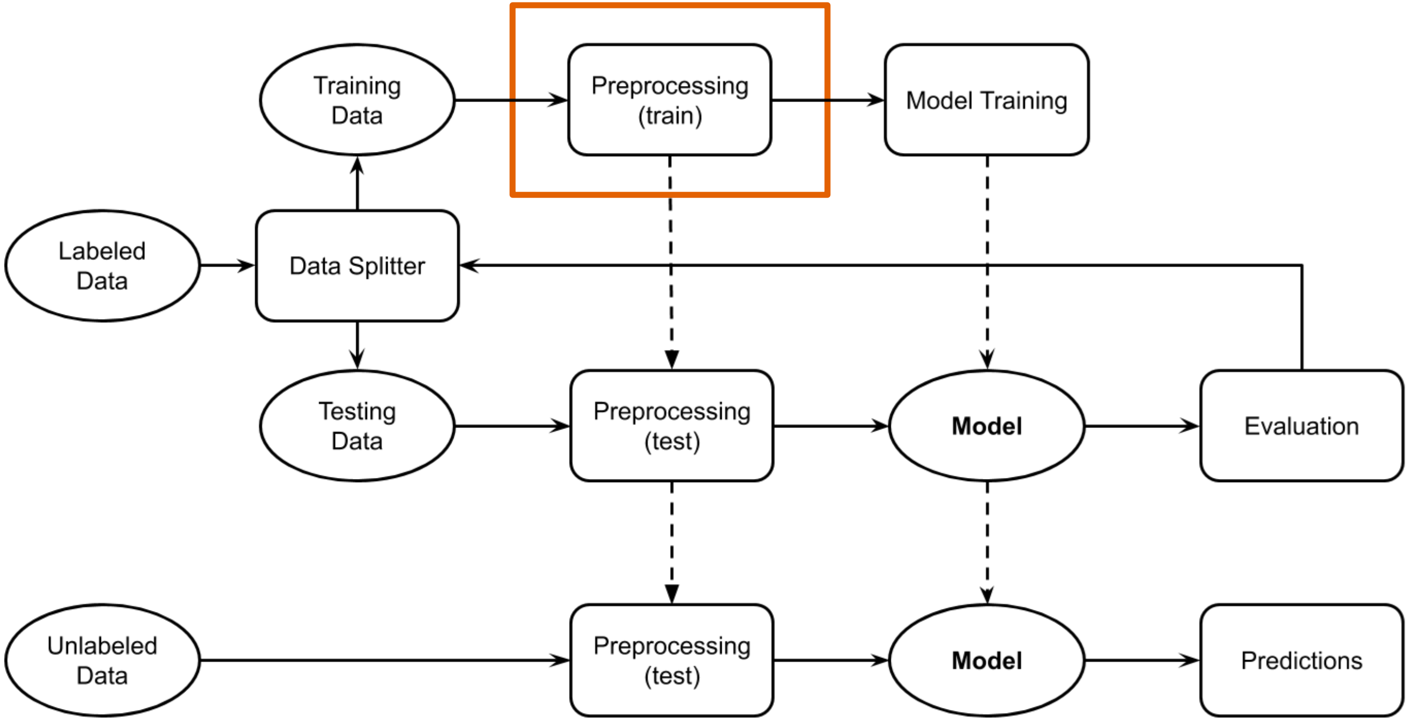
Foundations of Data Science & Analytics: Preprocessing

Ezgi Siir Kibris

Introduction to Data Mining, 2nd Edition
by
Tan, Steinbach, Karpatne, Kumar

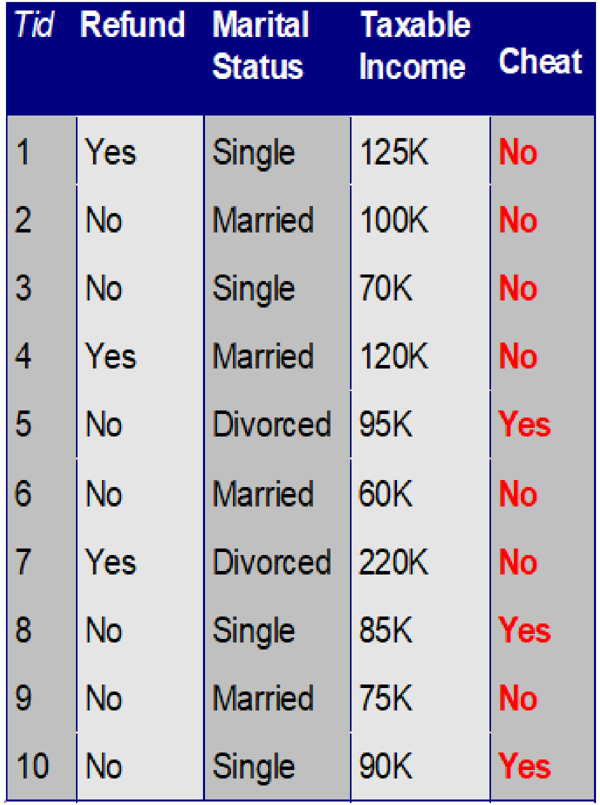
Data Mining / Machine Learning Pipeline



Preprocessing

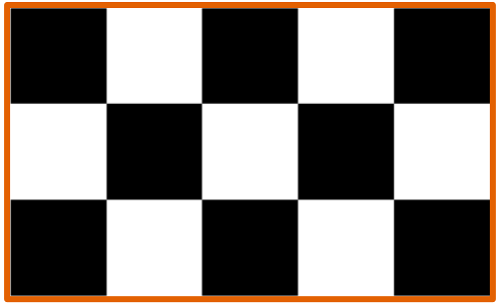
Goal:

Transform raw data to a format that machine learning / data mining models can (easily) learn from.



