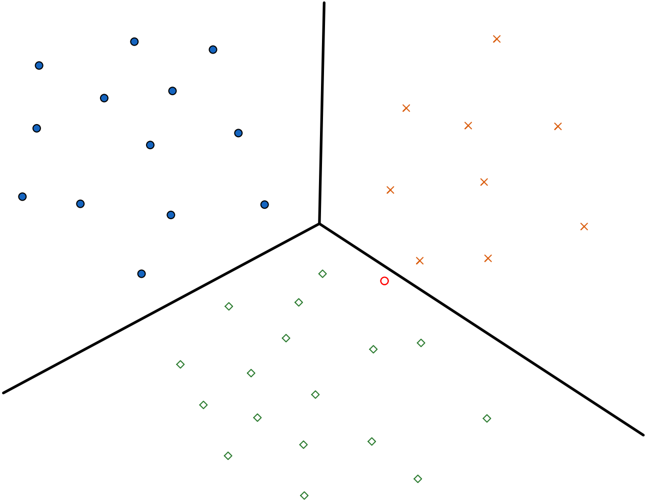
# BUAN 3500: Data Visualization and Descriptive Analytics in Business

## Descriptive Data Mining

Lauren M. Nelsen, Ph.D.

University of Colorado Colorado Springs

# Classification



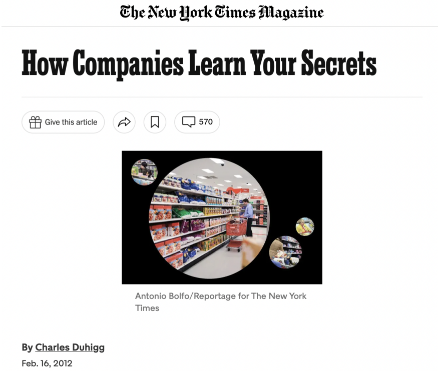
# Classification

**Examples of situations where classifying is helpful:**

- **Medical Trials**
  - Want to predict if someone's disease will be cured by a particular medical treatment
- **College Admission**
  - Want to predict whether or not someone will be admitted to a particular college (maybe based on something like GPA)

# Classification

- **Customer segmentation**
  - Marketing wants to know how to target advertising to particular groups of customers



The New York Times Magazine

How Companies Learn Your Secrets

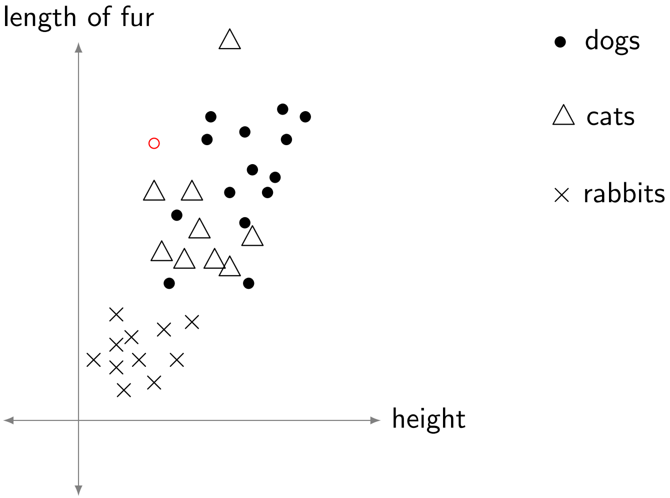
Antonio Bolfo/Reportage for The New York Times

By Charles Duhigg

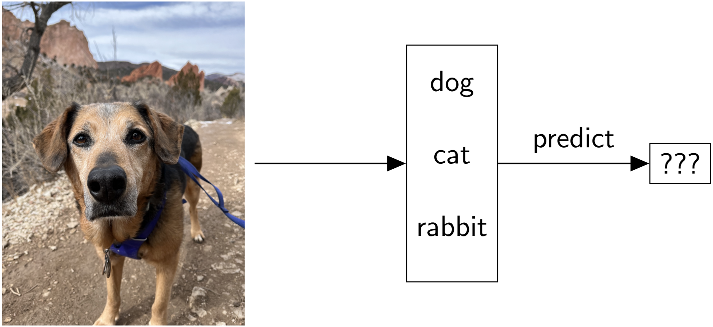
Feb. 16, 2012

# Classification

We want to be able to predict qualitative information.



