Quiz 1

Date: Monday, October 17, 2016

Name: ______________________________

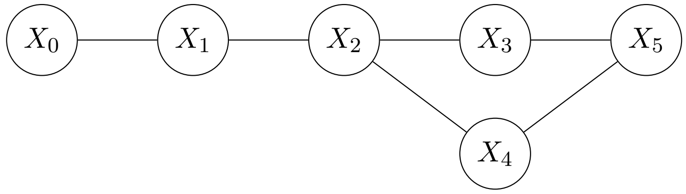
Andrew ID: ______________________________

Department: ______________________________

Guidelines:

1. **PLEASE DO NOT TURN THIS PAGE UNTIL INSTRUCTED.**
2. Write your name, Andrew ID, and department in the spaces provided above.
3. You have **sixty (60)** minutes for this exam.
4. This exam has **seven (7)** pages on seven (7) sheets of paper, including this one.
5. This exam has a total of 50 possible points. The number of points allocated to each question is indicated next to each question.
6. This exam is open notes. You may use any materials such as cheat sheets, class notes, etc. No electronic devices are permitted.
7. The questions vary in difficulty. The points allocated to a question do not entirely reflect its difficulty. Do not spend too much time on one question.
8. Questions only appear on one side of each sheet of paper. You may use any blank space for your answer or scratch work, but please clearly indicate your answers.

1. **[10 points]** Consider the undirected graphical model shown below:



(a) Which of the following statements are always true? No justification is needed.

i. **(2 points)** $H(X_0|X_1) \leq H(X_0|X_2)$

ii. **(2 points)** $I(X_0; X_3) \leq I(X_0; X_5)$

iii. **(2 points)** $I(X_2; X_5|X_4) \leq I(X_2; X_3|X_4)$

(b) **(4 points)** Suppose we observe n IID samples from the joint distribution of $(X_0, \ldots, X_5)$, and use the Chow-Liu algorithm with a consistent mutual information estimator. Explain why we never recover the above graph structure, even as $n \to \infty$.

Solution:

(a) i. True; $H(X_0|X_1) = H(X_0) - I(X_0; X_1) \leq H(X_0) - I(X_0; X_2) = H(X_0|X_2)$.

ii. False.

iii. True; for any fixed value $X_4 = x$, $X_2 - X_3 - X_5$ is a Markov chain, so $I(X_2; X_5|X_4 = x) \leq I(X_2; X_3|X_4 = x)$. Now take expectations over X_4.

(b) The Chow-Liu algorithm only outputs tree-shaped graphical models, and hence it cannot recover the cycle on the right.

2. **[12 points]** Let X_i for $i \in [d] = \{1, \ldots, d\}$ be independent random variables. Show that

(a) **(3 points)** Show that $I(X_i; X_i + X_j) = H(X_i + X_j) - H(X_j)$.

(b) **(4 points)** Show that $I(X_i; X_i + X_j) \geq I(X_i; X_i + X_j + X_k)$.

(c) **(5 points)** Define $f : 2^{[d]} \to \mathbb{R}$ by

$$f(S) := H\left(\sum_{i \in S} X_i\right), \quad \forall S \subseteq [d].$$

Show that f is submodular.

Solution:

(a) Since the distribution $X_i + X_j | X_i$ is a shifted version of the distribution of X_j, $H(X_i + X_j | X_i) = H(X_j)$. Thus,

$$I(X_i, X_i + X_j) = H(X_i + X_j) - H(X_i + X_j | X_i) = H(X_i + X_j) - H(X_j).$$

(b) Since $X_i \to X_i + X_j \to X_i + X_j + X_k$ is a Markov chain, this follows from the data processing inequality.

(c) For any $S \subseteq [d]$, $i, j \in [d]$, by the previous parts,

$$\begin{aligned} f(S \cup \{i,j\}) - f(S \cup \{j\}) &= H\left(X_i + X_j + \sum_{k \in S} X_k\right) - H\left(X_j + \sum_{k \in S} X_k\right) \\ &= I\left(X_i; X_i + X_j + \sum_{k \in S} X_k\right) \\ &\leq I\left(X_i; X_i + \sum_{k \in S} X_k\right) \\ &= H\left(X_i + \sum_{k \in S} X_k\right) - H\left(\sum_{k \in S} X_k\right) = f(S \cup \{i\}) - f(S). \end{aligned}$$

3. **[8 points]** Suppose you flip a coin independently n times and observe cn heads and $(1-c)n$ tails. Explain how to use the MDL principle to choose the best model amongst $M = \cup_\ell M_\ell$, where

$$M_\ell : \text{ The probability the coin lands heads is } z2^{-\ell} \text{ for some integer } z \in [0, 2^\ell).$$

Write the MDL rule in terms of n, c, z and ℓ only.

Solution: We can encode a model in M_ℓ with $\log_2(2^\ell) = \ell$ bits, and use another ℓ bits to encode ℓ itself.[1] Thus, each model can be encoded with 2ℓ bits. Encoding the data takes $cn\log(1/(z2^{-\ell})) + (1-c)n\log(1/(1-z2^{-\ell}))$ bits. Hence, the MDL rule is

$$\arg\min_\ell \; cn\log(1/(z2^{-\ell})) + (1-c)n\log(1/(1-z2^{-\ell})) + 2\ell$$

[1] One can actually encode ℓ with $\log \ell$ bits. Either is acceptable, since this term is asmptotically negligible.

