Learning with Incomplete Data

返回第 #4 周课程

4/4 得分（100%）

1 / 1 分

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)?

- [x] One of the students entered her name as "ROBERTA'); CLEAR TABLE Surveys;--" which led to all surveys entered before hers being removed from the results.

正确
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 depends on random loss based on survey order, which is unrelated to salary.

- [ ] One of the students submitted his name as "ROBERT'); foreach student in SURVEYS{ if (student.salary>80000) REMOVE student;}" causing all students who submitted surveys with a salary entry over 80,000 to be removed from the results.

未选择的是正确的

- [ ] Students who accepted a low-salaried job offer tended not to reveal it.

未选择的是正确的

- [x] The person recording the information accidentally lost some of the completed survey forms.

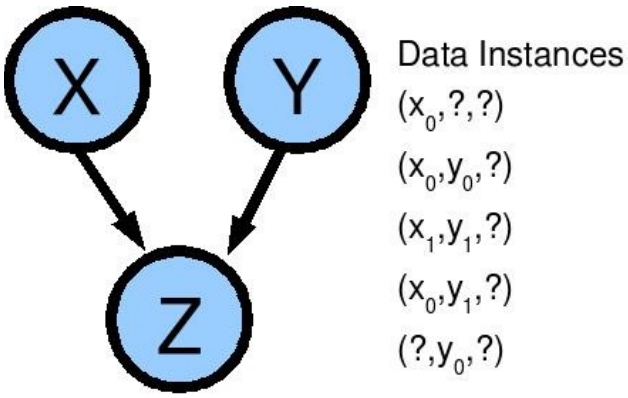
正确
We say data is MAR if whether the data is missing is independent of the missing values themselves given the observed values. This is MAR because whether the data is missing depends on random loss that does not correspond to salary. In fact, this is MCAR.

1 / 1 分

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?



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

正确
The expected sufficient statistics for the assignment (x_0, y_0, z_0) are the sum of the probability that each instance is consistent with that assignment (which is 0 for all inconsistent instances).

- $\bar{M}[x_0, y_0, z_0] = P(z_0 \mid x_0, \theta) + P(z_0 \mid x_0, y_0, \theta) + P(z_0 \mid y_0, \theta)$
- $\bar{M}[x_0, y_0, z_0] = 3$

✔ 1 / 1 分

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...

- (selected) requires probabilistic inference, while it DOES NOT in the case of fully observed data.

正确
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.

- cannot be achieved by probabilistic inference, while it CAN in the case of fully observed data.
- requires probabilistic inference, AS IN the case of fully observed data.

✔ 1 / 1 分

4。

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.

- [ ] Choosing which edges involving the hidden nodes to add to the graph.

未选择的是正确的

- [ ] Choosing the number of hidden variables to add to the graphical model.

未选择的是正确的

- [x] Given a fixed set of edges, learning the parameters in the table CPDs of each hidden node.

正确
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.

- [x] Given a fixed set of edges, learning the parameters in the table CPDs of observed nodes that have hidden nodes as parents.

正确
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.