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Preprocessing



1	0	1	0	1
0	1	0	1	0
1	0	1	0	1

I need to talk to you

i	need	to	talk	you
1	1	2	1	1

Preprocessing

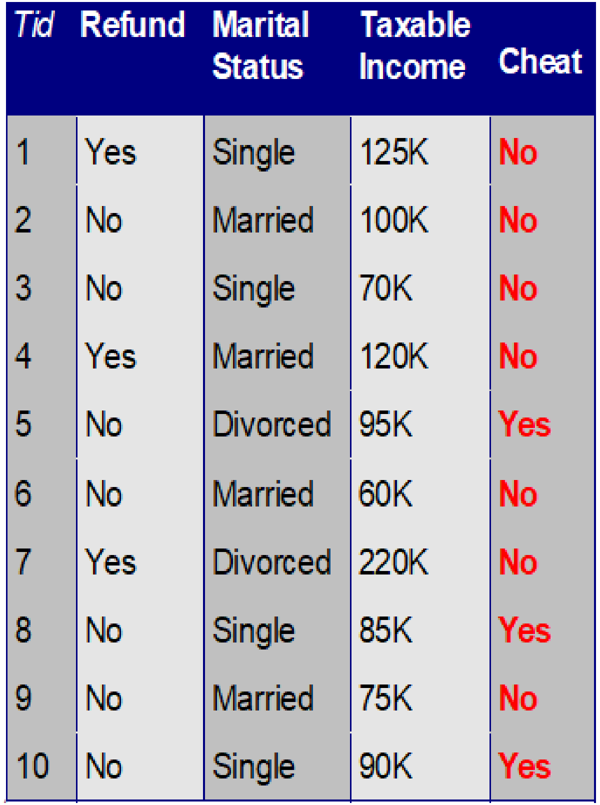
- Manipulating Data (rows)
  - Sampling

Only on training

- Manipulating Values
  - Discretization
  - Normalization

Same on training, and test data

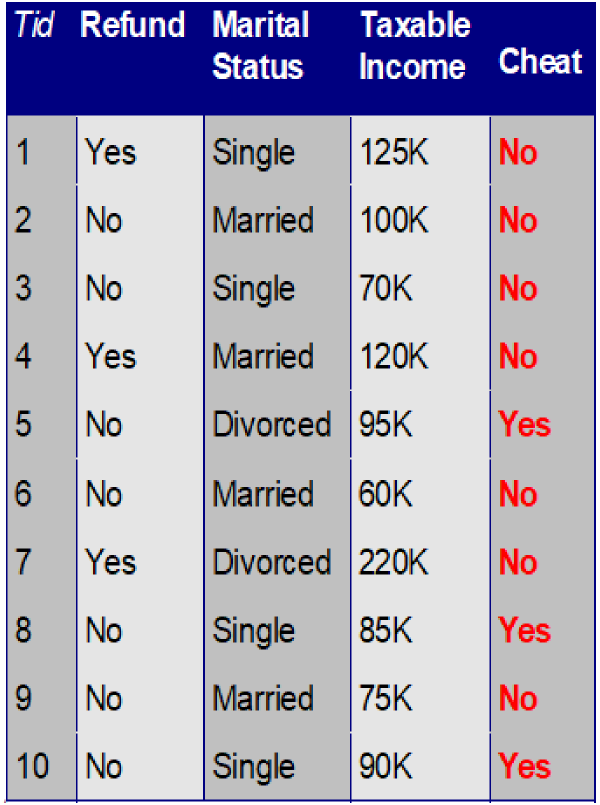
- Manipulating Features (columns)
  - Dimensionality Reduction
  - Feature Selection
  - Feature Creation



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Preprocessing

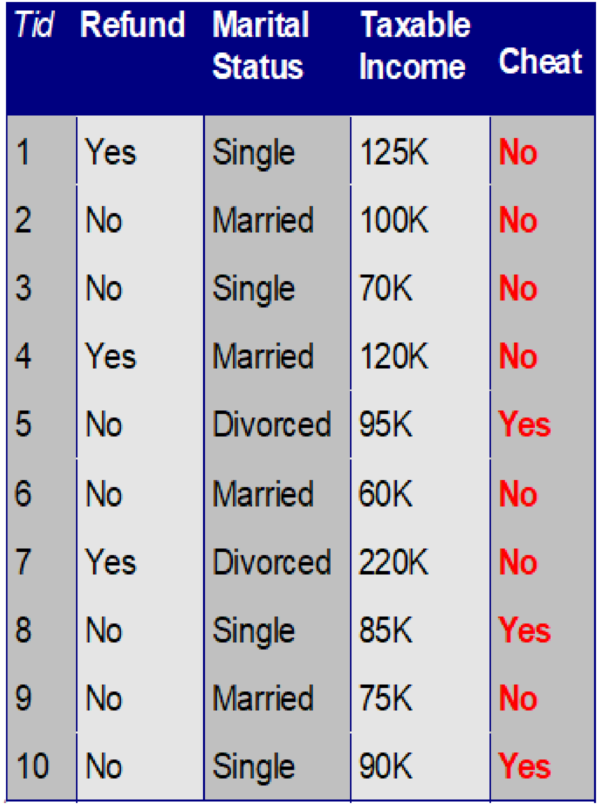
- **Supervised**
  - Requires labels
- **Unsupervised**
  - Does not rely on labels



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Preprocessing

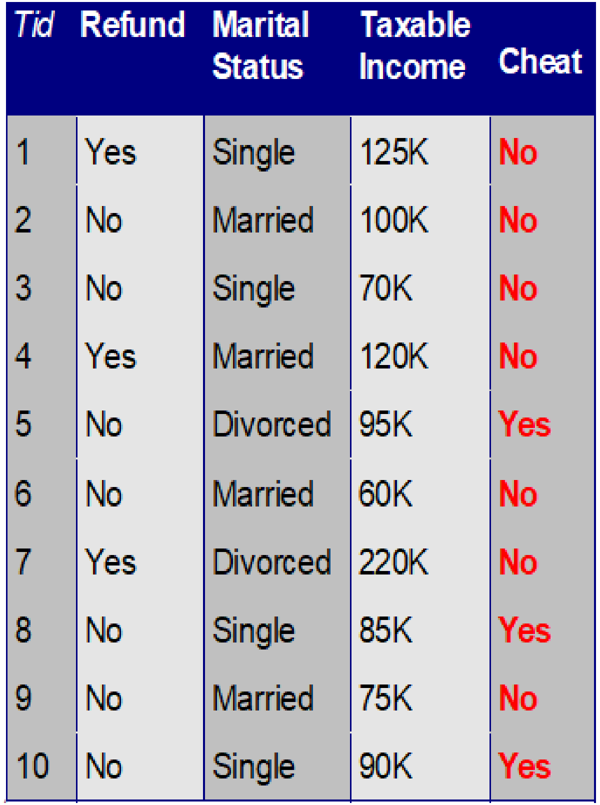
- Manipulating Data (rows)
  - Sampling
- Manipulating Values
  - Discretization
  - Normalization
- Manipulating Feature (columns)
  - Dimensionality Reduction
  - Feature Selection
  - Feature Creation