4. **[10 points]** Suppose we already have an estimate $\widehat{p}$ for some probability density p on $\mathcal{X}$. Using n new IID samples $X_1, \ldots, X_n \sim p$, we want to estimate the squared L_2-norm

$$\|p\|_2^2 = \int_{\mathcal{X}} p^2(x)\,dx = \mathbb{E}_{X\sim p}\left[p(X)\right].$$

of p. Show that the first-order von Mises estimator is identical to the re-substitution estimator:

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{p}(x),$$

Solution: Unfortunately the problem, as stated, is incorrect. The correct von Mises estimator is derived as follows, based on the first-order von Mises expansion of the squared L_2-norm:

$$\begin{aligned}
\|p\|_2^2 &= \|\widehat{p}\|_2^2 + \langle \nabla_p \|\widehat{p}\|_2^2, p - \widehat{p}\rangle + O\left(\|p - \widehat{p}\|_2^2\right) \\
&= \|\widehat{p}\|_2^2 + 2\langle \widehat{p}, p - \widehat{p}\rangle + O\left(\|p - \widehat{p}\|_2^2\right) \\
&\approx \|\widehat{p}\|_2^2 + 2\langle \widehat{p}, p - \widehat{p}\rangle \\
&= \|\widehat{p}\|_2^2 + \mathop{\mathbb{E}}_{p}\left[\widehat{p}\right] - \|\widehat{p}\|_2^2 = 2 \mathop{\mathbb{E}}_{X\sim p}\left[\widehat{p}(X)\right] - \|\widehat{p}\|_2^2.
\end{aligned}$$

The first term can be replaced by an empirical expectation, while the second is directly (perhaps approximately) computable from $\widehat{p}$. Thus, the first-order von Mises estimator for the L_2-norm is

$$\frac{2}{n}\sum_{i=1}^{n} \widehat{p}(X_i) - \int_{\mathcal{X}} \widehat{p}^2(x)\,dx.$$

Note that, for some standard estimators $\widehat{p}$, such as orthogonal series estimators with an appropriate number of basis elements, the above estimator has the *same convergence rate* as the resubstitution estimator; the difference $\mathbb{E}_{X\sim p}\left[\widehat{p}(X)\right] - \|\widehat{p}\|_2^2$ is negligibly small.

5. **[10 points]** Consider a set of k variables $X_1, \ldots, X_k$, and suppose we know the pairwise distributions $p(X_i, X_{i+1})$, for $i \in \{1, \ldots, k-1\}$, of consecutive pairs. Show that the MaxEnt joint distribution $p(X_1, \ldots, X_k)$ is a first-order Markov chain (i.e., for any $i_1 < i_2 < i_3$ in $\{1, \ldots, k\}$, X_{i_1} and X_{i_3} are conditionally independent given X_{i_2}). *(Hint: Write the joint distribution $H(X_1, \ldots, X_k)$ in terms of $\sum_{i=1}^k H(X_i|X_{i-1})$+another term.)*

   Solution: By the chain rule,

$$\begin{aligned} H(X_1, \ldots, X_k) &= \sum_{i=1}^{k} H(X_i|X_1, \ldots, X_{i-1}) \\ &= \sum_{i=1}^{k} H(X_i|X_{i-1}) - I(X_i; X_1, \ldots, X_{i-2}|X_{i-1}). \end{aligned}$$

   Since mutual information is non-negative, this is clearly maximized when

$$I(X_i; X_1, \ldots, X_{i-2}|X_{i-1}) = 0, \quad \forall i \in \{2, \ldots, k\}. \tag{1}$$

   This occurs precisely when each X_i is conditionally independent of $(X_1, \ldots, X_{i-2})$ given X_{i-1}. Thus, in the undirected graphical model of the MaxEnt distribution, for $i_1 < i_3$, every path from X_{i_1} to X_{i_3} goes through X_{i_3-1}, and it follows by induction that every path from X_{i_1} to X_{i_3} goes through X_{i_2}, for all $i_1 < i_2 < i_3$.

   Finally, note that (1) is achievable, for instance, by the process that draws X_1 from the marginal of $p(X_1, X_2)$, and then, for each $i \in \{1, \ldots, k-1\}$, recursively draws X_{i+1} from the conditional $p(X_{i+1}|X_i)$.

6. **[Optional - no credit]** If you found the quiz too easy, prove the following for problem 1. Assume all the edge weights are distinct. Argue that, as $n \to \infty$, we always recover the edge $X_0 - X_1$. *(Hint: Argue by means of contradiction.)*

   Solution: As $n \to \infty$, we almost surely (with probability 1) estimate the edge weights exactly. Let T denote the Chow-Liu tree, and suppose, for sake for contradiction, that T does not contain $X_0 - X_1$. Let X_i denote any neighbor of X_0. Construct a new graph T' by removing all edges adjacent to X_0, re-attaching all but $X_0 - X_i$ to X_i, and adding $X_0 - X_1$. Since the original graph was a tree, the new graph is still a tree. By the data processing inequality, and since the edge weights are distinct, the total weight of T is strictly less than the total weight of T'. This contradicts the fact that the Chow-Liu algorithm chooses the maximum weight spanning tree.

Please do not mark below this line.

Problem	Max	Points
Q1	10	
Q2	12	
Q3	8	
Q4	10	
Q5	10	
Total	50	