Stanford University
Probabilistic Graphical Models

Daphne Koller, Kevin Murphy
Winter 2011-2012

Feedback — Structure Scores

You achieved a score of 6.00 out of 6.00

Question 1

***Score Consistency.** Assume that the dataset D has m examples, each drawn independently from the distribution P^*, for which the graph G^* is a perfect map. What do we mean when we say that the BIC score $\text{Score}_{BIC}(G:D)$, measured with respect to D, is **consistent**?

(Note that more than one statement below might hold true for the BIC score. We're not looking at which statements are true; instead, we're looking for the statement that is **equivalent**, i.e., necessary and sufficient, for score consistency.)

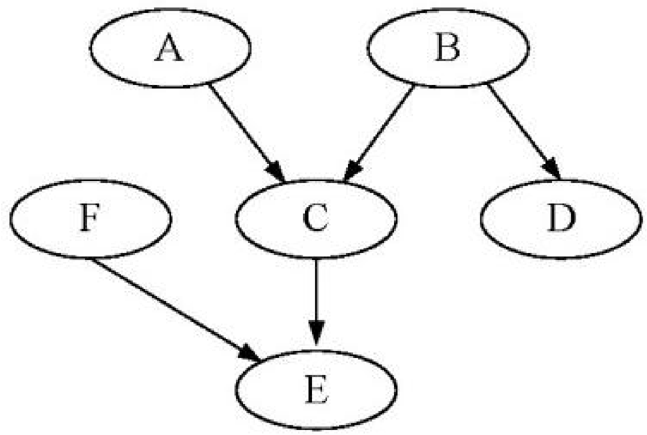
Your Answer	Score	Explanation
As $m \to \infty$, with probability 1 we will draw a dataset D from P^* such that the inequality $\text{Score}_{BIC}(G^*:D) > \text{Score}_{BIC}(G:D)$ holds for all other graphs G which are not I-equivalent to G^*.	1.00	It is still possible, though extremely unlikely for large m, to draw a pathological dataset D that does not reflect the underlying distribution P^* at all. Hence, any statement we make about consistency can only be about the probability of obtaining a dataset D for which the desired consistency conditions hold. Now, remember that consistent scoring metrics like the BIC score allow us to uniquely identify the correct I-

equivalence class (in this case, that of G^*) as the number of data points grows large. For this, we need strict inequality in the BIC scores of G^* versus other graphs not in its I-equivalence class.

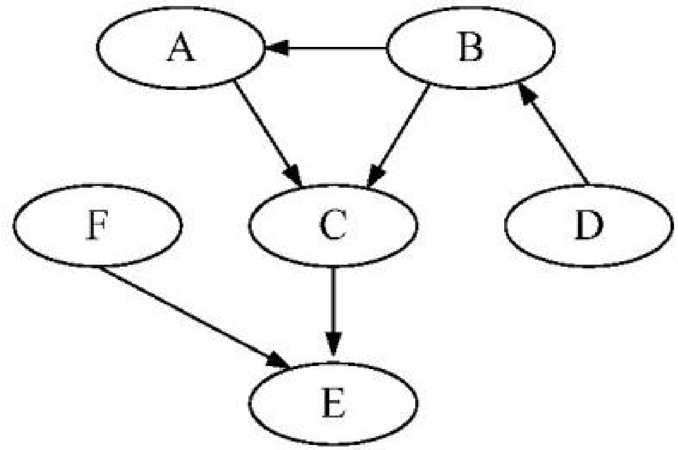
Total	1.00 / 1.00

Question 2

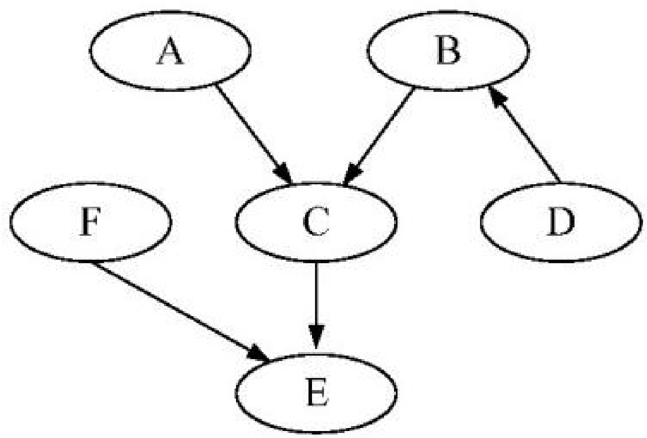
Likelihood Scores. Consider the following 4 graphs:



