Applied Analytics and Predictive Modeling

Spring 2021

Lecture-15

Lydia Manikonda

manikl@rpi.edu

Rensselaer

Some of the slides adapted from Intro to Data Mining Tan et al. 2nd edition

Today’s agenda

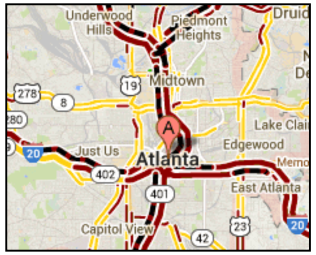
- Recap of the topics we focused until now using questions from the sample Midterm

Large-Scale Data is Everywhere

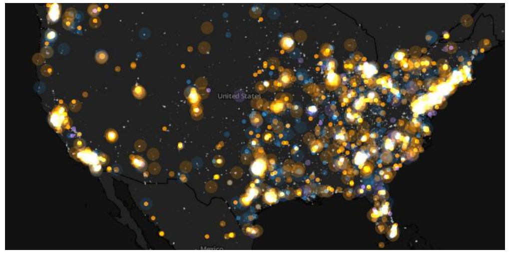
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

Cyber Security

E-Commerce

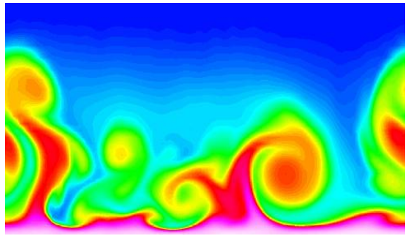


Traffic Patterns



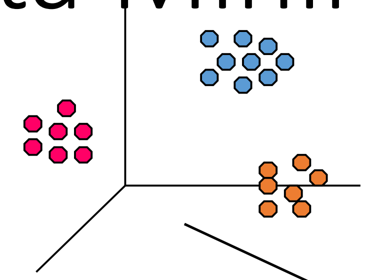
Social Networking: Twitter

Sensor Networks

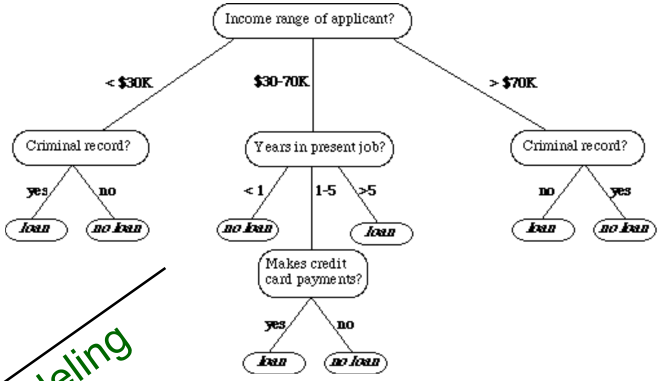


Computational Simulations

Data Mining Tasks



Clustering



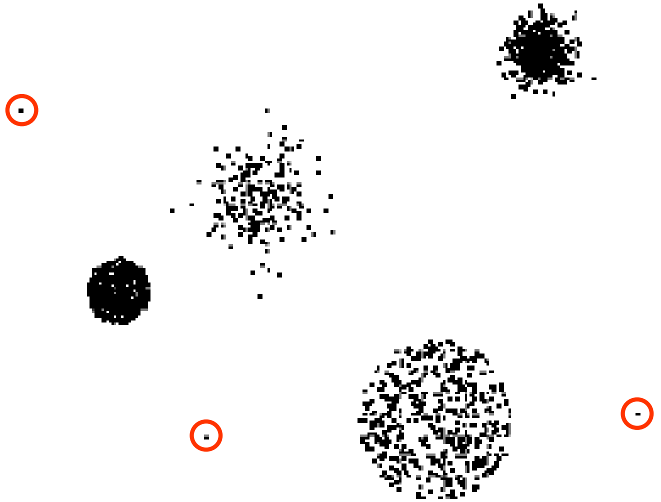
Predictive Modeling

Data

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
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

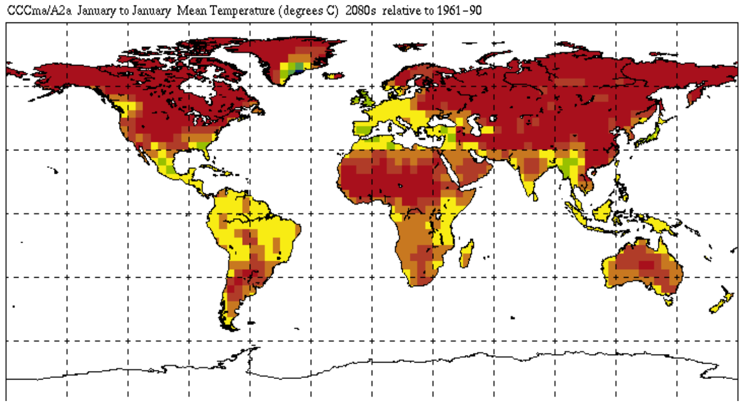
Association Rules

Anomaly Detection



Great Opportunities to solve Society's Major Problems

Improving health care and reducing costs



Predicting the impact of climate change

Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

What is NOT Data Mining?

- What is not Data Mining?
  - Look up phone number in phone directory
  - Query a Web search engine for information about “Amazon”

- What is Data Mining?
  - Certain names are more prevalent in certain US locations (O’Brien, O’Rourke, O’Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g., Amazon rainforest, Amazon.com)

Q2. Sample midterm

A. Looking up phone number – not data mining

B. Group together similar documents returned by search engine according to their context – data mining

C. Certain names are more prevalent in certain US locations – data mining

D. Query a Web search engine – not data mining

What is data?

- Collection of **data objects** and their **attributes**
- According to Tan et al.,

- An **attribute** is a property or characteristic of an object
  - Also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an **object**
  - Also known as tuple, record, point, case, sample, etc.