# Classification



dog

cat

rabbit

predict

???

# Classification

**Question:** Why do we not want to use linear regression when we have qualitative response variables?

**Example:** We could try letting our response variable take on the corresponding values:

$$Y = \begin{cases} 1 & \text{Dog} \\ 2 & \text{Cat} \\ 3 & \text{Rabbit} \end{cases}.$$

How far apart are "Cat" and "Rabbit"? What about "Dog" and "Cat"?

# Classification

- We need methods for classifying our data.

# Cluster Analysis

Two commonly used clustering methods:

- ***k*-means clustering**
- **hierarchical clustering**

Both of these are examples of **unsupervised machine learning**. → the idea of a “good” clustering is subjective and depends on what the analyst hopes to learn

# Cluster Analysis

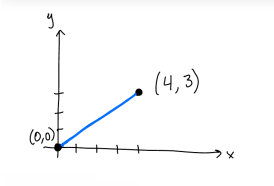
**Question:** What does it mean for observations to be “close” or “similar”? We need to be able to measure “similary/dissimilarity” between observations.

→ We need a way of calculating distance between observations.

# Cluster Analysis

Examples of ways to measure distance:

- **Euclidean distance:**



If $u = (u_1, u_2, ..., u_n)$ and $v = (v_1, v_2, ..., v_n)$, then

$$d_{uv}^{\text{euclid}} = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_n - v_n)^2}$$

(Note: We often want to use "standardized distance" instead.)

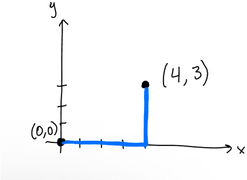
**Example**

Let's look at a problem on the Chapter 6 homework assignment related to this.

# Cluster Analysis

Examples of ways to measure distance:

- **Manhattan distance:**



# Cluster Analysis

**k-means clustering:**

- iteratively assigns each observation to one of *k* clusters in an attempt to achieve clusters that contain observations as similar to each other as possible

# Cluster Analysis

**k-means clustering:**

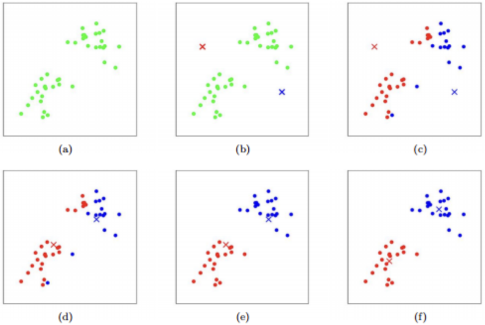


*Figure 1: K-means algorithm. Training examples are shown as dots, and cluster centroids are shown as crosses. (a) Original dataset. (b) Random initial cluster centroids. (c-f) Illustration of running two iterations of k-means. In each iteration, we assign each training example to the closest cluster centroid (shown by "painting" the training examples the same color as the cluster centroid to which is assigned); then we move each cluster centroid to the mean of the points assigned to it. Images courtesy of Michael Jordan.*

(https://stanford.edu/ cpiech/cs221/handouts/kmeans.html)

# Cluster Analysis

**k-means clustering:**

Tableau uses *k*-means clustering!

https://help.tableau.com/current/pro/desktop/en-us/clustering.htm

## Example

Use Tableau to do *k*-means clustering in the following examples:

- Salary Data
- Online Retail Data

https://archive.ics.uci.edu/dataset/352/online+retail

# Cluster Analysis

**k-means clustering:**

Example

Let's look at a problem on the Chapter 6 Assignment related to this.

# Cluster Analysis

**k-means clustering:**

Some disadvantages:

- Choosing $k$ manually.
- Being dependent on initial values.
- Clustering data of varying sizes and density.
- Clustering outliers.