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Sampling

- Reducing size of data
  - Random sampling
  - Stratified sampling



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Reducing Size of Data

- Random Sampling
  - Sampling without replacement
    - As each item is selected, it is removed from the population
  - Sampling with replacement
    - Objects are not removed from the population as they are selected for the sample.
    - The same object can be picked up more than once

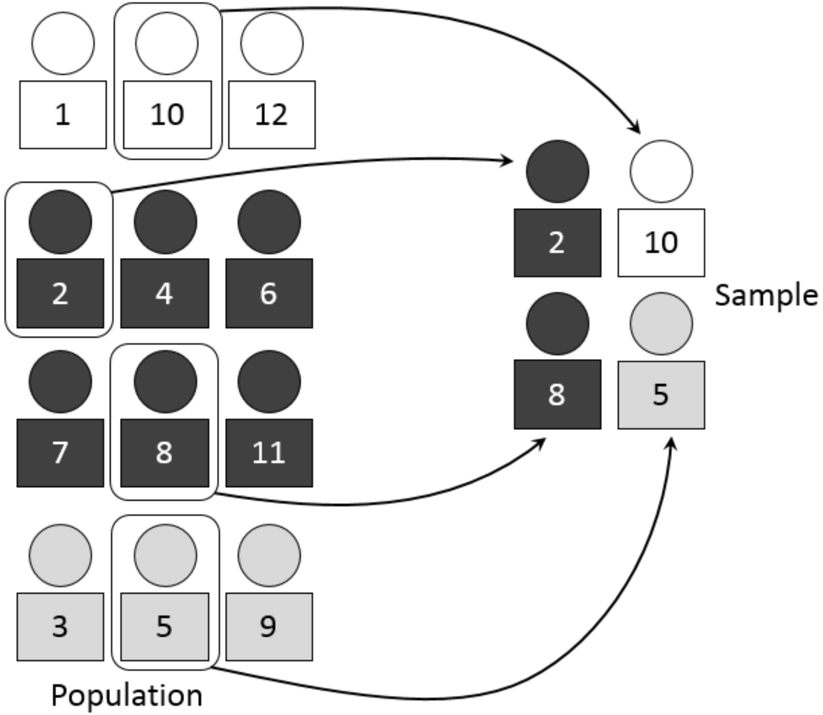
Unsupervised

- Stratified sampling
  - Random sample from each class.
  - Keep the same distribution of classes.
  - Avoid the sampled data to miss some classes.