Attributes

Objects

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

Types of Attributes

- Nominal
  - Examples: ID numbers, zip codes, eye color
- Ordinal
  - Examples: Rankings (expertise level on a scale of 1-10), grades, height {tall, medium, short}
- Interval
  - Examples: Calendar dates, temperature in Celsius or Fahrenheit
- Ratio
  - Examples: Temperature in Kelvin, length, time, counts

Q3. Important characteristics of data

- Dimensionality (number of attributes)
  - High dimensional data brings a number of challenges
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Size
  - Type of analysis may depend on size of data

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Aggregation

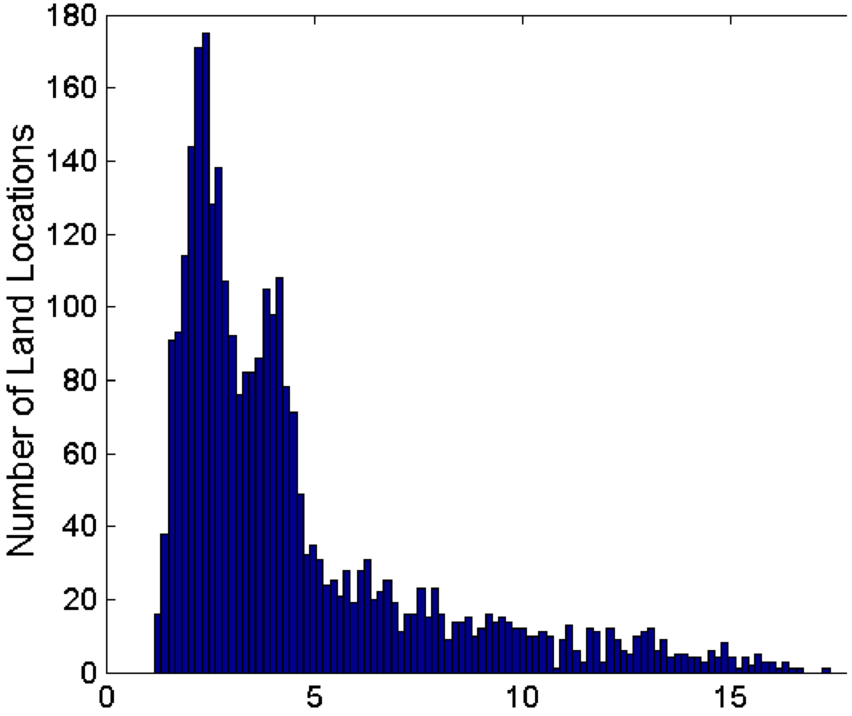
- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc.
    - Days aggregated into weeks, months, or years
  - More “stable” data
    - Aggregated data tends to have less variability

Example: Precipitation in Australia

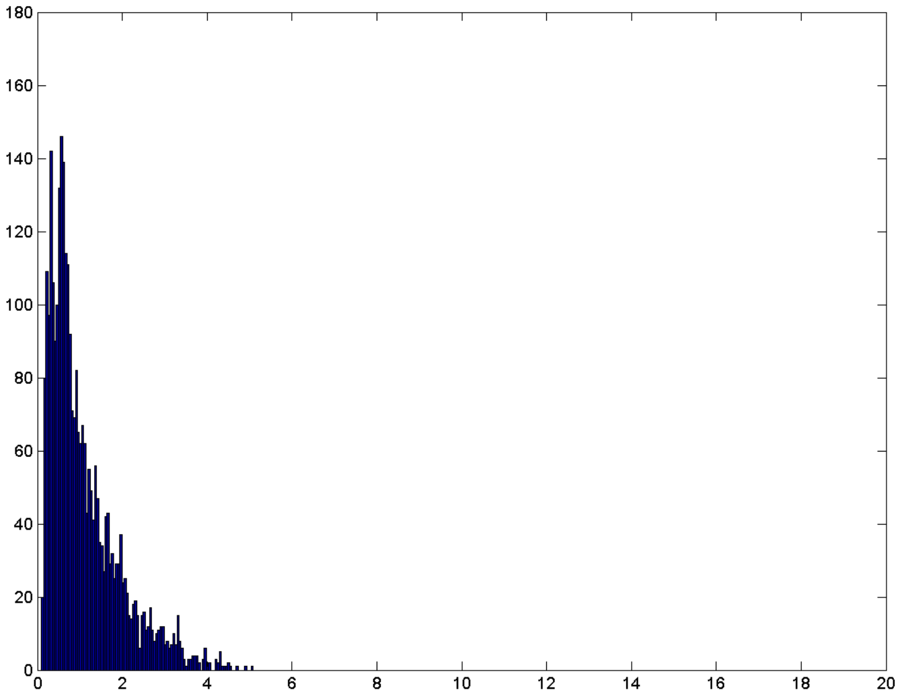
- This example is based on precipitation in Australia from the period 1982 to 1993.
- The next slide shows
  - A histogram for the standard deviation of average monthly precipitation for 3,030 0.5° by 0.5° grid cells in Australia, and
  - A histogram for the standard deviation of the average yearly precipitation for the same locations.
- The average yearly precipitation has less variability than the average monthly precipitation.
- All precipitation measurements (and their standard deviations) are in centimeters.

Example: Precipitation in Australia..

- Variation of precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

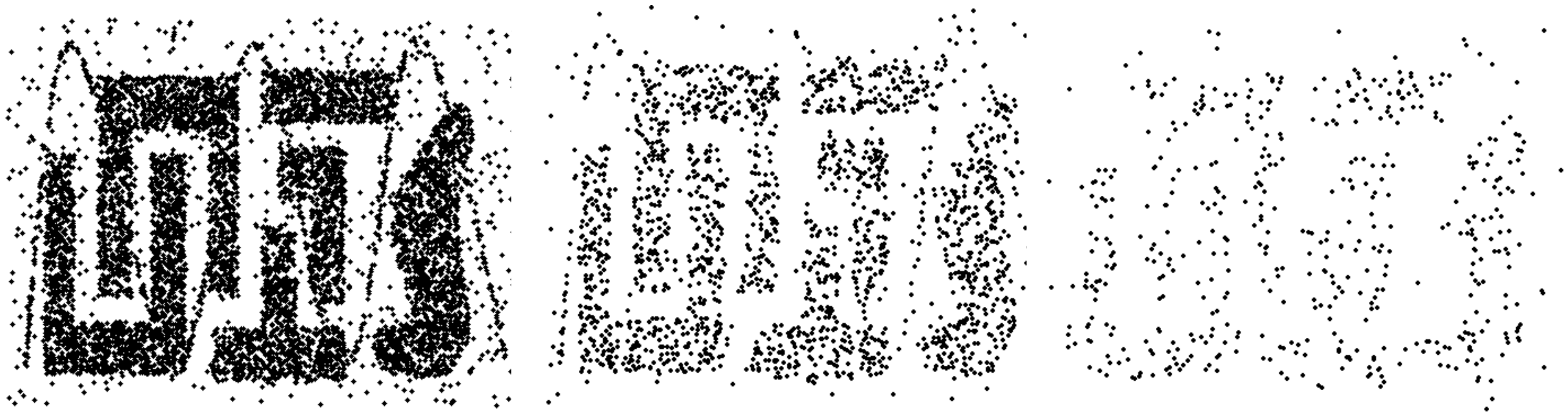
Sampling

- Sampling is the main technique employed for data reduction.
  - It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians often sample because obtaining the entire set of data of interest is too expensive or time consuming.
- Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming.

Sampling

- The key principle for effective sampling is the following:
  - Using a sample will work almost as well as using the entire data set, if the sample is representative
  - A sample is representative if it has approximately the same properties (of interest) as the original set of data

Sample size



8000 points   2000 Points   500 Points

Types of Sampling

- Simple Random Sampling
  - There is an equal probability of selecting any particular item
  - 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.
    - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

Curse of dimensionality

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

Dimensionality Reduction

- Purpose:
  - 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

Q4. Sample midterm

A. Why do we need to perform dimensionality reduction?

Addressing sparsity due to the curse of dimensionality, reduce redundancy, sometimes overfitting, etc.