# Cluster Analysis

**hierarchical clustering:**

- starts with each observation belonging to its own cluster and then sequentially merges the most similar clusters to create a series of nested clusters

# Cluster Analysis

## hierarchical clustering:

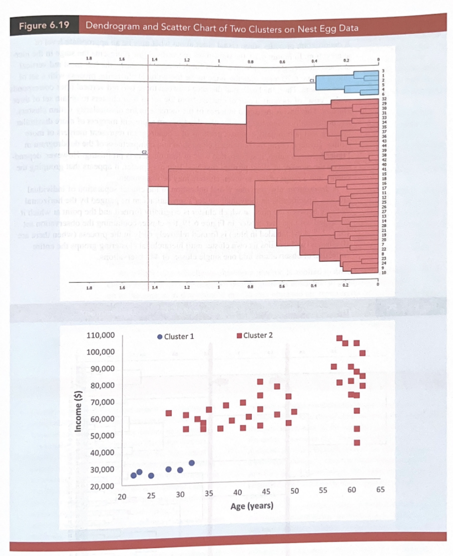


**Figure 6.19** Dendrogram and Scatter Chart of Two Clusters on Nest Egg Data

# Cluster Analysis

**hierarchical clustering:**

### Example

Let's look at problem 1(a) from the Chapter 5 homework assignment.

# Cluster Analysis

**Hierarchical Clustering versus *k*-Means Clustering:**

Hierarchical clustering is a good choice when:

- you want to easily examine solutions with a wide range of clusters
- you want to observe how clusters are nested

Downsides to hierarchical clustering:

- hierarchical clustering can be very sensitive to outliers
- clusters may change dramatically if observations are eliminated from or added to the data set
- hierarchical clustering may be a less appropriate option as the number of the observations in the data set grows large because it is relatively computationally expensive

# Cluster Analysis

**Hierarchical Clustering versus *k*-Means Clustering:**

- *k*-Means is a good choice when you want to cluster data on the basis of numerical variables
- *k*-Means is computationally efficient enough to handle a large number of observations
- *k*-means is generally not appropriate for categorical or ordinal data, where it doesn't make sense to talk about an "average"

No matter what clustering method is used... it is up to the analyst to determine if the clusters obtained are actually meaningful and provide insight! (They might just be relatively arbitrary groupings of observations.)

# Association Rules

## Definition

**association rules:** probabilistic if–then statements which convey the likelihood of certain items being purchased together

- an important tool in **market basket analysis**
- also applicable outside of marketing
  - can assist medical researchers in understanding which treatments have been commonly prescribed to certain patient symptoms

# Association Rules

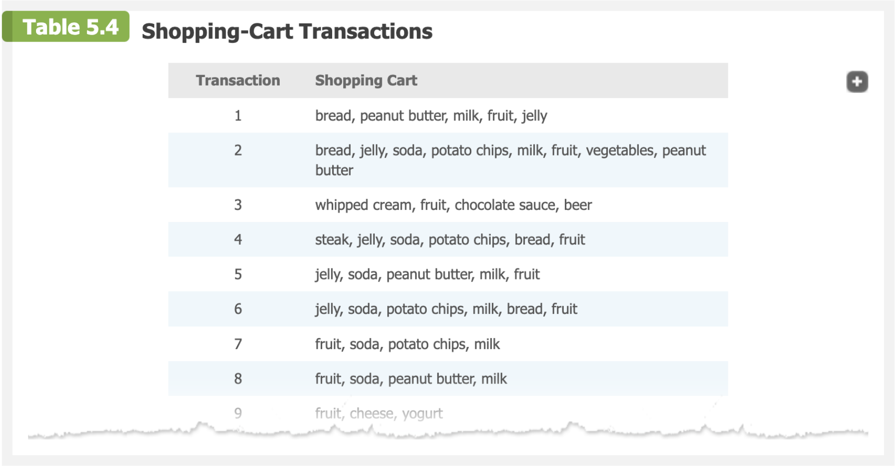


**Table 5.4** **Shopping-Cart Transactions**

| Transaction | Shopping Cart |
|---|---|
| 1 | bread, peanut butter, milk, fruit, jelly |
| 2 | bread, jelly, soda, potato chips, milk, fruit, vegetables, peanut butter |
| 3 | whipped cream, fruit, chocolate sauce, beer |
| 4 | steak, jelly, soda, potato chips, bread, fruit |
| 5 | jelly, soda, peanut butter, milk, fruit |
| 6 | jelly, soda, potato chips, milk, bread, fruit |
| 7 | fruit, soda, potato chips, milk |
| 8 | fruit, soda, peanut butter, milk |
| 9 | fruit, cheese, yogurt |