Supervised

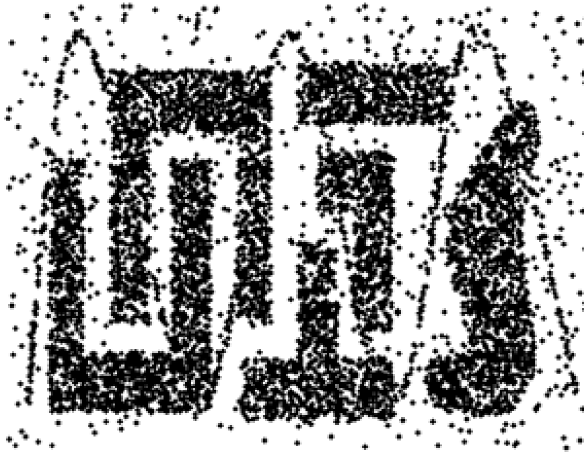
Stratified sampling

1 : 2 : 1

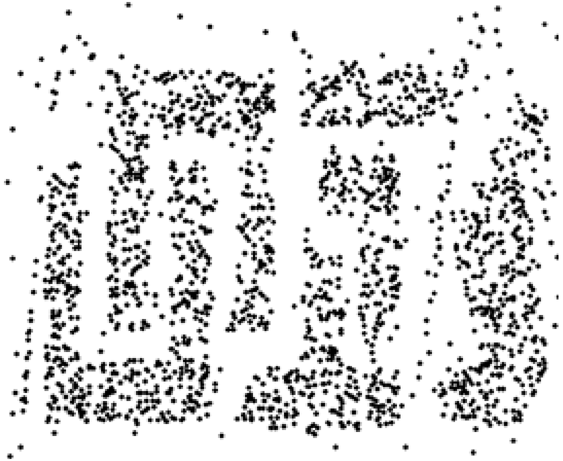


1 : 2 : 1

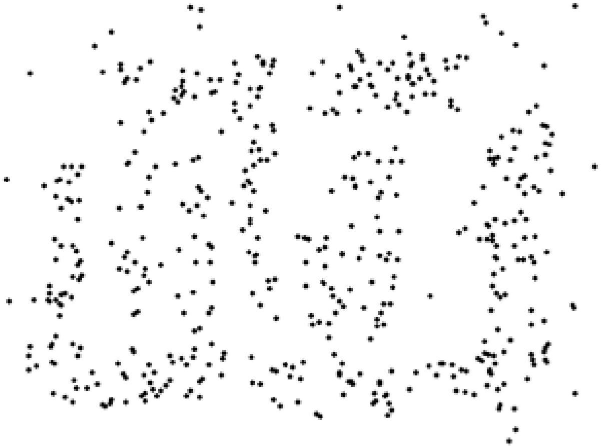
Example



8000 points



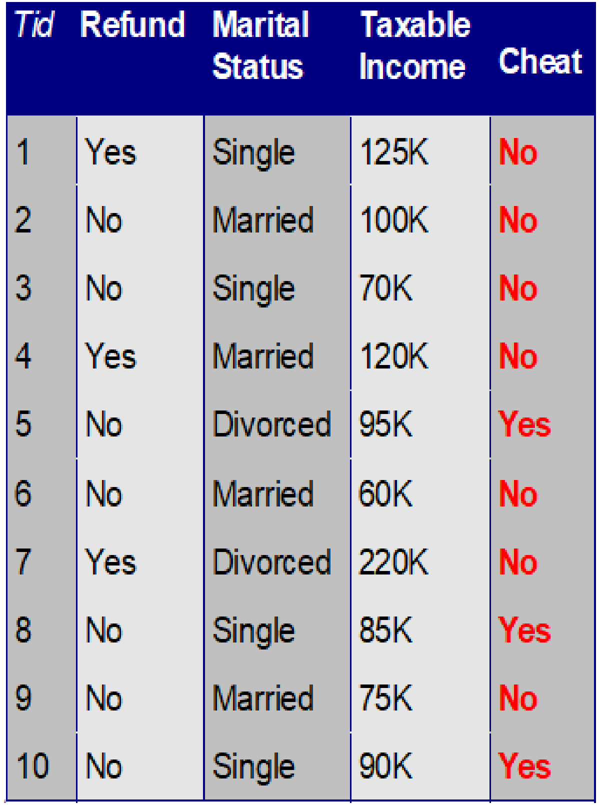
2000 Points



500 Points

Preprocessing

- Manipulating Data (rows)
  - Sampling
- Manipulating Values
  - Discretization
  - Normalization
- Manipulating Feature (columns)
  - Dimensionality Reduction
  - Feature Selection
  - Feature Creation



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Discretization

Discretization is the process of converting a continuous (Interval, Ratio) feature into an **ordinal** feature

- A potentially infinite number of values are mapped into a small number of categories
- Discretization is commonly used in classification
- Many classification algorithms work best if both the independent and dependent variables have only a few values

Types of features

- **Nominal**
  - Examples: ID numbers, eye color, zip codes
- **Ordinal**
  - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
- **Interval**
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- **Ratio**
  - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

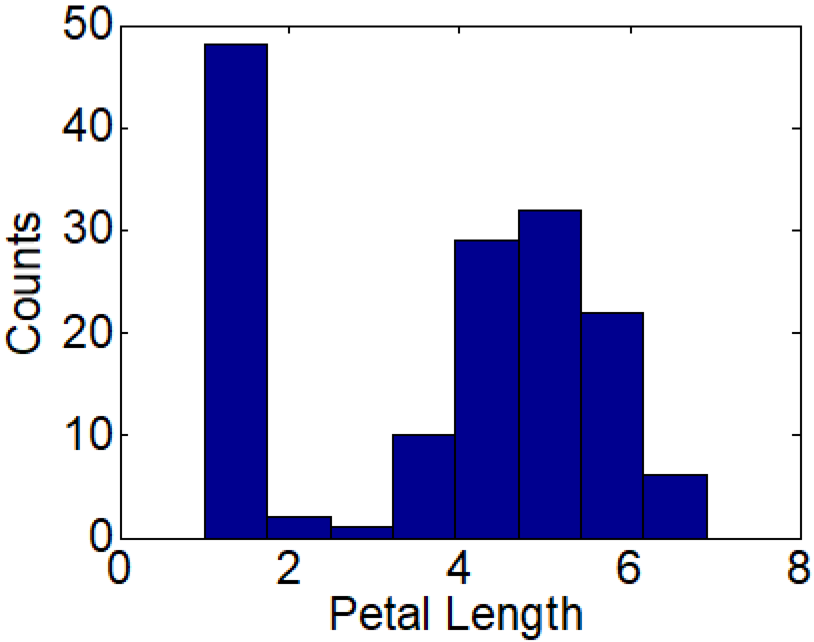
Types of features

- **Distinctness**: =
- **Order**: < >
- **Differences** are meaningful : + -
- **Ratios** are meaningful * /

- **Nominal** feature: distinctness
- **Ordinal** feature: distinctness & order
- **Interval** feature: distinctness, order & meaningful differences
- **Ratio** feature: all 4 properties/operations

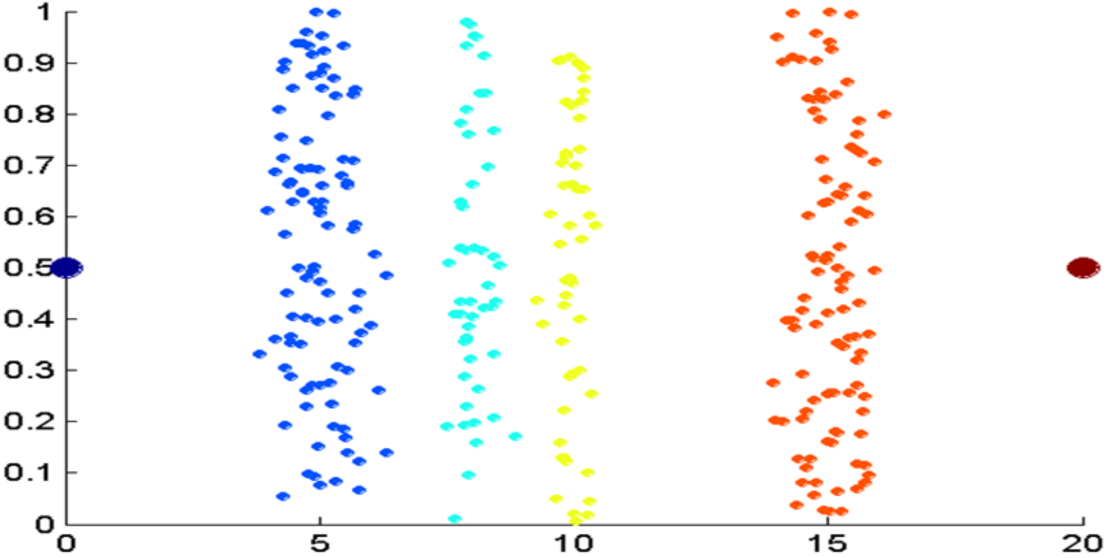
Discretization

- How can we tell what the best discretization is?
    - **Unsupervised** discretization: find breaks in the data values
    - Example:
      Petal Length



