



Stanford University

Probabilistic Graphical Models

Daphne Koller, Kevin Murphy
Winter 2011-2012

[Home](#)[Review Quizzes](#)[Theory Problems](#)[Assignments](#)[Assignment Questions](#)[Video Lectures](#)[Discussion Forums](#)[Course Wiki](#)[Lecture Slides](#)[Course Schedule](#)[Course Logistics](#)[Course Information](#)[Course Staff](#)[Octave Installation](#)

Feedback — Learning with Incomplete Data

You achieved a score of 4.75 out of 6.00

Question 1

Missing At Random. Suppose we are conducting a survey of job offers and salaries for Stanford graduates. We already have the major of each of these students recorded, so in the survey form, each graduating student is only asked to list up to two job offers and salaries he/she received. Which of the following scenarios is/are missing at random (MAR)?

| Your Answer | Score | Explanation |
|---|--|--|
| <input checked="" type="checkbox"/> The person recording the information didn't care about humanities students and neglected to record their salaries. | <input checked="" type="checkbox"/> 0.25 | We say data is MAR if whether the data is missing is independent of the missing values themselves given the observed values. Whether the data is missing was determined by major, which is observed. |
| <input type="checkbox"/> The database software ignored salaries of 0 submitted by students who were taking unpaid positions with community service organizations. | <input checked="" type="checkbox"/> 0.25 | This is not MAR because whether the data is missing depends on the missing value (the salary). |
| <input type="checkbox"/> Students who accepted a low-salaried job offer tended not to reveal it. | <input checked="" type="checkbox"/> 0.25 | This is not MAR because whether the data is missing depends on the missing value (the salary). |

✓ CS students get more offers than other majors and found selecting only two too stressful, so they neglected to list any.

✓ 0.25

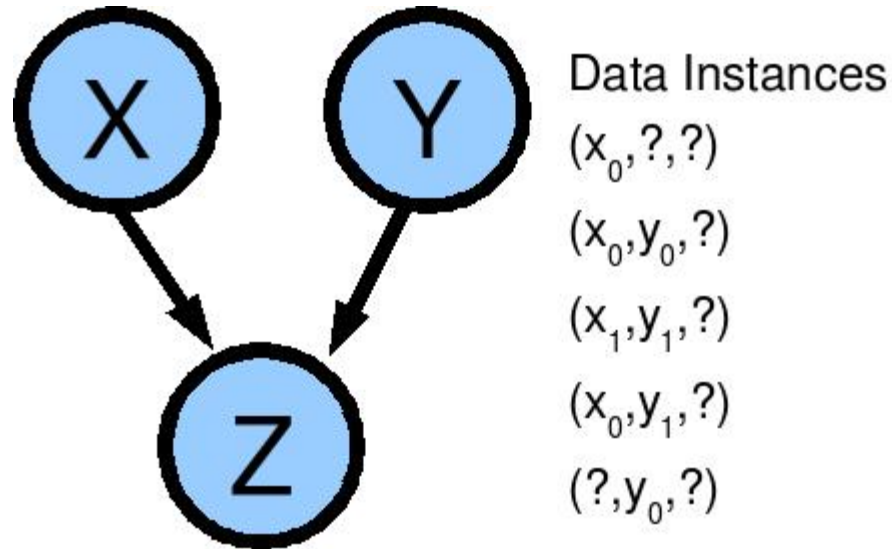
We say data is MAR if whether the data is missing is independent of the missing values themselves given the observed values. This is MAR because whether the data is missing depends on the major (cs) which is observed.

Total

1.00 /
1.00

Question 2

Computing Sufficient Statistics. Given the network and data instances shown below, how do we compute the expected sufficient statistics for a particular value of the parameters?