B. Provide one approach to reduce dimensionality

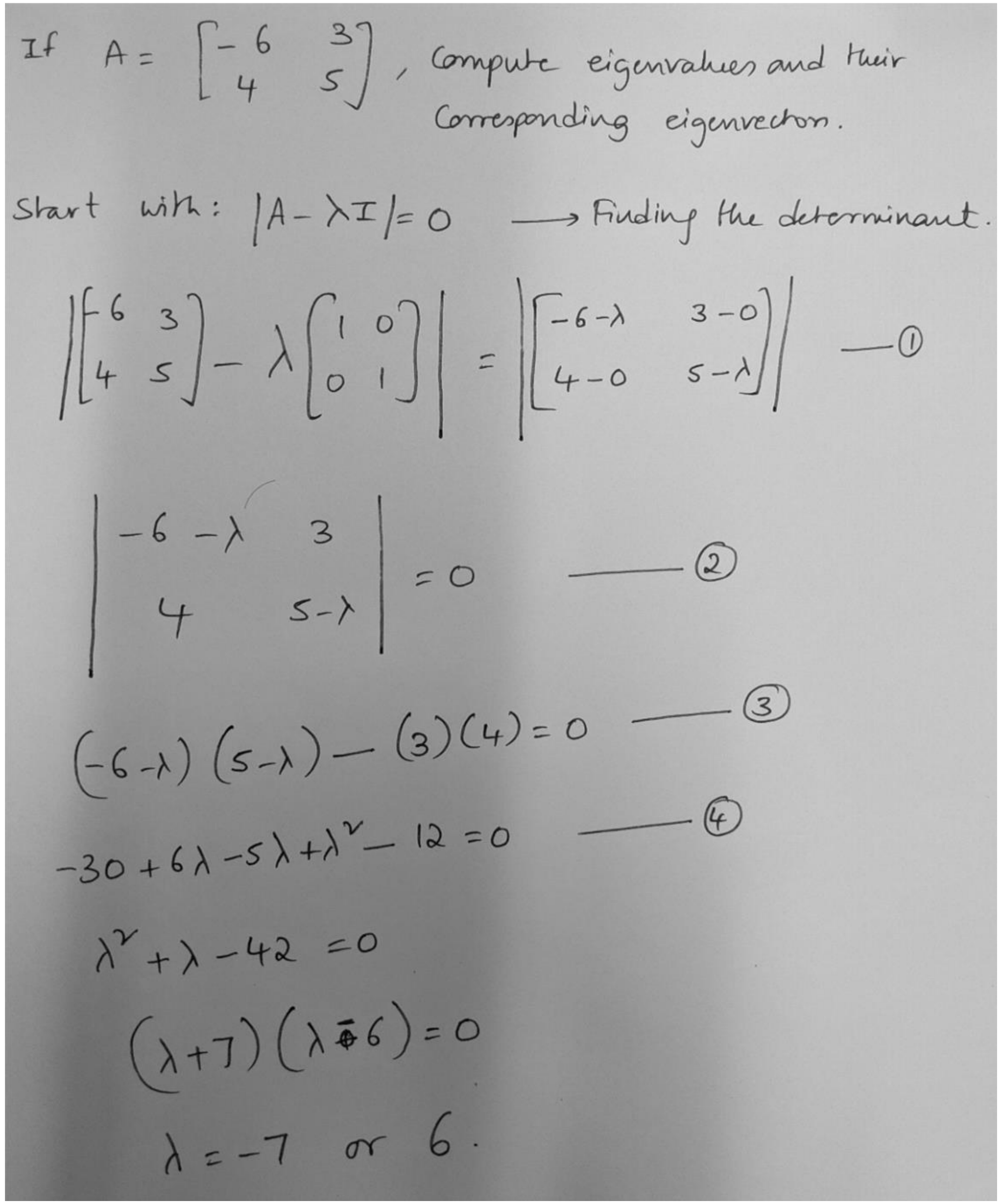
Principal Component Analysis

- Standardize; Covariance matrix; Compute eigenvalues and eigenvectors; top-K pcs; Map original data to new PCs space

Q4. Sample midterm

C. Most important eigenvector of the matrix

Most important eigenvector will be corresponding to eigenvalue=6



If $A = \begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix}$, Compute eigenvalues and their Corresponding eigenvectors.

Start with: $|A - \lambda I| = 0$ ⟶ Finding the determinant.

$$\left| \begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = \left| \begin{bmatrix} -6-\lambda & 3-0 \\ 4-0 & 5-\lambda \end{bmatrix} \right| \quad \text{— (1)}$$

$$\begin{vmatrix} -6-\lambda & 3 \\ 4 & 5-\lambda \end{vmatrix} = 0 \quad \text{— (2)}$$

$$(-6-\lambda)(5-\lambda) - (3)(4) = 0 \quad \text{— (3)}$$

$$-30 + 6\lambda - 5\lambda + \lambda^2 - 12 = 0 \quad \text{— (4)}$$

$$\lambda^2 + \lambda - 42 = 0$$

$$(\lambda + 7)(\lambda - 6) = 0$$

$$\lambda = -7 \text{ or } 6.$$

Q4. Sample midterm

Eigenvector with eigenvalue=6

Case 1: $\lambda = 6$:

$$AV = \lambda V$$

$$\begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 6 \begin{bmatrix} x \\ y \end{bmatrix} \quad (\rightarrow \text{Multiply})$$

$$\left.\begin{aligned} -6x + 3y &= 6x \\ 4x + 5y &= 6y \end{aligned}\right\} \text{——— } ①$$

$$-12x + 3y = 0$$

$$4x - y = 0$$

$$\Downarrow$$

$$4x = y \quad \text{or} \quad y = 4x.$$

So, Eigenvector is any non-zero multiple of

$$\begin{bmatrix} 1 \\ 4 \end{bmatrix}.$$

Feature subset Selection

- Another way to reduce dimensionality of data
- Redundant features
  - Duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - Contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA
- Many techniques developed, especially for classification

Feature Creation

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

- 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 and wavelet analysis

Discretization

- Discretization is the process of converting a continuous attribute into an ordinal attribute
  - 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

Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables
- Typically used for association analysis
- Often convert a continuous attribute to a categorical attribute and then convert a categorical attribute to a set of binary attributes
  - Association analysis needs asymmetric binary attributes
  - Examples: eye color and height measured as {low, medium, high}

Attribute Transformation

- An attribute transform is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k, $\log(x)$, e^x, $|x|$
  - Normalization
    - Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range
    - Take out unwanted, common signal, e.g., seasonality
  - In statistics, standardization refers to subtracting off the means and dividing by the standard deviation