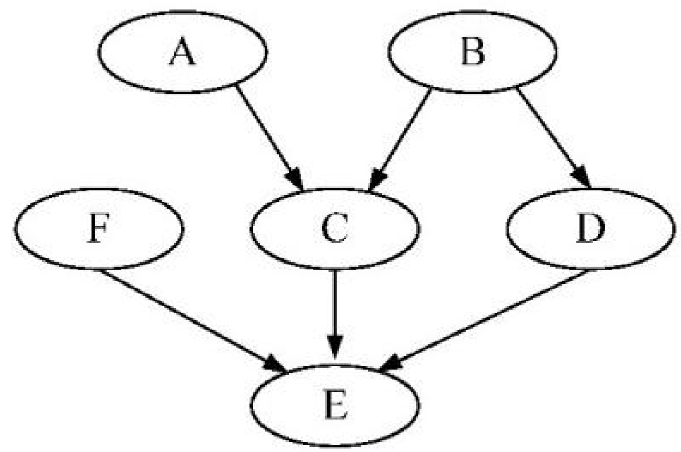
G1



G2



G3



G4

Which of the following statements about the likelihood scores of the different graphs is/are true? You may choose more than 1 option (or none, if you think none are true).

Your Answer	Score	Explanation

✔ $\text{Score}_L(G4 : D) \geq \text{Score}_L(G3 : D)$ for every dataset D	✔	0.20	$G4$ is an I-map of $G3$, that is, every independence relation that is in $G4$ is also in $G3$. Hence, $G4$ can represent all distributions that $G3$ can, and so its likelihood score will not be lower than that of $G3$.
✔ $\text{Score}_L(G2 : D) \geq \text{Score}_L(G3 : D)$ for every dataset D	✔	0.20	$G2$ is an I-map of $G3$, that is, every independence relation that is in $G2$ is also in $G3$. Hence, $G2$ can represent all distributions that $G3$ can, and so its likelihood score will not be lower than that of $G3$.
$\text{Score}_L(G1 : D) \geq \text{Score}_L(G4 : D)$ for every dataset D	✔	0.20	$G4$ is an I-map of $G1$, that is, every independence relation that is in $G4$ is also in $G1$. Hence, $G4$ can represent all distributions that $G1$ can, and so its likelihood score will not be lower than that of $G1$.
$\text{Score}_L(G2 : D) \geq \text{Score}_L(G4 : D)$ for every dataset D	✔	0.20	$G2$ and $G4$ encode different sets of independence relations, neither of which is contained in the other. For example, $G2$ satisfies $D \perp E \mid B$, while $G4$ satisfies $A \perp B$. Hence, there will be some datasets for which $\text{Score}_L(G2 : D)$ is larger, and others for which $\text{Score}_L(G4 : D)$ is larger.
None of the above	✔	0.20	
Total		1.00 / 1.00	

Question 3

BIC Scores. Consider the same 4 graphs as in the previous question, but now think about the BIC score.

Which of the following statements is/are true?

