Lecture 11: Missing and relational data

STATS 202: Data mining and analysis

Rajan Patel

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.

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.

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.

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.

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.

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.

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.

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}.

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.

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 missing variable is predicted well by X_{-j}.

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 missing variable 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.

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 $X_1, X_2, \ldots, X_p$?

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}.

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.

Some practical considerations

- It is important to visualize summaries or plots for the pattern of missingness.

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.

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?

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 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.

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?

Relational data

The observations have the form of a graph.

Examples.

Relational data

The observations have the form of a graph.

Examples.

- Links between websites.

Relational data

The observations have the form of a graph.

Examples.

- Links between websites.
- Relationships between accounts in social networks.

Relational data

The observations have the form of a graph.

Examples.

- Links between websites.
- Relationships between accounts in social networks.
- Transmission networks for contagious diseases.

Relational data

The observations have the form of a graph.

Examples.

- Links between websites.
- Relationships between accounts in social networks.
- Transmission networks for contagious diseases.

The links can be **directed** or **undirected**.

Relational data

The observations have the form of a graph.

Examples.

- Links between websites.
- Relationships between accounts in social networks.
- Transmission networks for contagious diseases.

The links can be **directed** or **undirected**.

There can be different types of link (friend, follower, followed).

Relational data

The observations have the form of a graph.

Examples.

- Links between websites.
- Relationships between accounts in social networks.
- Transmission networks for contagious diseases.

The links can be **directed** or **undirected**.

There can be different types of link (friend, follower, followed).

We can observe the graph in time (how do social networks grow?).

Relational data

The observations have the form of a graph.

Examples.

- Links between websites.
- Relationships between accounts in social networks.
- Transmission networks for contagious diseases.

The links can be **directed** or **undirected**.

There can be different types of link (friend, follower, followed).

We can observe the graph in time (how do social networks grow?).

Each vertex can have additional features or **metadata**.

PageRank algorithm

- Invented by Sergei Brin and Larry Page of Google.

PageRank algorithm

- Invented by Sergei Brin and Larry Page of Google.
- Uses a graph of links between websites to rank websites by “importance”.

PageRank algorithm

- Invented by Sergei Brin and Larry Page of Google.
- Uses a graph of links between websites to rank websites by "importance".
- **Motivation:**
  - Consider the problem of searching the web using the query "birth control".

PageRank algorithm

- Invented by Sergei Brin and Larry Page of Google.
- Uses a graph of links between websites to rank websites by “importance”.
- **Motivation:**
  - Consider the problem of searching the web using the query "birth control".
  - There are millions of pages containing the term.

PageRank algorithm

- Invented by Sergei Brin and Larry Page of Google.
- Uses a graph of links between websites to rank websites by "importance".
- **Motivation:**
  - Consider the problem of searching the web using the query "birth control".
  - There are millions of pages containing the term.
  - Analyzing the content of each website semantically to infer which one is more likely to satisfy the user is very expensive.

PageRank algorithm

- Invented by Sergei Brin and Larry Page of Google.
- Uses a graph of links between websites to rank websites by “importance”.
- **Motivation:**
  - Consider the problem of searching the web using the query "birth control".
  - There are millions of pages containing the term.
  - Analyzing the content of each website semantically to infer which one is more likely to satisfy the user is very expensive.
  - We need a way to rank websites, to filter out all those that are rarely visited. This information is given by links.

PageRank algorithm

Consider a hypothetical **surfer** who is jumping from website to website by clicking on random links.

PageRank algorithm

Consider a hypothetical **surfer** who is jumping from website to website by clicking on random links. Intuitively, the websites that are visited more frequently can be considered more important in the network of links.

PageRank algorithm

Consider a hypothetical **surfer** who is jumping from website to website by clicking on random links. Intuitively, the websites that are visited more frequently can be considered more important in the network of links.

Will the surfer visit every website eventually?

PageRank algorithm

Consider a hypothetical **surfer** who is jumping from website to website by clicking on random links. Intuitively, the websites that are visited more frequently can be considered more important in the network of links.