Data Quality …

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data
  - Wrong data

Information and Probability

- Information relates to possible outcomes of an event
  - transmission of a message, flip of a coin, or measurement of a piece of data

- The more certain an outcome, the less information that it contains and vice-versa
  - For example, if a coin has two heads, then an outcome of heads provides no information
  - More quantitatively, the information is related the probability of an outcome
    - The smaller the probability of an outcome, the more information it provides and vice-versa
  - Entropy is the commonly used measure

Entropy

- For
  - a variable (event), X,
  - with n possible values (outcomes), $x_1, x_2 \ldots, x_n$
  - each outcome having probability, $p_1, p_2 \ldots, p_n$
  - the entropy of X , $H(X)$, is given by

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- Entropy is between 0 and $\log_2 n$ and is measured in bits
  - Thus, entropy is a measure of how many bits it takes to represent an observation of X on average

Entropy Examples

- For a coin with probability p of heads and probability $q = 1 - p$ of tails

$$H = -p \log_2 p - q \log_2 q$$

  - For p= 0.5, q = 0.5 (fair coin) $H = 1$
  - For $p = 1$ or $q = 1$, $H = 0$

- What is the entropy of a fair four-sided die ?

Entropy for Sample Data: Example

Hair Color	Count	p	$-p\log_2 p$
Black	75	0.75	0.3113
Brown	15	0.15	0.4105
Blond	5	0.05	0.2161
Red	0	0.00	0
Other	5	0.05	0.2161
Total	100	1.0	1.1540

Entropy for Sample Data

- Suppose we have
  - a number of observations (m) of some attribute, X, e.g., the gpa (assuming rounded values) of students in the class,
  - where there are n different possible values
  - And the number of observation in the i^{th} category is m_i
  - Then, for this sample

$$H(X) = -\sum_{i=1}^{n} \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

- For continuous data, the calculation is harder

Mutual Information

- Information one variable provides about another -- it quantifies the "amount of information" obtained about one random variable through observing the other random variable

Formally, $I(X,Y) = H(X) + H(Y) - H(X,Y)$, where

$H(X,Y)$ is the joint entropy of X and Y,

$$H(X,Y) = -\sum_i \sum_j p_{ij} \log_2 p_{ij}$$

Where p_{ij} is the probability that the i^{th} value of X and the j^{th} value of Y occur together

- For discrete variables, this is easy to compute

- Maximum mutual information for discrete variables is $\log_2(\min(n_X, n_Y))$, where n_X (n_Y) is the number of values of X (Y)

Mutual Information Example

Student Status	Count	p	$-p\log_2 p$
Undergrad	45	0.45	0.5184
Grad	55	0.55	0.4744
Total	100	1.00	0.9928

Grade	Count	p	$-p\log_2 p$
A	35	0.35	0.5301
B	50	0.50	0.5000
C	15	0.15	0.4105
Total	100	1.00	1.4406

Student Status	Grade	Count	p	$-p\log_2 p$
Undergrad	A	5	0.05	0.2161
Undergrad	B	30	0.30	0.5211
Undergrad	C	10	0.10	0.3322
Grad	A	30	0.30	0.5211
Grad	B	20	0.20	0.4644
Grad	C	5	0.05	0.2161
Total		100	1.00	2.2710

Mutual information of Student Status and Grade = 0.9928 + 1.4406 - 2.2710 = 0.1624

Similarity and Dissimilarity Measures

- Similarity measure
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity measure
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

Similarity/Dissimilarity for Simple Attributes

The following table shows the similarity and dissimilarity between two objects, *x* and *y*, with respect to a single, simple attribute.

Attribute Type	**Dissimilarity**	**Similarity**
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n - 1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Euclidean Distance

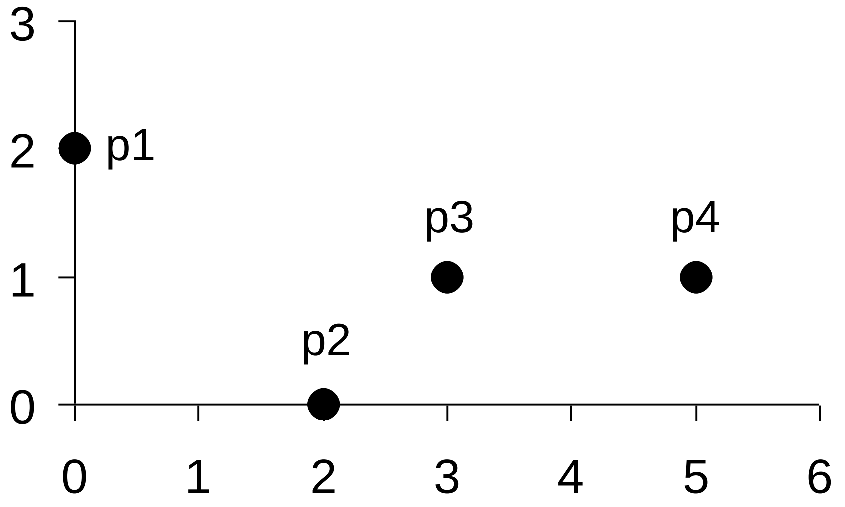
- Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

where n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\mathbf{x}$ and $\mathbf{y}$.

- Standardization is necessary, if scales differ.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	**p2**	**p3**	**p4**
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\boldsymbol{x}$ and $\boldsymbol{y}$.

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors

- $r = 2$. Euclidean distance

- $r \to \infty$. “supremum” (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component of the vectors

- Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of dimensions.

Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

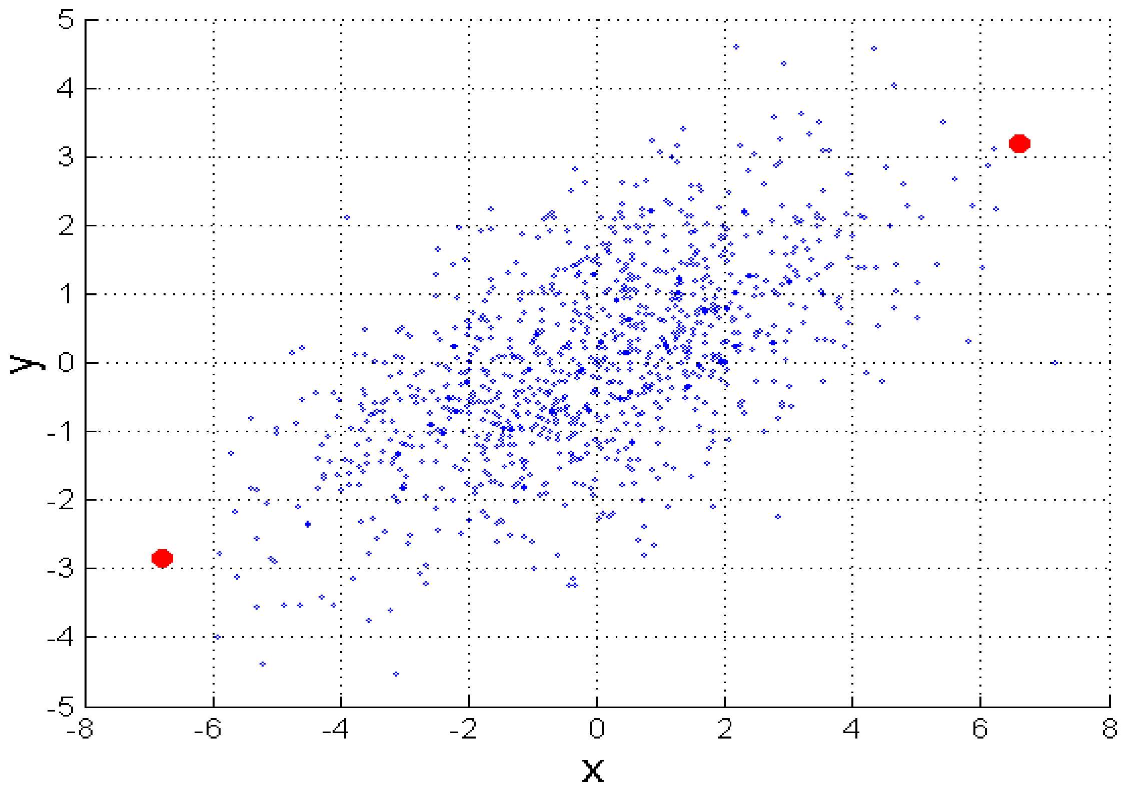
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

Mahalanobis Distance

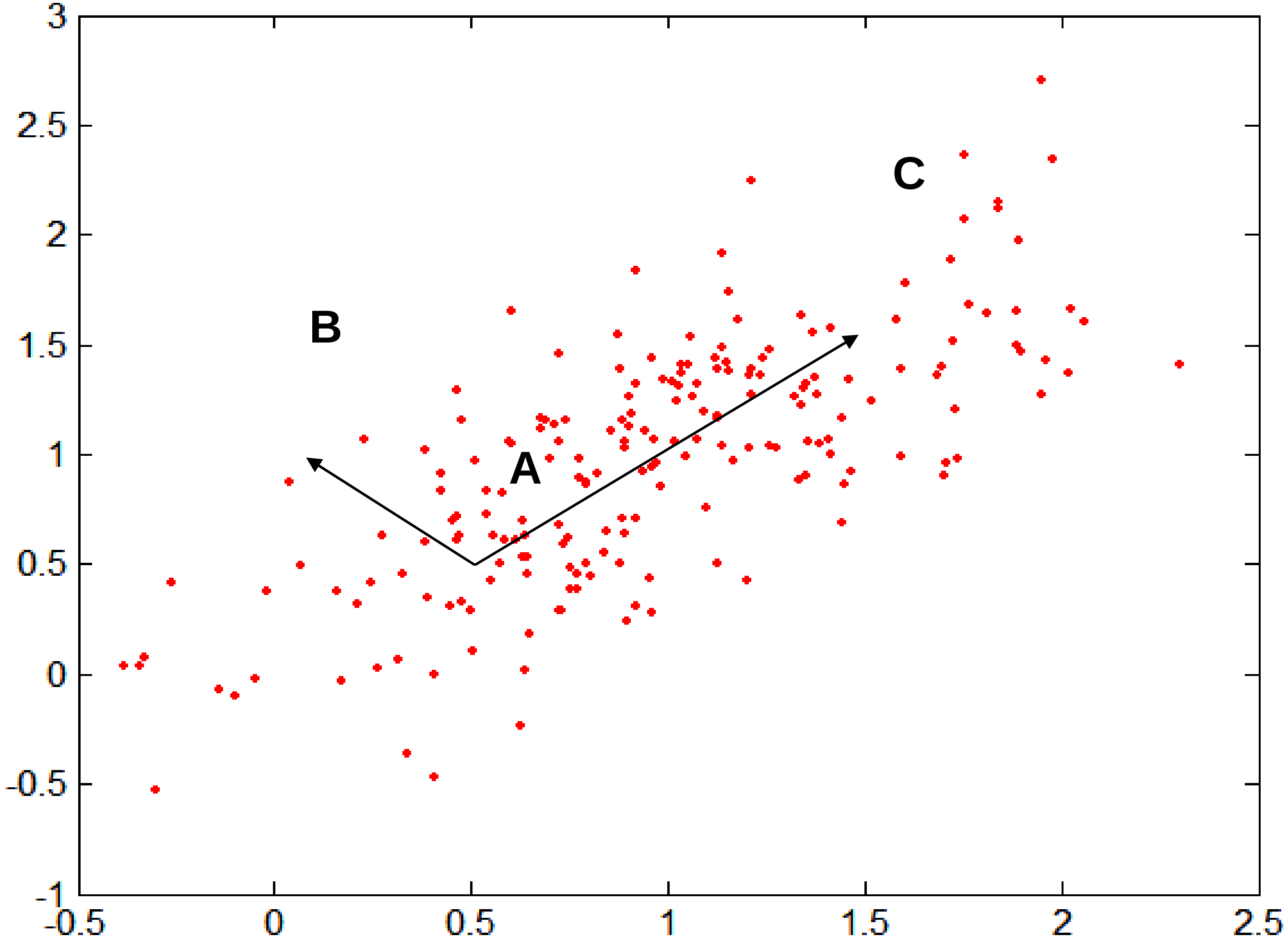
$$\mathbf{mahalanobis}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \Sigma^{-1} (\mathbf{x} - \mathbf{y})$$



Σ is the covariance matrix

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Q1. Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

  1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all x and y and $d(\mathbf{x}, \mathbf{y}) = 0$ only if $\mathbf{x} = \mathbf{y}$. (Positive definiteness)
  2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)
  3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. (Triangle Inequality)

  where $d(\mathbf{x}, \mathbf{y})$ is the distance (dissimilarity) between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.

- A distance that satisfies these properties is a metric

Common Properties of a Similarity

- Similarities, also have some well known properties.

1. $s(\mathbf{x}, \mathbf{y}) = 1$ (or maximum similarity) only if $\mathbf{x} = \mathbf{y}$.

2. $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)

where $s(\mathbf{x}, \mathbf{y})$ is the similarity between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.

