Lecture 13: Missing data

STATS 202: Data mining and analysis

Sergio Bacallado

October 23, 2013

Announcements

- Homework drop box will be set up on Wednesday in the second floor of Sequoia hall. No late homework rule still applies.
- The midterm is next Monday. A practice exam will be posted on Wednesday.
- *Material:* Anything said in lecture until Friday October 18, and anything in Chapters 2, 3, 4, 5, and 10 of the book is fair game.
- The exam will be closed book and closed notes.
- No calculators or computers necessary.

Announcements

- How much math?
  - One problem will be a simple proof or derivation.
  - Most equations needed will be provided.
- How much R?
  - You will be asked to interpret code without documentation. This shouldn't be difficult if you have done the homework independently.
- *SCPD*: instructions for the exam (delivery times, mechanisms, rules) will be sent this week via email. Please, contact SCPD directly with questions on how to choose a proctor.

Missing data is everywhere

- Survey data (nonresponse).
- Longitudinal studies and clinical trials (dropout).
- Recommendation systems.
- Data integration.

Mechanisms for missing data

- **Missing completely at random:** We remove elements from a column X_j of X at random.
  Example. We run a taste study for 20 different drinks. Each subject was asked to rate only 4 drinks chosen at random.
- **Missing at random:** The pattern of missingness depends on other predictors.
  Example. In a survey, poor subjects were less likely to answer a question about drug use than wealthy subjects.
  - Missingness is related to observed predictors (income).
  - Missingness is related to unobserved predictors.
- **Censoring:** The pattern of missingness is closely related to the missing variable.
  Example. High earners less likely to report their income.

Dealing with missing data

- Some tree-based methods can deal with missing data naturally.
- **Single imputation**: We replace each missing value with a single number.
  1. Replace with the mean or median of the column.
  2. Replace with a random sample from the non-missing values in the column.
  3. Replace missing values in X_j with a regression estimate from other predictors, X_{-j}.

  - Methods 1 and 2 can give biased coefficients if the data is not missing completely at random. Method 3 does not have bias if the missingness is predicted well by X_{-j}.
  - Method 3 yields standard errors that are artificially small.

Dealing with missing data

- **Multiple imputation**: We replace each missing value in X_j with a regression estimate from the other predictors X_{-j}, plus some noise. This is repeated several times.
  - If the regression fit of X_j onto X_{-j} is good, the standard errors from this method can be unbiased.

Missing data in more than one variable

Problem: What if we have missing data in almost every column $X_1, X_2, \ldots, X_p$?

- **Iterative multiple imputation**: Start with a simple imputation. Then, iterate the following:
  1. Multiple imputation of X_1 from X_{-1}.
  2. Multiple imputation of X_2 from X_{-2}.

     ...
  3. Multiple imputation of X_p from X_{-p}.
- **Model based imputation**: Fit the missing values to a joint statistical model for all the predictors. Rarely worth the trouble.

Missing data in more than one variable

Problem: What if we have missing data in almost every column $X_1, X_2, \ldots, X_p$?

- **Matrix completion**:

  In linear regression, $\hat{y}$ can be understood as a projection of y onto the space spanned by the columns of X. In a sense, what matters is this column space.

  Matrix completion algorithms find a matrix X' which is similar to X in its non-missing values, and has low rank (a low dimensional column space). For example,

$$\min_{\text{subject to rank}(X')=k} \|X' - X\|.$$

  The appropriate rank can be set as a tuning parameter.

Some practical considerations

- It is important to visualize summaries or plots for the pattern of missingness.
- If the pattern of missingness is informative, include it as a dummy variable.
- If a variable has too many missing values, it is worth it to include it?
- If we are using a method that allows it, consider weighting variables according to the rate of missing data.

  Example. In nearest neighbors, scale each variable and multiply by $(100-$ % missing).
- Some variables are restricted to be positive, or bounded above.
- Are there any variables that are non-linear functions of others?