Learning: Final Exam

12 试题

1 point

1。
***Multiplexer CPDs.** What is the form of the independence that is implied by the multiplexer CPD and that we used in our derivation of the posterior over the parameters of the simple Bayesian Network $x \rightarrow y$? (i.e. the factorization of $P(\theta_{\mathbf{x}}, \theta_{\mathbf{Y}|\mathbf{x}^1}, \theta_{\mathbf{Y}|\mathbf{x}^0} \mid \mathbf{D})$). Recall that a CPD is defined as a multiplexer if it has the structure $P(Y|A, Z_1, \ldots, Z_k) = \mathbf{I}\{Y = Z_a\}$ where the values of A are the natural numbers 1 through k . **Also note that the answer is specific to multiplexer CPDs and is not implied by the graph structure alone.**

- $\theta_{\mathbf{Y}|\mathbf{X}} \perp \theta_{\mathbf{X}}$
- $\theta_{\mathbf{Y}|\mathbf{x}^0} \perp \theta_{\mathbf{Y}|\mathbf{x}^1} \mid \mathbf{X}, \mathbf{Y}$
- $\theta_{\mathbf{Y}|\mathbf{x}^0} \perp \theta_{\mathbf{Y}|\mathbf{x}^1} \mid \mathbf{X}$
- $\theta_{\mathbf{Y}|\mathbf{x}^0}, \theta_{\mathbf{Y}|\mathbf{x}^1} \perp \mathbf{X}$
- $\theta_{\mathbf{Y}|\mathbf{X}} \perp \mathbf{X} \mid \theta_{\mathbf{X}}$

1 point

2。
***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**?

Hint: We are looking for a definition that will always be true, not just probably be true.

- As $m \rightarrow \infty$, no matter which examples were drawn from P^* into the dataset D , the inequality

  $$\text{Score}_{BIC}(G^* : D) > \text{Score}_{BIC}(G : D)$$

  will always be true for all other graphs $G \neq G^*$.
- As $m \rightarrow \infty$, 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 .

- As $m \to \infty$, no matter which examples were drawn from P^* into the dataset D, the inequality

  $$\text{Score}_{BIC}(G^* : D) > \text{Score}_{BIC}(G : D)$$

  will always be true for all graphs G which are not I-equivalent to G^*.

- 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 $G \neq G^*$.

---

1 point

3。

EM and Convergence.

When checking for the convergence of the EM algorithm, we can choose to measure changes in either the log-likelihood function or in the parameters.

For a generic application, we typically prefer to check for convergence using the log-likelihood function.

However, this is not always the case, especially when the values of the parameters are important in and of themselves.

In which situations would we also be concerned about reaching convergence in terms of the parameters? Do not worry about the implementation details in the following models.

- [ ] We are trying to transcribe human speech by building a Hidden Markov Model (HMM) and learning its parameters with the EM algorithm. The end-goal is correctly transcribing raw audio input into words.

- [x] We are building a graphical model for medical diagnosis, where nodes can represent symptoms, diseases, predisposing factors, and so on. This system will not be deployed in the clinic; our only aim is to understand how various predisposing factors can interact with each other in increasing disease risk.

- [x] We are trying to better understand high-energy physics by using a graphical model to analyze time-series data from particle accelerators. The hope is to elucidate the types of interactions between different particle types.

- [ ] We have a graphical model in which each node is a superpixel, and we are using EM to learn the parameters that specify the relations between superpixels.

  Our end-goal is to build an image segmentation pipeline that is highly accurate.

- [ ] We are building a graphical model for medical diagnosis, where nodes can represent symptoms, diseases, predisposing factors, and so on. Our only aim is to maximize our chances of correctly predicting diseases that patients are suffering from.

- [x] We are building a graphical model to represent a biological network, where each node corresponds to a gene. We want to learn the interactions between genes by finding the parameters that maximize the likelihood of a given training dataset of gene expression measurements. The interactions we find will then be further studied by biologists.

- [ ] We have a graphical model in which each node represents an object part, and we are using EM to learn the parameters that specify the relations between object parts.

  Our end-goal is to build an image classification system that can accurately recognize the image as one of several known objects.

---

1 point

4.
Parameter Estimation with Missing Data.

The process of learning Bayesian Network parameters with missing data (partially observed instances) is more difficult than learning with complete data for which of the following reasons? You may select one or more options.

- [ ] We require more training data, because we must throw out all incomplete instances.

- [ ] Because there can be multiple optimal values, we must always run our learning algorithm multiple times from different initializations to make sure we find ALL of them.

- [x] We lose local decomposition, whereby each CPD can be estimated independently.

- [ ] While there is still always a single optimal value for the parameters, it can only be found using an iterative method.

---

1 point

5.
Optimality of Hill Climbing. Jack and Jill come up to you one day with a worried look on their face. "All this while we've been climbing hills, trying to improve upon our graph structure," they say. "We've been considering edge deletions, reversals, and additions at each step. Today, we found that no single edge deletion, reversal, or addition could give us a higher-scoring structure. Are we guaranteed that our current graph is the best graph structure?"

