tree would be X_1—X_3—X_2

3. **(15 Points)** We have samples $x_1, \ldots, x_n$ from a discrete distribution π on support $\mathcal{X}$. Let $D = |\mathcal{X}|$. We wish to estimate π using d features $\mathbf{f}$. Here $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$ and $\mathbf{f}(x) = [f_1(x), \ldots, f_d(x)]$. Since $D \gg n$, in order to improve generalization performance we wish to find the MaxEnt distribution subject to an ℓ_∞-ball constraint. I.e. we wish to solve,

$$\begin{aligned} \min_{p\in\Delta^{D-1}} \quad & D(p\|q_0) \\ \text{subject to} \quad & |\mathbb{E}_p[f_i(X)] - \mathbb{E}_n[f_i(X)]| \le \beta \quad \forall i = 1, \ldots, d \end{aligned}$$

where $\Delta^{D-1} = \{x \in \mathbb{R}^D; \mathbf{1}^\top x = 1, x_i \ge 0\}$ and $\mathbb{E}_p$, $\mathbb{E}_n$ denote the expectation w.r.t p and the empirical expectations respectively.

Obtain the dual of this problem as a penalized Maximum Likelihood problem in the Exponential family $E_{\mathbf{f},q_0} = \{q(x) \propto q_0(x)\exp(\lambda^\top \mathbf{f}(x)) \,:\, \lambda \in \mathbb{R}^d\}$. *The dual should be in terms of empirically computable quantities.*
You may use any results we covered in class. You *do not* need to derive the dual from first principles.

Hint: The Fenchel conjugate of $U(u) = \mathbb{1}(\|r - u\|_\infty \le \beta)$ is $U^*(\lambda) = \lambda^\top r + \beta\|\lambda\|_1$.

Solution Note that we can rewrite the problem as,

$$\min_{p\in\Delta^{d-1}} D(p\|q_0) + U(\mathbb{E}_p[\mathbf{f}])$$

where $U(u) = \mathbb{1}(\|\mathbb{E}_n[\mathbf{f}] - u\|_\infty \le \beta)$. Then, the dual can be written as

$$\inf_{\lambda\in\mathbb{R}^d} L_t(\lambda) + U_t^*(\lambda)$$

for any distribution t in Δ^{D-1}. By taking t to be the empirical distribution we get,

$$L_n(\lambda) = -\frac{1}{n}\sum_{i=1}^n \log q_\lambda(x_i)$$
$$U_n^*(\lambda) = U^*(-\lambda) + \lambda^\top \mathbb{E}_n[\mathbf{f}] = \beta\|\lambda\|_1$$

Then, we need to solve

$$\max_{\lambda\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \log q_\lambda(x_i) - \beta\|\lambda\|_1$$