Your Answer		Score	Explanation
☐ $\text{Score}_{BIC}(G4 : D) \geq \text{Score}_{BIC}(G1 : D)$ for every dataset D	✔	0.20	While the likelihood score of $G4$ is always greater than or equals to that of $G1$, $G4$ is also a more complex graph. As the BIC score is essentially the likelihood score minus a penalty for more complex models, these two components are in opposition. For large datasets in which $G4$ is a much better model than $G1$, the likelihood score could dominate; conversely, for small datasets or for datasets that are generated by a distribution that $G1$ can encode, the BIC score would be in $G1$'s favor.
☐ $\text{Score}_{BIC}(G2 : D) \neq \text{Score}_{BIC}(G3 : D)$ for every dataset D	✔	0.20	While the likelihood score of $G2$ is always greater than or equals to that of $G3$, $G2$ is also a more complex graph. As the BIC score is essentially the likelihood score minus a penalty for more complex models, these two components are in opposition. In general, there can exist datasets for which the advantage of $G2$ in terms of the likelihood score is exactly canceled by the penalty it incurs for being more complex.
☑ $\text{Score}_{BIC}(G1 : D) = \text{Score}_{BIC}(G3 : D)$ for every dataset D	✔	0.20	I-equivalent graphs have the same likelihood score, and have the same complexity (in terms of the number of independent parameters). Hence, they have the same BIC score.
☐ $\text{Score}_{BIC}(G1 : D) \neq \text{Score}_{BIC}(G3 : D)$	✔	0.20	I-equivalent graphs have the same likelihood score and the same

for every dataset D			complexity (in terms of the number of independent parameters). Hence, they have the same BIC score.
None of the above	✓	0.20	
Total		1.00 / 1.00	

Question 4

Likelihood Guarantees. Consider graphs $G2$ and $G3$. We have a dataset D generated from some probability distribution P, and the likelihood scores for $G2$ and $G3$ are $\mathrm{Score}_L(G2 : D)$ and $\mathrm{Score}_L(G3 : D)$, respectively. Let $\theta^*_{D,2}$ and $\theta^*_{D,3}$ be the maximum likelihood parameters for each network, taken with respect to the dataset D. Now let $L(X : G, \theta)$ represent the likelihood of dataset X given the graph G and parameters θ, so $\mathrm{Score}_L(G2 : D) = L(D : G2, \theta^*_{D,2})$ and $\mathrm{Score}_L(G3 : D) = L(D : G3, \theta^*_{D,3})$.

Suppose that $L(D : G2, \theta^*_{D,2}) > L(D : G3, \theta^*_{D,3})$. If we draw a new dataset E from the distribution P, which of the following statements can we guarantee? If more than one statement holds, choose the more general statement.

Your Answer		Score	Explanation
None of the above	✓	1.00	$\theta^*_{D,2}$ and $\theta^*_{D,3}$ correspond to the ML estimation from dataset D. Given the new dataset E, they might not be the ML estimation parameters any longer. Since the dataset D and E might not be sufficiently large enough to accurately characterize P, there is no guarantee on the relation of likelihood scores of the new dataset.
Total		1.00 / 1.00	

Question 5

Hidden Variables. Consider the case where the generating distribution has a naive Bayes structure, with an unobserved class variable C and its binary-valued children $X_1, \ldots, X_{100}$. Assume that C is strongly correlated with each of its children (that is, distinct classes are associated with fairly different distributions over each X_i). Now suppose we try to learn a network structure directly on $X_1, \ldots, X_{100}$, **without including C in the network**. What network structure are we likely to learn if we have 10,000 data instances, and we are using table CPDs with the **likelihood** score as the structure learning criterion?

Your Answer		Score	Explanation
A fully connected network, i.e., one with an edge between every pair of nodes.	✔	1.00	In the generating distribution, for any pair of variables X_i, X_j, the trail $X_i \leftarrow C \rightarrow X_j$ is active. Thus, there are no independence relations of the form $X_i \perp X_j$. This means that when we try to use the likelihood score to learn a network structure over only the X_i's, we will end up with a fully connected network.
Total		1.00 / 1.00	

Question 6

Hidden Variables. Now suppose that we use the BIC score instead of the likelihood score in the previous question. What network structure are we likely to learn with the same 10,000 data instances?

Your Answer		Score	Explanation
Some connected network over	✔	1.00	Even though a fully connected network may be the best representation for the true underlying distribution, we don't

$X_1, \ldots, X_{100}$ that is not fully connected nor empty.		have enough data to learn it, and the BIC structure penalty will not allow the learning of a network with such high complexity, given only 10,000 instances.
Total	1.00 / 1.00	