What should you tell them? You may assume that their dataset is sufficiently large, and that your answer should hold for a general graph.

- ( ) No - greedy hill-climbing will only find the true graph structure if we restrict the number of parents for each node to at most 2.

- ( ) Yes, but only if we extend our range of available moves to allow for pairs of edges to be changed simultaneously.

- ( ) No - greedy hill-climbing will find only local maxima of the scoring function with respect to our available moves. While it might find the true graph structure on occasion, we cannot guarantee this.
- ( ) Yes - greedy hill-climbing provably finds the true graph structure, provided our dataset is large enough.
- ( ) Yes, but only if we use random restarts and tabu search.
- ( ) No - greedy hill-climbing can never find the true graph structure, only local maxima of the scoring function with respect to our available moves.

---

1 point

6。
***Latent Variable Cardinality.**

Assume that we are doing Bayesian clustering, and want to select the cardinality of the hidden class variable.

Which of these methods can we use?

Assume that the structure of the graph has already been fixed.

You may choose more than one option.

- [x] Training several models, each with a different cardinality for that hidden variable.

  For each model, we choose the (table CPD) parameters that maximize the likelihood on the **training set**.

  We then pick the model with the highest likelihood on a **held-out validation set**.

- [ ] Training several models, each with a different cardinality for that hidden variable.

  For each model, we choose the (table CPD) parameters that maximize the likelihood on the **training set**.

  We then pick the model that performs the best on some external evaluation task, using a **held-out validation set**.

  For example, say we are using Bayesian clustering to classify customers visiting an online store, with the aim of giving class-specific product recommendations.

  We could run each model in an alpha-beta testing framework (where different customers may see the result of different models), and measure the percentage of customers that end up purchasing what each model recommends.

- [ ] Training several models, each with a different cardinality for that hidden variable.

  For each model, we choose the (table CPD) parameters that maximize the likelihood on the **training set**.

  We then pick the model with the highest **training set** likelihood.

- [x]

- [ ] Training several models, each with a different cardinality for that hidden variable.

  For each model, we choose the (table CPD) parameters that maximize the likelihood on the **training set**.

  We then pick the model with the highest **test set** likelihood.

- [ ] If we have relevant prior knowledge, we can simply use this to set the cardinality by hand.

---

1 point

7。

EM Stopping Criterion.

When learning the parameters $\theta \in \mathbf{R}^n$ of a graphical model using the EM algorithm, an important design decision is choosing when to stop training.

Let $\ell_{\text{Train}}(\theta)$, $\ell_{\text{Valid}}(\theta)$, and $\ell_{\text{Test}}(\theta)$

be the log-likelihood of the parameters θ on the training set, a held-out validation set, and the test set, respectively.

Let θ^t be the parameters at the t -th iteration of the EM algorithm.

We can denote the change in the dataset log-likelihoods at each iteration with $\Delta\ell^t_{\text{Train}} = \ell_{\text{Train}}(\theta^t) - \ell_{\text{Train}}(\theta^{t-1})$

and the corresponding analogues for the validation set and the test set. Likewise, let $\Delta\theta^t = \theta^t - \theta^{t-1}$ be the vector of changes in the parameters at time step t .

Which of the following would be reasonable conditions for stopping training at iteration t ? You may choose more than one option.

- [ ] $\Delta\ell_{\text{Test}}$ becomes negative.
- [x] $\|\Delta\theta^t\|_2^2$ becomes small, i.e., it falls below a certain tolerance $\epsilon > 0$.

  Note: The ℓ_2 norm, also known as the Euclidean norm, is defined for any vector $x \in \mathbf{R}^n$ as

  $\|x\|_2^2 = \sum_{i=1}^n x_i^2$.
- [x] $\Delta\ell^t_{\text{Valid}}$ becomes negative.
- [ ] $\Delta\ell_{\text{Test}}$ becomes small, i.e., it falls below a certain tolerance $\epsilon > 0$

---

1 point

8。

EM Parameter Selection.

Once again, we are using EM to estimate parameters of a graphical model. We use n random starting points $\{\theta_i^0\}_{i=1,2,\ldots,n}$, and run

EM to convergence from each of them to obtain a set of candidate parameters $\{\theta_i\}_{i=1,2,\ldots,n}$. We wish to select one of these candidate parameters for

use. As in the previous question, let $\ell_{\text{Train}}(\theta)$, $\ell_{\text{Valid}}(\theta)$, and $\ell_{\text{Test}}(\theta)$

be the log-likelihood of the parameters θ on the training set, a held-out validation set, and the test set, respectively.

Which of the following methods of selecting final parameters θ would be a reasonable choice? You may pick more than one option.

- [x] Pick $\theta = \text{argmax}_{i=1,2,\ldots,n} \ell_{\text{Valid}}(\theta_i)$.
- [x] Pick $\theta = \text{argmax}_{i=1,2,\ldots,n} \ell_{\text{Train}}(\theta_i)$.
- [ ] Pick $\theta = \text{argmax}_{i=1,2,\ldots,n} \ell_{\text{Test}}(\theta_i)$.
- [ ] Any one; the θ_i are all equivalent, since all of them are local maxima of the log-likelihood function.