## Definition

**itemset:** collection of items

# Association Rules

*if* portion of the rule → *then* portion of the rule

**antecedent** **consequent**

**Example:**

*if* {bread, jelly} → *then* {peanut butter}

**antecedent** **consequent**

# Association Rules

The number of possible association rules can be overwhelming.
→ We typically investigate only association rules that involve antecedent and consequent itemsets that occur together frequently.

What is “frequent”?

# Association Rules

## Definition

**support** of an itemset: percentage of transactions in the data that include that itemset.

**Example:** Shopping basket transactions

| **Transaction** | **Shopping Cart** |
|---|---|
| 1 | bread, socks, eggs, broccoli |
| 2 | milk, bananas, towels, celery |
| 3 | bread, cheese, yogurt |
| 4 | yogurt, eggs, broccoli, bread |
| 5 | socks, towels, eggs, milk |

support of the itemset {socks, eggs}: $\frac{2}{5} = .4$

# Association Rules

- By only considering rules involving itemsets with large enough support, we can generally avoid inexplicable rules capturing random noise in the data.
  → Rule of thumb: consider only association rules with a support of at least 20% of the total number of transactions.

- Property of a reliable association rule:
  → given a transaction contains the antecedent itemset, there is a high probability that it contains the consequent itemset
  → this is called the **confidence** of a rule

# Association Rules

**Confidence:**

$$P(\text{consequent}|\text{antecedent}) = \frac{P(\text{consequent and antecedent})}{P(antecedent)}$$
$$= \frac{\text{support of }\{\text{consequent and antecedent}\}}{\text{support of antecedent}}$$

# Association Rules

We also want to be able to measure how effective an association rule is at identifying transactions in which the consequent itemset occurs versus a randomly selected transaction.

**Lift Ratio:**

$$\begin{aligned}\frac{P(\text{consequent}|\text{antecedent})}{P(\text{consequent})} &= \frac{P(\text{consequent and antecedent})}{P(\text{consequent}) \times P(\text{antecedent})} \\ &= \frac{\text{confidence of rule}}{\text{support of consequent}}\end{aligned}$$

# Association Rules

## Example

**Association Rules for Hy-Vee:**
Let's apply these ideas to the Hy-Vee data files on Canvas.

# Association Rules

## Example

If we have time, we can see how to implement the **Apriori algorithm** to automate the process for analyzing transaction data in Python.

https://bit.ly/45M037i

# Text Mining

We're going to talk about some of the ideas from this section in the context of an example in Python.

But first, we need to discuss another classification method!

# Classification

- Now we're going to talk about a classification method designed for a **binary response variable**.

$$\rightarrow \boxed{\textbf{Logistic Regression}}$$

# Logistic Regression

Logistic regression is a classification method designed for a **binary response**.

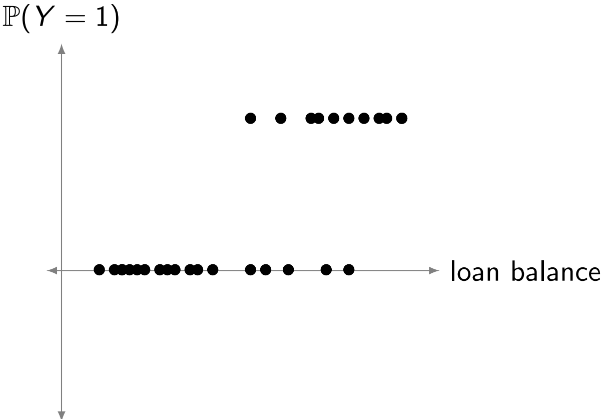
**Examples:**

- A company has many customer reviews, and they want to be able to quickly categorize each review as positive or negative.
  - positive: 1
  - negative: 0
- An organization wants to look at factors that influence whether or not a customer will renew their membership.
  - renews: 1
  - does not renew: 0
- You want to be able to predict whether or not a movie will be profitable.
  - profitable: 1
  - not profitable: 0