Will the surfer visit every website eventually? No. It is possible to get stuck in a website with no outgoing links, or to be stuck in a loop between two websites, for example.

PageRank algorithm

Consider a hypothetical **surfer** who is jumping from website to website by clicking on random links. Intuitively, the websites that are visited more frequently can be considered more important in the network of links.

Will the surfer visit every website eventually? No. It is possible to get stuck in a website with no outgoing links, or to be stuck in a loop between two websites, for example.

To avoid this problem, we modify the random walk, such that at every step, with probability $1 - q$, we pick a website at random, and with probability q we go through one of the links in the current website at random.

PageRank algorithm

- The **surfer**'s random walk is a Markov chain on the set of websites.

PageRank algorithm

- The **surfer**'s random walk is a Markov chain on the set of websites.
- It is a fact that the frequency with which the surfer visits any website converges to some limit.

PageRank algorithm

- The **surfer**'s random walk is a Markov chain on the set of websites.
- It is a fact that the frequency with which the surfer visits any website converges to some limit.
- The PageRank of a website is this limiting frequency.

PageRank algorithm

Let P_{ij} be the probability of jumping from website i to website j, then

$$P_{ij} = (1-q)\frac{1}{n} + q\left[\frac{\text{\# of links from } i \text{ to } j}{\text{\# of links out of } i}\right]$$

PageRank algorithm

Let P_{ij} be the probability of jumping from website i to website j, then

$$P_{ij} = (1-q)\frac{1}{n} + q\left[\frac{\text{\# of links from } i \text{ to } j}{\text{\# of links out of } i}\right]$$

The limiting frequency of website j, π_j, must satisfy

$$\pi_j = \sum_{i=1}^{n} \pi_i P_{ij}$$

or in matrix notation $\pi = \pi P$.

PageRank algorithm

Let P_{ij} be the probability of jumping from website i to website j, then

$$P_{ij} = (1-q)\frac{1}{n} + q\left[\frac{\text{\# of links from } i \text{ to } j}{\text{\# of links out of } i}\right]$$

The limiting frequency of website j, π_j, must satisfy

$$\pi_j = \sum_{i=1}^{n} \pi_i P_{ij}$$

or in matrix notation $\pi = \pi P$. That is, π is an eigenvector of the transition probability matrix P with eigenvalue 1.

Finding the limiting frequencies π

In principle, finding the limiting frequencies could require solving the eigendecomposition of a matrix P which is $n \times n$, and this has a complexity which grows as n^3.

Finding the limiting frequencies π

In principle, finding the limiting frequencies could require solving the eigendecomposition of a matrix P which is $n \times n$, and this has a complexity which grows as n^3.

However, it is possible to compute π by starting with the approximation $\pi^{(0)} = (1/n, \ldots, 1/n)$, and iterating:

$$\pi^{(t)} = \pi^{(t-1)} P.$$

The number of iterations necessary for convergence is typically small.

Finding the limiting frequencies π

In principle, finding the limiting frequencies could require solving the eigendecomposition of a matrix P which is $n \times n$, and this has a complexity which grows as n^3.

However, it is possible to compute π by starting with the approximation $\pi^{(0)} = (1/n, \ldots, 1/n)$, and iterating:

$$\pi^{(t)} = \pi^{(t-1)} P.$$

The number of iterations necessary for convergence is typically small.

The matrix-vector multiplication in each iteration can be sped up using sparse matrix techniques.

How can PageRank be used in web search?

One idea:

1. Find all websites that contain all query terms.
2. Display them in order of their PageRank.

How can PageRank be used in web search?

One idea:

1. Find all websites that contain all query terms.
2. Display them in order of their PageRank.

A more likely approach:

1. Use PageRank to select the 10,000 most important pages which contain the query terms.
2. Rank these 10,000 pages by analyzing their content, integrating information about the user, etc.

Working with graphs in R and Python

The package `igraph` implements a lot of utilities for analyzing graphs in R and Python.

It has a function page.rank, among many others.