---

1 point

9.

Greedy Hill-Climbing.

Your friend is performing greedy hill-climbing structure search over a network with three variables using three possible operations and the BIC score with dataset $\mathcal{D}$:

- Add an edge
- Delete an edge
- Reverse an edge

She tells you that after examining $\mathcal{D}$, she took a single step and got the following graph:

She also tells you that for the next step she has determined that there is a **unique** optimal greedy operation o to take. Which of the following steps could o be?

Hint: The fact that it is unique eliminates some possibilities for o.

- [ ] Add edge $B \to C$
- [ ] Reverse edge $A \to B$
- [x] Add edge $A \to C$
- [ ] Add edge $C \to A$
- [ ]

- [ ] Delete edge $A \rightarrow B$
- [x] Add edge $C \rightarrow B$

---

1 point

10.
Graph Structure Search.

Consider performing graph structure search using a decomposable score. Suppose our current candidate is graph G below.

We want to compute the changes of scores associated with applying three different operations:

- Delete the edge $A \rightarrow D$
- Reverse the edge $C \rightarrow E$
- Add the edge $F \rightarrow E$

Let $\delta(G : o_1), \delta(G : o_2), \delta(G : o_3)$ denote the score changes associated with each of these three operations, respectively. Which of the following equations is/are true for all datasets $\mathcal{D}$?

- [x] $\delta(G : o_3) = \text{FamScore}(C, \{A, E\} : \mathcal{D})$
- [ ] $$\begin{aligned}\delta(G : o_2) = \text{FamScore}(C, \{A, E\} : \mathcal{D}) \ &+ \text{FamScore}(E, \emptyset : \mathcal{D}) \\ &- \text{FamScore}(C, \{A\} : \mathcal{D}) \ - \ \text{FamScore}(E, \{C\} : \mathcal{D})\end{aligned}$$
- [x] $\delta(G : o_1) = \text{FamScore}(D, \{B, C\} : \mathcal{D})$

- [ ] $\delta(G : o_3) = \text{FamScore}(E, \{C, F\} : \mathcal{D}) \; - \; \text{FamScore}(E, \{C\} : \mathcal{D})$
- [ ] $\delta(G : o_2) = \text{FamScore}(C, \{A, E\} : \mathcal{D}) \; - \; \text{FamScore}(C, \{A\} : \mathcal{D}) \; - \; \text{FamScore}(E, \{C\} : \mathcal{D})$
- [ ] $\delta(G : o_1) = \text{FamScore}(D, \{B, C\} : \mathcal{D}) \; - \; \text{FamScore}(D, \{A, B, C\} : \mathcal{D})$

---

1 point

11。

Structure Learning with Incomplete Data.

After implementing the pose clustering algorithm in PA9, your friend tries to pick the number of pose clusters K for her data by running EM and evaluating the log-likelihood of her data for different values of K . What happens to her log-likelihood as she varies K ?

- (●) The log-likelihood (almost) always increases as K increases.
- ( ) The log-likelihood remains the same regardless of K .
- ( ) The log-likelihood (almost) always decreases as K increases.
- ( ) Impossible to say - depends on the data and on what K is.

---

1 point

12。

Calculating Likelihood Differences. While doing a hill-climbing search, you run into the following two graphs, and need to choose between them using the likelihood score.

What is the difference in likelihood scores, $\text{score}_L(G_1 : \mathbf{D}) - \text{score}_L(G_2 : \mathbf{D})$, given a dataset $\mathbf{D}$ of size M ?

Give your answer in terms of the entropy H and mutual information I . The subscripts below denote empirical values according to $\mathbf{D}$: for example, $H_{\mathbf{D}}(X)$ is the empirical entropy of the variable X in the dataset $\mathbf{D}$.

- (●) $M \times [I_{\mathbf{D}}(D; C) + I_{\mathbf{D}}(E; D) - I_{\mathbf{D}}(B; A) - I_{\mathbf{D}}(D; C, E)]$

- $M \times [I_{\mathbf{D}}(A;B) - H_{\mathbf{D}}(A,B)]$
- $M \times [I_{\mathbf{D}}(C;A,B) + I_{\mathbf{D}}(D;C) + I_{\mathbf{D}}(E;D) - I_{\mathbf{D}}(B;A) - I_{\mathbf{D}}(D;C,E)]$
- $M \times I_{\mathbf{D}}(A;B)$
- $M \times [I_{\mathbf{D}}(D;C) + I_{\mathbf{D}}(E;D) - I_{\mathbf{D}}(A;B) - I_{\mathbf{D}}(D;C,E) - H_{\mathbf{D}}(A,B,C,D,E)]$

☑ 我（**伟臣 沈**）了解提交不是我自己完成的作业 将永远不会通过此课程或导致我的 Coursera 帐号被关闭。 了解荣誉准则的更多信息

提交测试