Your Answer

$$\bar{M}[x_0, y_0, z_0] = P(y_0, z_0 \mid x_0, \theta) + P(z_0 \mid x_0, y_0, \theta) + P(z_0 \mid x_1, y_1, \theta) + P(z_0 \mid x_0, y_1, \theta)$$

Total

Question 3

Likelihood of Observed Data. In a Bayesian Network with partially observed training data, computing the likelihood of observed data for a given set of parameters...

| Your Answer | Score | Explanation |
|--|---|--|
| <input checked="" type="radio"/> requires probabilistic inference, while it DOES NOT in the case of fully observed data. |  1.00 | With missing data, inference is required to complete the expected sufficient statistics (ESS) for the expected likelihood function. Thus, inference is not needed to compute the ESS in the case of fully observed data. |
| Total | 1.00 / 1.00 | |

Question 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, or none if you think none apply.

| Your Answer | Score | Explanation |
|-------------|-------|-------------|
|-------------|-------|-------------|

| | | | |
|---|---|-------------|--|
| <input type="checkbox"/> None of these reasons. | ✓ | 0.25 | At least one of the other options is correct. |
| <input checked="" type="checkbox"/> We lose local decomposition, whereby each CPD can be estimated independently. | ✓ | 0.25 | When all values are observed, we can ignore the values of nodes that are not directly connected, but when there are unobserved values, the CPD of parent nodes can affect the optimal parameters in child nodes. |
| <input checked="" type="checkbox"/> 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. | ✗ | 0.00 | While there may be cases where we want to run from different initializations in order to find a "good" set of parameters, it is generally not our goal to find all optima nor can we ensure that our parameters are globally optimal. |
| <input type="checkbox"/> While there is still always a single optimal value for the parameters, it can only be found using an iterative method. | ✓ | 0.25 | There can be more than a single optimal value for the parameters. |
| Total | | 0.75 / 1.00 | |

Question 5

***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 (or none, if you think none apply).

| Your Answer | Score | Explanation |
|--|--------|---|
| <input checked="" type="checkbox"/> 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 | ✓ 0.25 | If the purpose of the model is to eventually perform this |

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.



external task, then measuring its performance on the true task is arguably more effective than simply measuring data likelihood. Since the external dataset was previously unseen, we will not run into the problem of always picking the model with the highest cardinality.

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



0.25

In some cases (e.g., we are modeling clothing preference in a population, and we introduce a hidden variable that we hope will pick up the differences between genders), we have sufficient prior knowledge to choose a reasonable value for the cardinality of our hidden variable.





| | | |
|--|--|--|
| <input checked="" type="checkbox"/> 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 . |  0.25 | This is the closest we can get to measuring the performance of each model on the test set (in the sense of maximizing data likelihood), without actually using the test set. |
| <input type="checkbox"/> 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. |  0.25 | Training set likelihood will always increase with the cardinality of each hidden variable, so we will always end selecting the model which has the largest cardinality. |
| Total | 1.00 / 1.00 | |

Question 6

PGM with latent variables. Adding hidden variables to a model can significantly increase the expressiveness of a model. However, there are also some issues that arise when we try to add hidden variables.

For which of these problems can we learn a reasonable model by simply choosing the parameters that

maximize training likelihood? Assume that all variables, hidden or (partially) observed, are discrete and follow a table CPD. You may choose more than one option (or none, if you think none apply).

| Your Answer | Score | Explanation |
|--|--|--|
| <input checked="" type="checkbox"/> Given a fixed set of edges, learning the parameters in the table CPDs of each hidden node. |  0.25 | This is a standard parameter estimation with missing data problem that we can solve with methods such as the EM algorithm. While using only training likelihood runs the risk of overfitting, given a large enough training set, the parameters found are still likely to perform reasonably well. |
| <input checked="" type="checkbox"/> Given a fixed set of edges, learning the parameters in the table CPDs of observed nodes that have hidden nodes as parents. |  0.25 | This is a standard parameter estimation with missing data problem that we can solve with methods such as the EM algorithm. While using only training likelihood runs the risk of overfitting, given a large enough training set, the parameters found are still likely to perform reasonably well. |
| <input type="checkbox"/> Choosing the number of values (cardinality) that each hidden variable can take on. |  0.25 | Training set likelihood will always increase with the cardinality of each hidden variable, so we will end up with hidden variables with infinite cardinality. This is because hidden variables with a larger cardinality are always more expressive than hidden variables with a smaller cardinality. |
| <input type="checkbox"/> Choosing the number of hidden variables to add to the graphical model. |  0.25 | Training set likelihood will always increase with the number of hidden variables we add, so we will end up with infinitely many hidden variables. This is because for any two integers $N > n$, the set of graphs with N hidden nodes is more expressive; it can represent whatever probability distributions the set of graphs with n hidden nodes can, and possibly more. |
| Total | 1.00 / 1.00 | |