# Logistic Regression

**Example:** A bank wants to be able to predict if a customer will default on a loan based on the balance of their loan.

$$Y = \begin{cases} 1 & \text{if customer defaults on loan} \\ 0 & \text{if customer does not default on loan} \end{cases}$$

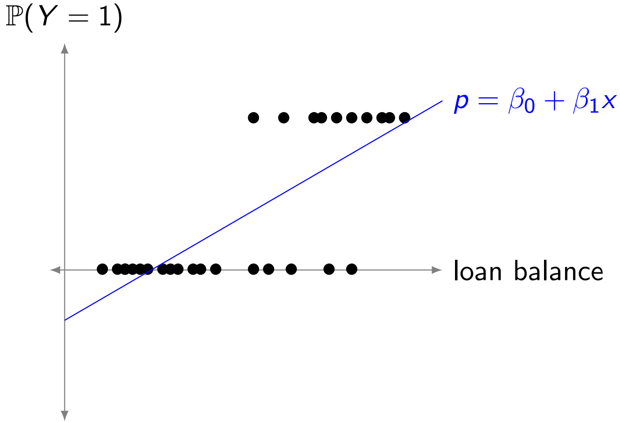


# Logistic Regression

**linear regression:**

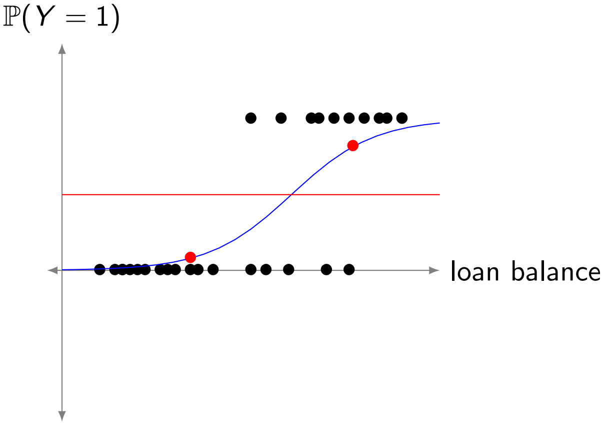
$$y = p = \beta_0 + \beta_1 x$$

Why isn't this good for categorical responses?



# Logistic Regression

It would be nice to have a function like this:



# Logistic Regression

**logistic regression:**

$$\boxed{???} = \beta_0 + \beta_1 x$$

- $p \leftarrow$ only in $[0, 1]$
- $\dfrac{p}{1-p} \leftarrow$ in $[0, \infty)$
- $\log\left(\dfrac{p}{1-p}\right) \leftarrow$ in $(-\infty, \infty)$

$\log\left(\dfrac{p}{1-p}\right)$ is called “log odds” or “logit($p$)”

# Logistic Regression

$$\boxed{\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x}$$

**Question:** How would we solve this for $p$?

$$\begin{aligned}
\left(\frac{p}{1-p}\right) &= e^{\beta_0+\beta_1 x} \\
p &= (1-p)e^{\beta_0+\beta_1 x} \\
p &= e^{\beta_0+\beta_1 x} - pe^{\beta_0+\beta_1 x} \\
p + pe^{\beta_0+\beta_1 x} &= e^{\beta_0+\beta_1 x} \\
p(1 + e^{\beta_0+\beta_1 x}) &= e^{\beta_0+\beta_1 x} \\
p &= \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \\
p &= \frac{1}{e^{-(\beta_0+\beta_1 x)} + 1}
\end{aligned}$$

# Logistic Regression

$$\boxed{\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x}$$

$$\boxed{p(x) = \frac{1}{e^{-(\beta_0+\beta_1 x)} + 1}}$$

## Example

Let's see if this looks the way we want it to in an example:
Use the following link or QR code:

`https://bit.ly/414tgZP`

# Logistic Regression

**In General:**

$$\boxed{\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}$$

$$\boxed{p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}}}$$

Assumptions:

- The $X_i$ are independent
- No multicollinearity
- Linearity between the independent variables and log odds
- Large sample size

# Logistic Regression

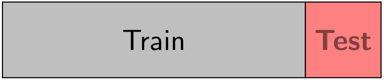
Now we're going to use Python to make predictions using logistic regression!