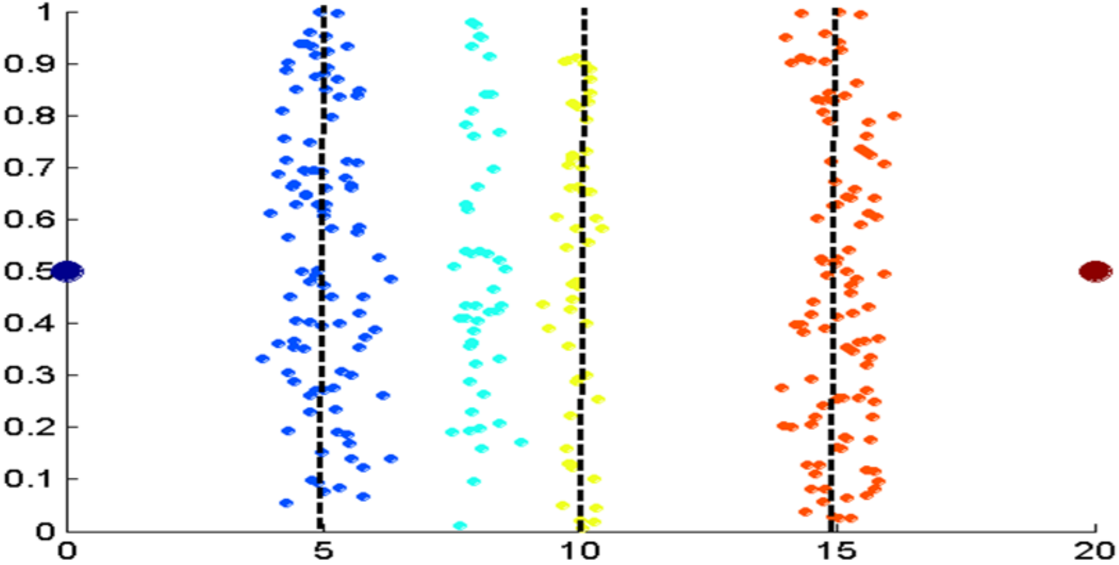
    - **Supervised** discretization: Use class labels to find breaks

Unsupervised Discretization



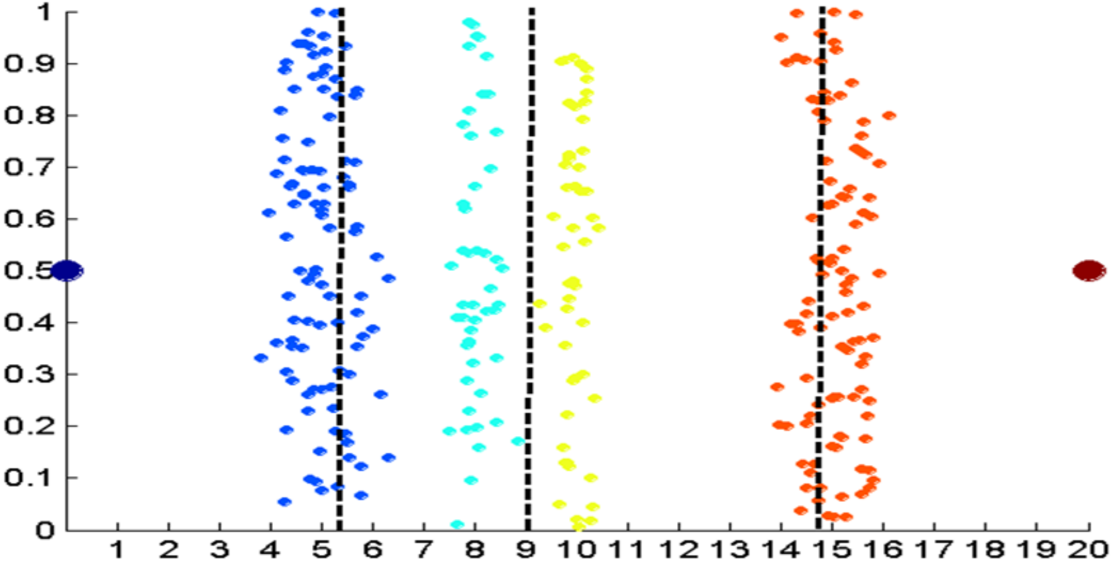
Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.

Unsupervised Discretization



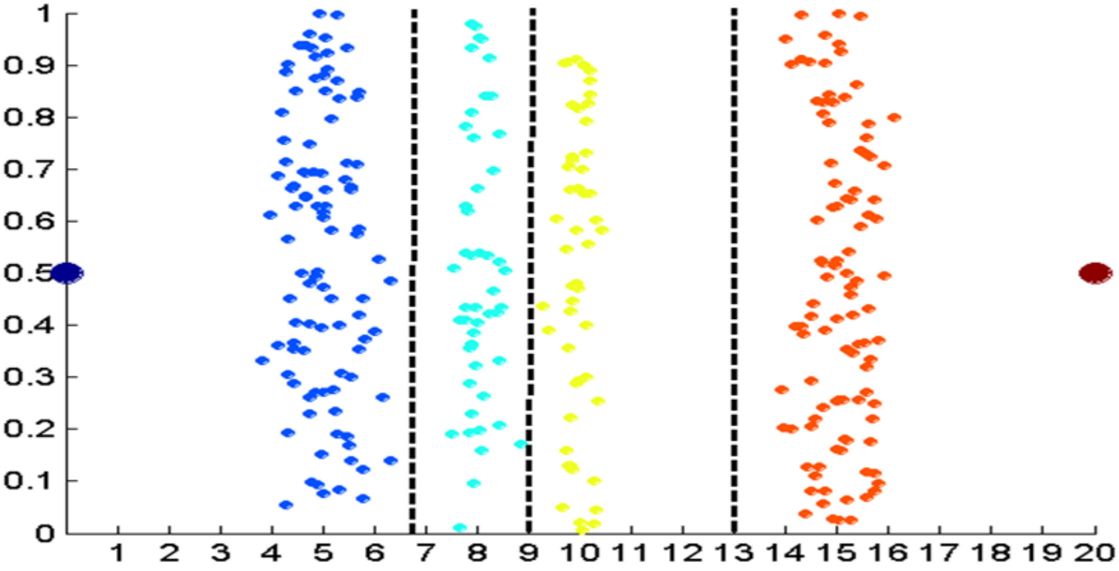
Equal interval width approach used to obtain 4 values.

