Data

Lyle Ungar

Learning objectives

Missing at random

Imputation

Indicator functions for missing

How to handle missing data?

x_1	x_2	x_3	x_4	y
1.1	4	T	3.0	1
NA	4	T	2.2	1
0.9	2	NA	0.8	0
1.0	3	F	NA	0

Simple imputation

x_1	x_2	x_3	x_4	y	x_1	x_2	x_3	x_4	y
1.1	4	T	3.0	1	1.1	4	T	3.0	1
NA	4	T	2.2	1	1.0	4	T	2.2	1
0.9	2	NA	0.8	0	0.9	2	T	0.8	0
1.0	3	F	NA	0	1.0	3	F	2.0	0

Replace with average or majority

Simple imputation

x_1	x_2	x_3	x_4	y	x_1	x_2	x_3	x_4	y
1.1	4	T	3.0	1	1.1	4	1	3.0	1
NA	4	T	2.2	1	1.0	4	1	2.2	1
0.9	2	NA	0.8	0	0.9	2	0.67	0.8	0
1.0	3	F	NA	0	1.0	3	0	2.0	0

Replace with average or majority

Fancier imputation

x_1	x_2	x_3	x_4	y	x_1	x_2	x_3	x_4	y
1.1	4	1	3.0	1	1.1	4	1	3.0	1
1.0	4	1	2.2	1	1.1	4	1	2.2	1
0.9	2	0.67	0.8	0	0.9	2	-1	0.8	0
1.0	3	0	2.0	0	1.0	3	0	?	0

$x_1 = c_0 + c_2x_2 + c_3x_3 + c_4x_4 = 0.7 + 0.1\ x_2$

$x_3 = c_0 + c_1x_1 + c_2x_2 + c_4x_4 = -3 + x_2$

Use regression to estimate missing values

Imputation

- **Often done using EM**
  - If you know the regression models to predict each feature as a function of the others, you can estimate the missing values
  - If you know all the missing values, you can fit the regression models

Missing at Random?

- Grades on front page of application to Penn
- Measured chemical composition (range 0.001-0.1)
- Sensor failure?
- Clicker: how valuable do you think attending lecture is?
- Tax return

Better: add indicators for missing

x_1	x_{1m}	x_2	x_{2m}	x_3	x_{3m}	x_4	x_{4m}	y
1.1	0	4	0	1	0	3.0	0	1
1.0	1	4	0	1	0	2.2	0	1
0.9	0	2	0	0.67	1	0.8	0	0
1.0	0	3	0	0	0	2.0	1	0

How to handle categorical data?

x_1	x_{1R}	x_{1G}	x_{1B}	x_{1NA}
R	1	0	0	0
G	0	1	0	0
B	0	0	1	0
R	1	0	0	0
NA	0	0	0	1

What if there are *lots* of categories?

- ZIP codes (42,000)
- FIPS codes
- SIC Codes

1623	Water, Sewer, Pipeline, Comm & Power Line Construction
1629	Heavy Construction, Not Elsewhere Classified[6]
1700	Construction - Special Trade Contractors
1731	Electrical Work
2000	Food and Kindred Products
2011	Meat Packing Plants
2013	Sausages & Other Prepared Meat Products
2015	Poultry Slaughtering and Processing
2020	Dairy Products
2024	Ice Cream & Frozen Desserts
2030	Canned, Frozen & Preserved Fruit, Veg & Food Specialties
2033	Canned, Fruits, Veg, Preserves, Jams & Jellies

What if there are *lots* of categories?

- **Dimensionality reduce: cluster, PCA, …**
- **Possible features**
  - Geolocation
  - Demographics
  - Co-occurrence
  - Product sales, twitter language, …
- **Often someone has already done the clustering**

Conclusions

- **Most data is not missing at random**
- **So add an indicator variable to indicate missing**
  - And fill in the missing value with mean or majority