Similarity Between Binary Vectors

- Common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities using the following quantities
  f_{01} = the number of attributes where *p* was 0 and *q* was 1
  f_{10} = the number of attributes where *p* was 1 and *q* was 0
  f_{00} = the number of attributes where *p* was 0 and *q* was 0
  f_{11} = the number of attributes where *p* was 1 and *q* was 1
- Simple Matching and Jaccard Coefficients
  SMC = number of matches / number of attributes
  $= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

  J = number of 11 matches / number of non-zero attributes
  $= (f_{11}) / (f_{01} + f_{10} + f_{11})$

SMC versus Jaccard: Example

x = 1 0 0 0 0 0 0 0 0 0
y = 0 0 0 0 0 0 1 0 0 1

$f_{01} = 2$ (the number of attributes where *p* was 0 and *q* was 1)
$f_{10} = 1$ (the number of attributes where *p* was 1 and *q* was 0)
$f_{00} = 7$ (the number of attributes where *p* was 0 and *q* was 0)
$f_{11} = 0$ (the number of attributes where *p* was 1 and *q* was 1)

SMC $= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$
$= (0+7) / (2+1+0+7) = 0.7$

$J = (f_{11}) / (f_{01} + f_{10} + f_{11}) = 0 / (2 + 1 + 0) = 0$

Cosine Similarity

- If $\mathbf{d}_1$ and $\mathbf{d}_2$ are two document vectors, then

$$\cos(\mathbf{d_1}, \mathbf{d_2}) = \langle \mathbf{d_1}, \mathbf{d_2} \rangle / ||\mathbf{d}_1|| \; ||\mathbf{d}_2|| \, ,$$

where $\langle \mathbf{d_1}, \mathbf{d_2} \rangle$ indicates inner product or vector dot product of vectors, $\mathbf{d_1}$ and $\mathbf{d_2}$, and $||\ \mathbf{d}\ ||$ is the length of vector $\mathbf{d}$.

- Example:

$$\mathbf{d_1} = \mathbf{3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0}$$

$$\mathbf{d_2} = \mathbf{1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2}$$

$\langle \mathbf{d_1}, \mathbf{d2} \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$|\ \mathbf{d_1}\,|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{\mathbf{0.5}} = (42)^{\mathbf{0.5}} = 6.481$

$||\ \mathbf{d_2}\,|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{\mathbf{0.5}} = (6)^{\mathbf{0.5}} = 2.449$

$\cos(\mathbf{d_1}, \mathbf{d_2}) = 0.3150$

Q1. Edit distance

S1 = “cat”;
S2 = “dog”

3 operations of replace a character in string S2. Edit distance of 3.

Common operations allowed for this purpose include:

a. insert a character into a string;
b. delete a character from a string;
c. replace a character in the string.

Linear Regression

The technique is used to **predict** the value of one variable (the dependent variable - y) **based on** the value of other variables (independent variables x_1, x_2,…x_k)

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Modeling

- The first order linear model

y = dependent variable

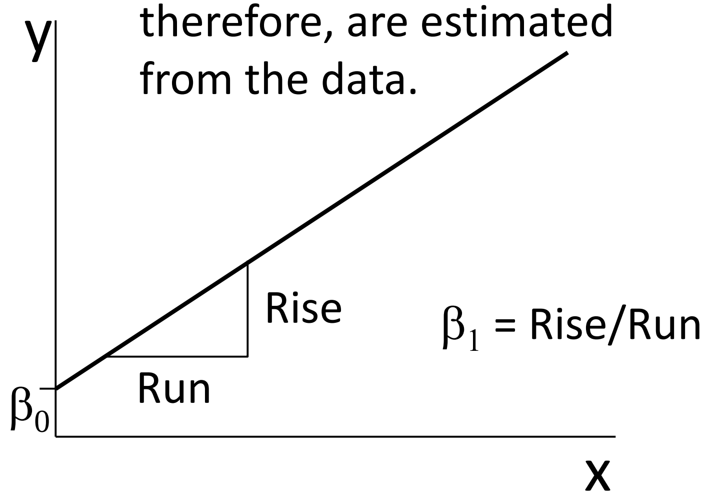
x = independent variable

β_0 = y-intercept

β_1 = slope of the line

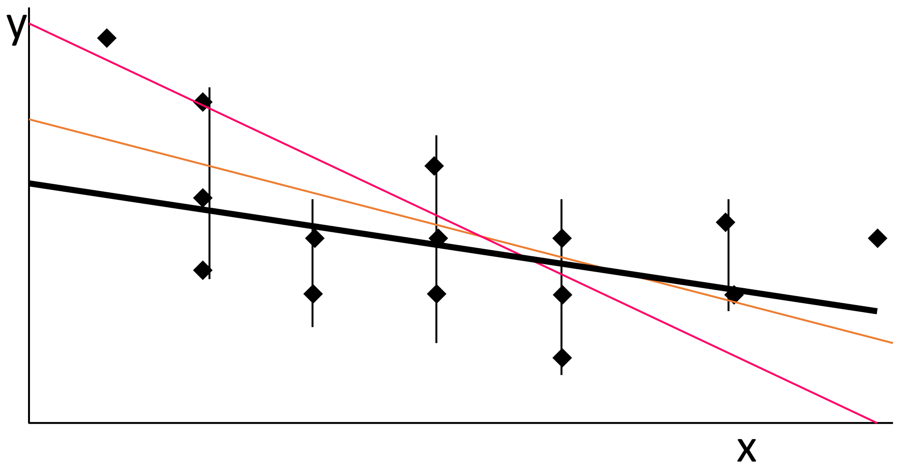
$\mathcal{E}$ = error variable

β_0 and β_1 are unknown, therefore, are estimated from the data.



Estimating the coefficients

- The estimates are determined by
  - drawing a sample from the population of interest,
  - calculating sample statistics.
  - producing a straight line that cuts into the data.



The question is:
Which straight line fits best?