Unsupervised Discretization



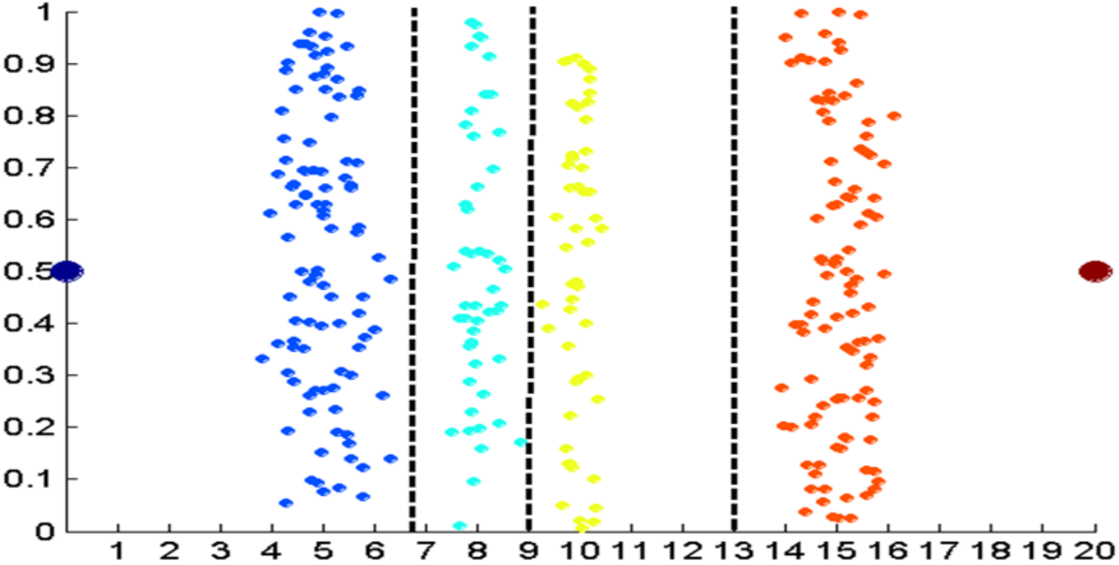
Equal frequency approach used to obtain 4 values.

Unsupervised Discretization



K-means approach to obtain 4 values.

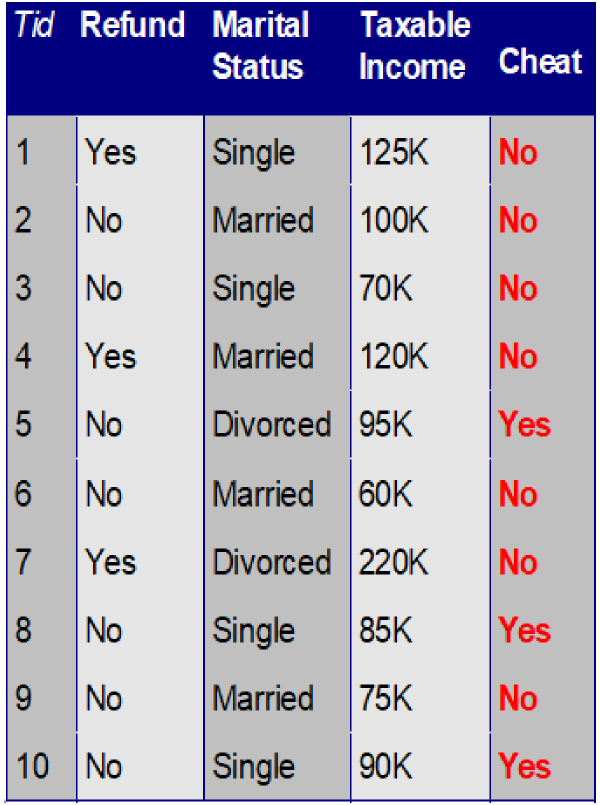
Supervised Discretization



Use entropy to find the best splits, like in decision trees.

Preprocessing

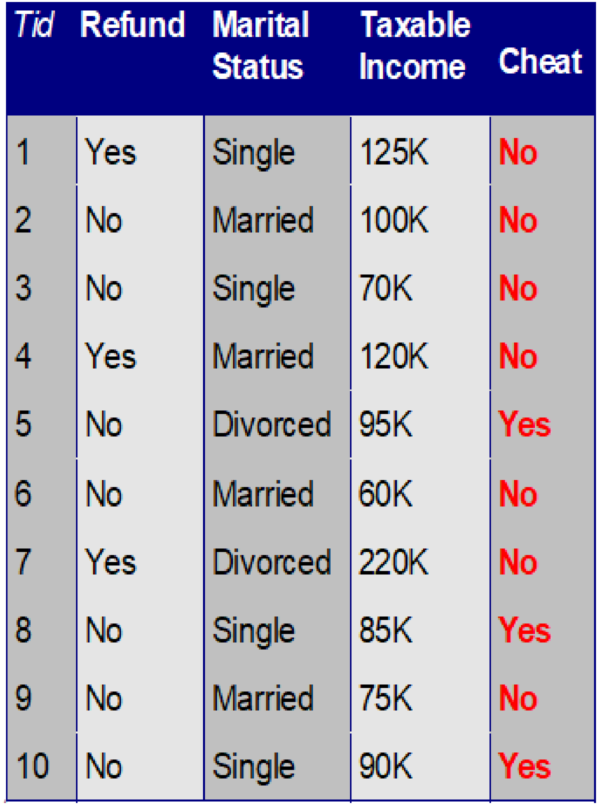
- Manipulating Data (rows)
  - Sampling
- Manipulating Values
  - Discretization
  - Normalization
- Manipulating Feature (columns)
  - Dimensionality Reduction
  - Feature Selection
  - Feature Creation



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Normalization

- Make all features of the same scale (**normalize columns**)
- Make each feature vector of unit length (**normalize rows**)



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Why Normalize?

