Missing Data

Unit 2 Live Session
MSDS 7333
Spring 2017

From STATS 202 course, Stanford University

Missing data are everywhere

- Survey data: nonresponse.
- Longitudinal studies and clinical trials: dropout.
- Recommendation systems: different individuals interact with or express preferences for different items.
- Data integration: different variables collected by different organizations or in different experiments or trials.

Mechanisms for Missing Data

- Missing completely at random: Pattern of missingness independent of missing values and the values of any measured variables.
  - 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, but conditional on observed variables, missingness is independent of missing value.
  - Example. In a survey, poor subjects were less likely to answer a question about drug use than wealthy subjects. Related to observed predictors (income) but not drug use.
- Missing not at random: The pattern of missingness is related to the missing variable, even after correcting for measured variables.
  - Example 1: High earners less likely to report their income.
  - Example 2: Record time until subjects have an accident but only follow for three years (censoring).

Dealing with Missing Data

- Categorical case: Treat "missing" as an additional category.
- Surrogate variables: Tree-based methods like CART can deal with missingness by introducing surrogate variables!
- Observation deletion: Delete observations with missing values.
  - Drawbacks: Reduces dataset size, can bias input feature space, doesn’t work at test time.
- Variable deletion: Delete variables with missing values
  - Drawbacks: May be throwing away valuable variable, can bias input feature space.

Single Imputation

- 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 .
- Drawbacks:
  – 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 missing variable is predicted well by X_-j .
  – Resulting inferences about estimated parameters or predictions do not account for uncertainty in missing values.

Multiple Imputation

- Multiple imputation: Form many imputed datasets by positing a distribution over unobserved variables and repeatedly sampling from that distribution.
  - For example, each sample could be obtained by replacing each missing value in Xj with a regression estimate from the other predictors X_-j , plus some noise. This is repeated several times.
  - Run entire analysis on each dataset, and use multiple results to get a better estimate of uncertainty.
- If the regression fit of Xj onto X_-j is good, the standard errors from this method can be unbiased.

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, you may want to exclude it from your analysis (you can still include a missingness indicator for that variable.)
- 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).
- When imputing, keep in mind that some variables are restricted to be positive or bounded.
- Some variables are well modeled as non-linear functions of other variables.

Breakout Sessions

- You will study 1 of 3 missing data scenarios.
- Read material and then have discussion (questions to discuss are included)
- Choose presenter and report back summary of scenario and discussion questions.