Q5. Best regression line

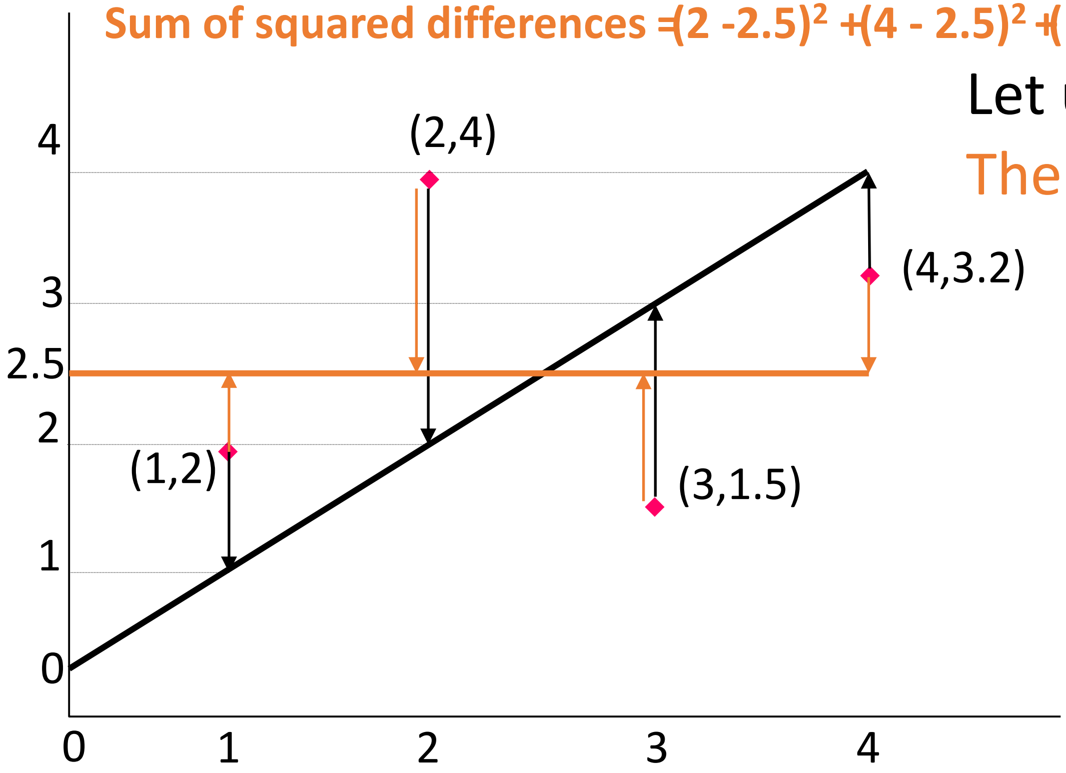
The best line is the one that minimizes the sum of squared vertical differences between the points and the line.

Sum of squared differences = $(2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$

Sum of squared differences = $(2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 - 2.5)^2 = 3.99$

Let us compare two lines

The second line is horizontal



The smaller the sum of squared differences the better the fit of the line to the data.

Logistic Regression

- Special case of linear regression where the target variable is categorical in nature
- Uses a log of odds as a dependent variable
- Predicts the probability of occurrence of an event using a sigmoid function (inverse of logit function)

$$p = 1/(1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2)})$$

Linear vs Logistic Regression

- Output for linear regression is continuous
  - For example, stock prices
  - Or real estate price estimation
- Output for logistic regression is estimated as a constant
  - For example, predicting if a sample is tested +ve or –ve
  - Output >0.5 is +ve or 1 or yes; ouput <=0.5 is –ve or 0 or no

Linear vs Logistic Regression

- Linear regression is estimated using ordinary least squares
  - Distance minimizing approximation approach
  - Fits a regression line on a given set of data points that has the minimum sum of squared deviations (least squared error)

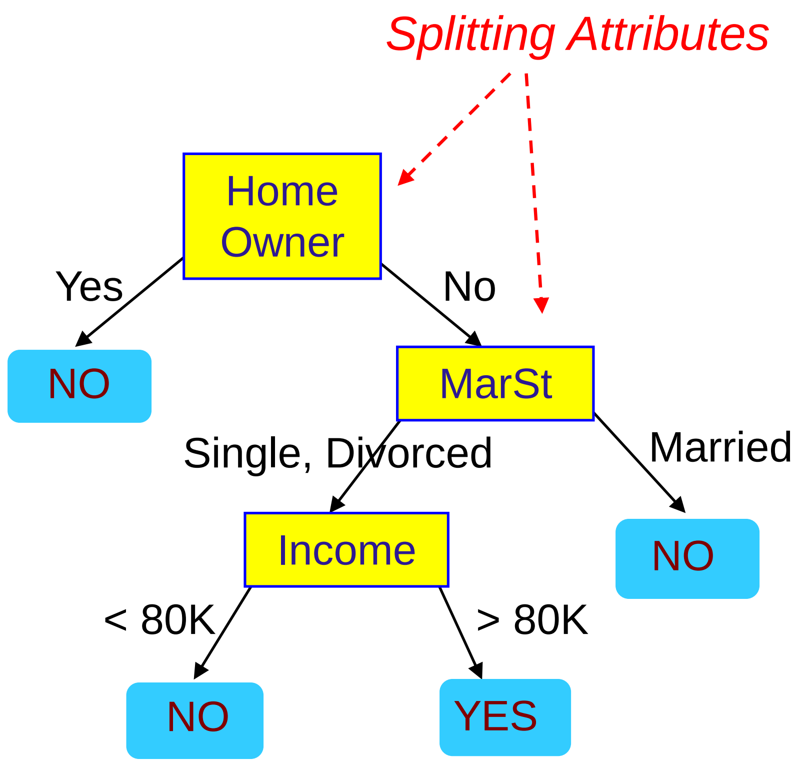
- Logistic regression is estimated using maximum likelihood estimation
  - “Likelihood” maximization method
  - Determines parameters (such as mean/variance) that are most likely to produce the set of data points.

Example of a Decision Tree

categorical categorical continuous class

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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

Training Data

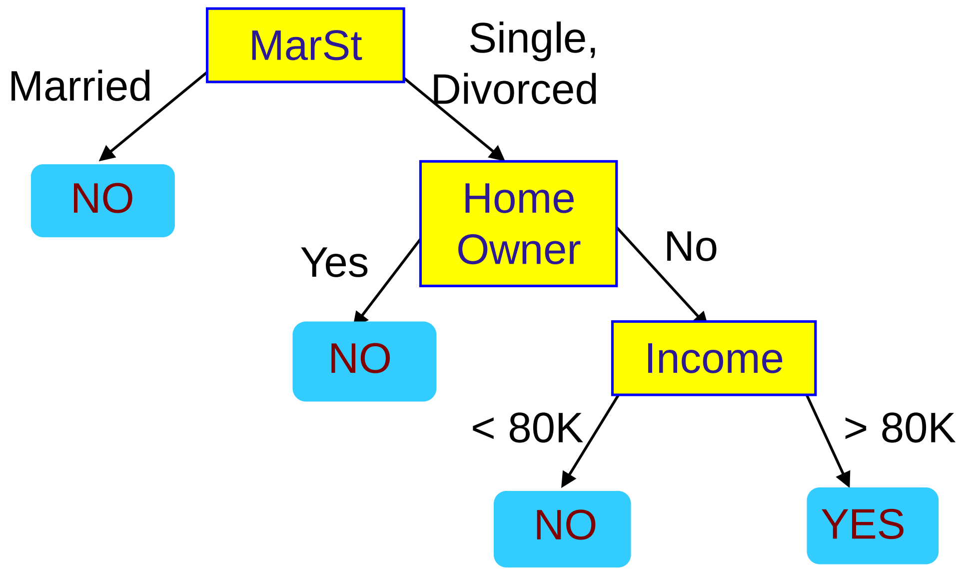


Model: Decision Tree

Another Example of Decision Tree

categorical categorical continuous class

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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**



There could be more than one tree that fits the same data!