- Features may have to be scaled to prevent distance measures from being dominated by one of the features
- Example:
  - height of a person may vary from 1.5 m to 1.8 m
  - weight of a person may vary from 90 lb to 300 lb
  - income of a person may vary from $10K to $1M

Normalization

- Standard score (based on normal distribution)

$$X' = \frac{X-\mu}{\sigma}$$

- Min-Max Feature scaling

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Normalization

- L2 Normalization $||\vec{x}'||_2 = 1$

$$\vec{x}' = \frac{\vec{x}}{||\vec{x}||_2} = \frac{\vec{x}}{\sqrt{\sum x_i^2}}$$

- L1 Normalization $||\vec{x}'||_1 = 1$

$$\vec{x}' = \frac{\vec{x}}{||\vec{x}||_1} = \frac{\vec{x}}{\sum |x_i|}$$

X = [1,2,3,4]

sum_square = 1+4+9+16

L2_norm = sqrt(sum_square) = 5.48

X_norm = X / L2_norm = [1/5.48, 2/5.48, 3/5.48, 4/5.48]

sqrt(sum_square(X_norm)) = 1

L2 Normalization on Columns

Training

X1	X2
3	30
2	120

X1_norm = 3.6
X2_norm = 123.7

X1	X2
0.83	0.24
0.55	0.97

Testing

X1	X2
2	100

X1_norm = 3.6
X2_norm = 123.7

X1	X2
0.55	0.81

L2 Normalization on Rows

Training

X1	X2
3	30
2	120

norm1 = 30.15
norm2 = 120.02

X1	X2
0.100	0.995
0.017	0.999

Testing

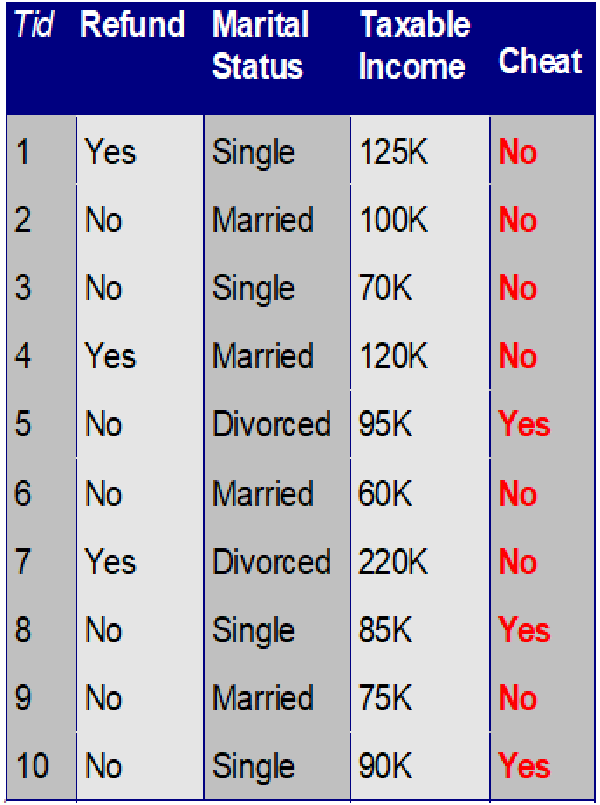
X1	X2
2	100

norm = 100.02

X1	X2
0.020	0.999

Preprocessing

- Manipulating Data (rows)
  - Sampling
- Manipulating Values
  - Discretization
  - Normalization
- Manipulating Feature (columns)
  - Dimensionality Reduction
  - Feature Selection
  - Feature Creation

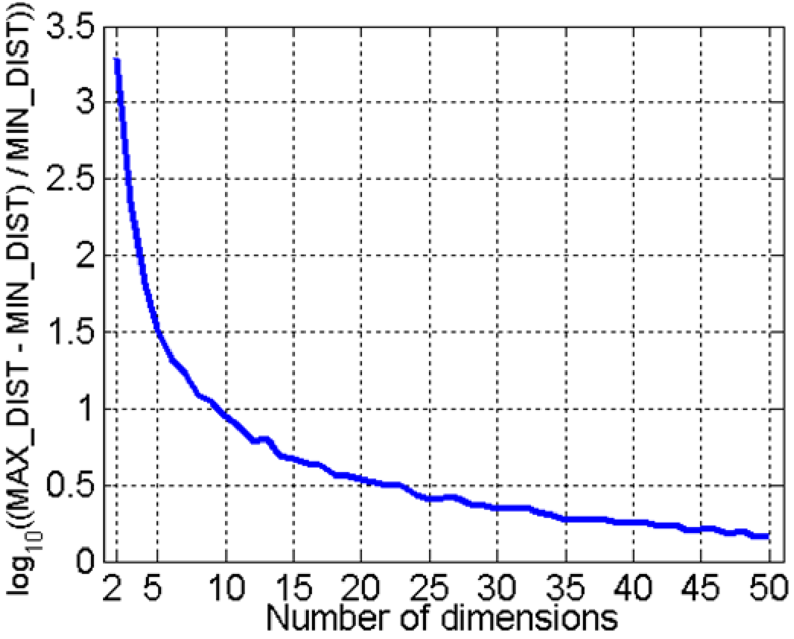


Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies

- Definitions of density and distance between points, which are critical for clustering and outlier detection, become less meaningful



- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points

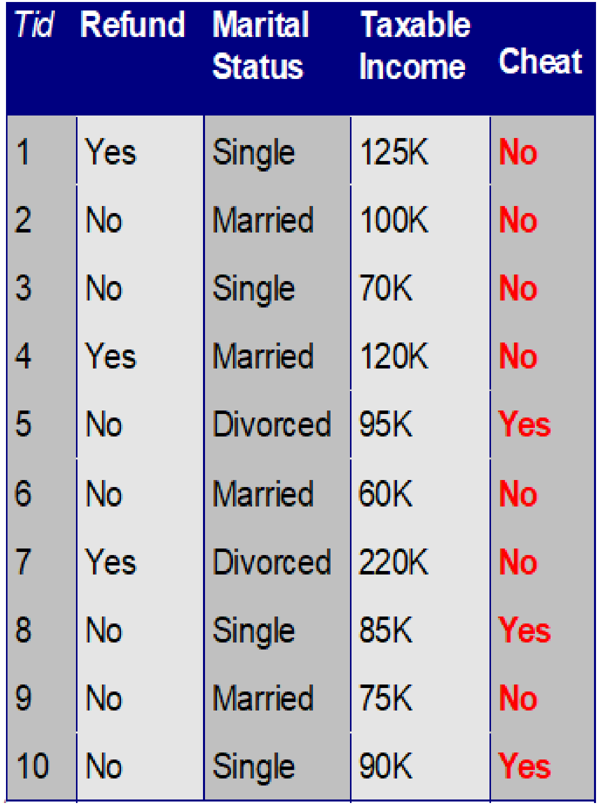
Dimensionality Reduction

- **Purposes**:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

- **Techniques**
  - Principal Components Analysis (PCA)
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

Preprocessing

- Manipulating Data (rows)
  - Sampling
- Manipulating Values
  - Discretization
  - Normalization
- Manipulating Feature (columns)
  - Dimensionality Reduction
  - Feature Selection
  - Feature Creation



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

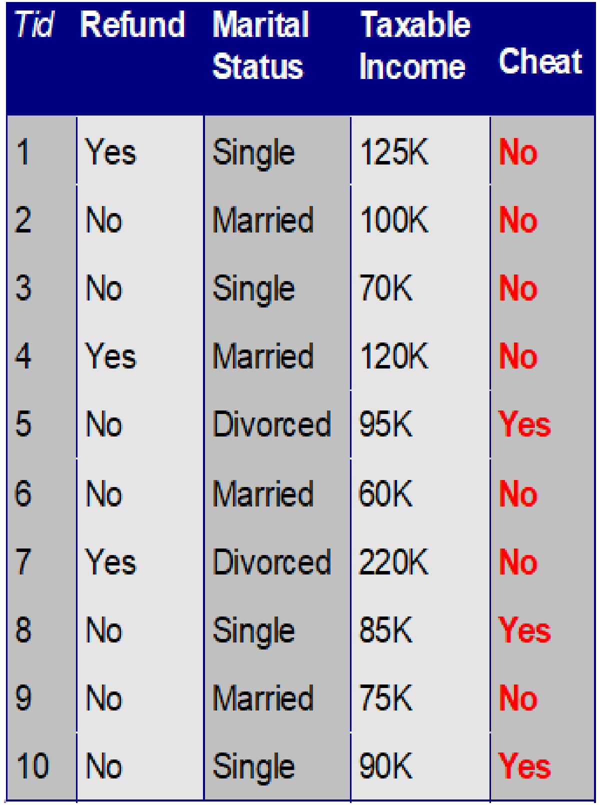
Feature Selection

- Another way to reduce dimensionality of data
- Redundant features (usually unsupervised)
  - Duplicate much or all of the information contained in one or more other features
  - E.g. purchase price of a product and the amount of sales tax paid
- Irrelevant features (usually supervised)
  - Contain no information that is useful for the data mining task at hand
  - E.g. students' ID is often irrelevant to the task of predicting students' GPA

- Correlation matrix

Preprocessing

- Manipulating Data (rows)
  - Sampling
- Manipulating Values
  - Discretization
  - Normalization
- Manipulating Feature (columns)
  - Dimensionality Reduction
  - Feature Selection
  - Feature Creation



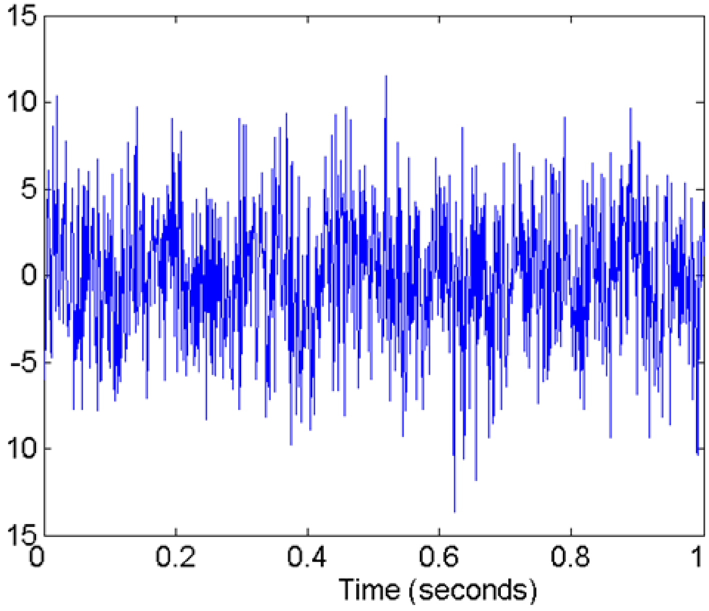
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Feature Creation (unsupervised)

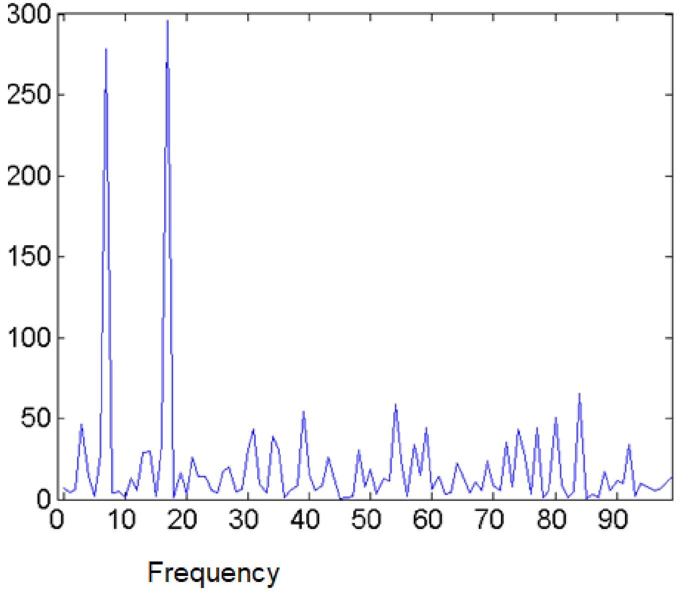
- Create new features that can capture the important information in a data set much more efficiently than the original features

- Three general methodologies:
  - Feature extraction
    - Example: extracting edges from images
  - Feature construction
    - Example: dividing mass by volume to get density
  - Mapping data to new space
    - Example: Fourier transform, kernel trick in SVM

Fourier Transform



Two Sine Waves + Noise



Frequency

Assignment 3

Github!!!