Design Issues of Decision Tree Induction

- How should training **records be split**?
  - Method for specifying test condition
    - depending on attribute types
  - Measure for evaluating the goodness of a test condition

- How should the **splitting procedure stop**?
  - Stop splitting if all the records belong to the same class or have identical attribute values
  - Early termination

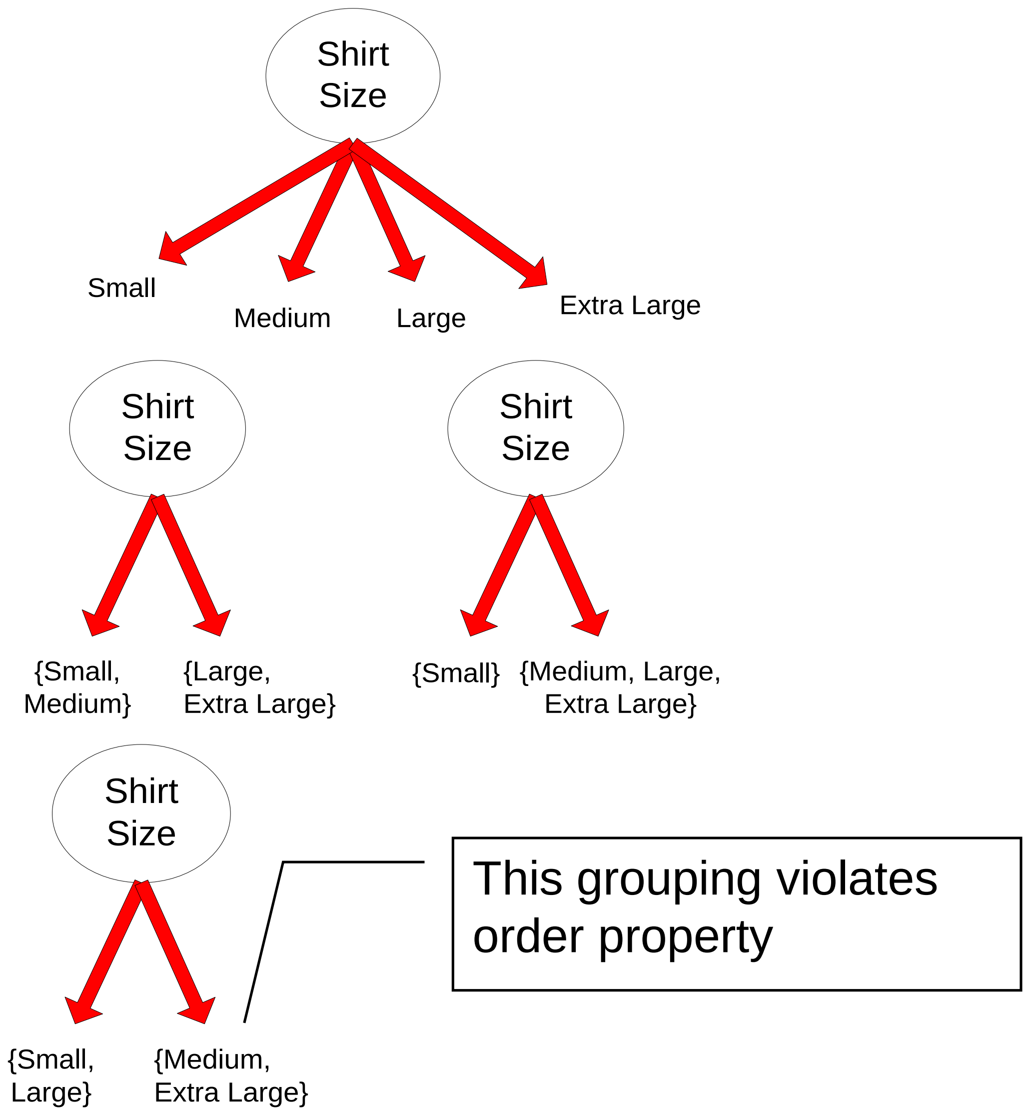
Methods for Expressing Test Conditions

- Depends on attribute types
  - Binary
  - Nominal
  - Ordinal
  - Continuous

- Depends on number of ways to split
  - 2-way split
  - Multi-way split

Test Condition for Ordinal Attributes

- Multi-way split:
  - Use as many partitions as distinct values.
- Binary split:
  - Divides values into two subsets
  - Preserve order property among attribute values



How to determine the best split

- Greedy approach:
  - Nodes with purer class distribution are preferred

- Need a measure of node impurity:

C0: 5 C1: 5	C0: 9 C1: 1
High degree of impurity	Low degree of impurity

Measures of Node Impurity

- Gini Index

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

- Entropy

$$Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$$

- Misclassification error

$$Error(t) = 1 - \max_i P(i \mid t)$$

Finding the best split

1. Compute impurity measure (P) before splitting
2. Compute impurity measure (M) after splitting
    1. Compute impurity measure of each child node
    2. M is the weighted impurity of children
3. Choose the attribute test condition that produces the highest gain

$$\text{Gain} = P - M$$

or equivalently, lowest impurity measure after splitting (M)

Measure of Impurity: Entropy

- Entropy at a given node t:

$$Entropy(t) = -\sum_{j} p(j\,|\,t) \log p(j\,|\,t)$$

  - (NOTE: $p(j \mid t)$ is the relative frequency of class j at node t).
  - Maximum ($\log n_c$) when records are equally distributed among all classes implying least information
  - Minimum (0.0) when all records belong to one class, implying most information

- Entropy based computations are quite similar to the GINI index computations

Computing Entropy of a Single Node

$$Entropy(t) = -\sum_j p(j|t)\log_2 p(j|t)$$

C1	**0**
C2	**6**

P(C1) = 0/6 = 0   P(C2) = 6/6 = 1

Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

C1	**1**
C2	**5**

P(C1) = 1/6   P(C2) = 5/6

Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (5/6) = 0.65

C1	**2**
C2	**4**

P(C1) = 2/6   P(C2) = 4/6

Entropy = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

Computing Information Gain after Splitting

- Information Gain

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

Parent Node, p is split into k partitions; n_i is number of records in partition i

- Choose the split that achieves most reduction (maximizes GAIN)
- Used in ID3 and C4.5 decision tree algorithms

Q6. Sample Midterm – Decision Trees

a. 10 transactions with

Class labels {Yes, No}

I(Y, N) = I(7, 3) = 0.98

P(<=5) = 3/10

P(5..10) = 3/10

P(>10) = 4/10

Attribute-A	Y_i	N_i
<=5	3	0
5..10	1	2
>10	2	2

Entropy_attr-A = 0.68

Info_gain_attr-A = 0.3

Q6. Sample midterm – Decision Trees

a. b. 10 transactions with

Class labels {Yes, No}

Attribute-B	Y_i	N_i
Yes	3	4
No	1	2

I(Y, N) = I(7, 3) = 0.98

Entropy_attr-B = 0.97

P(Yes) = 7/10

P(No) = 3/10

Info_gain_attr-B = 0.01