## Sentiment Analysis Example: Classifying Product Reviews

https://bit.ly/3Qwmvx0

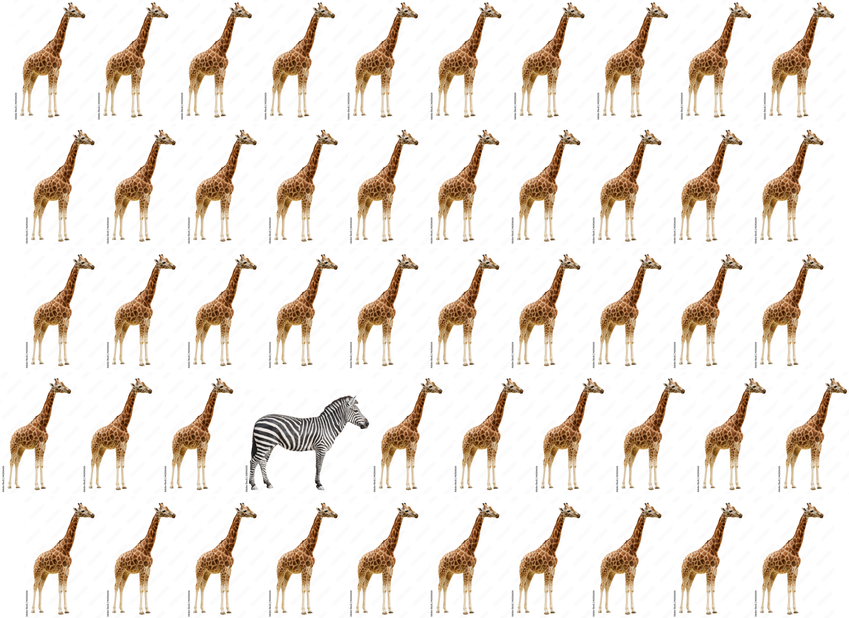
# Testing a Model

- **Validation Set Approach:**



Train

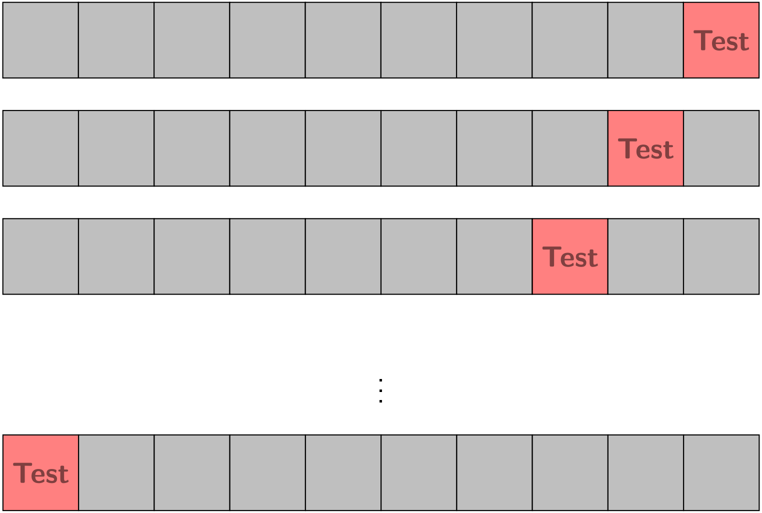
**Test**



# Testing a Model

- **Cross-Validation:**

# Testing a Model: Cross Validation



Test

Test

Test

Test

# Testing a Model: Cross Validation

**Question:** Why is this better than the validation set approach?

→ There is less bias. (It was trained on more data.)

# Supervised vs. Unsupervised Learning

- **Using Machine Learning to Make Predictions**
  → logistic regression is an example of supervised learning

- **Unsupervised Learning**

# Other Classification Models

There are many other classification models that are helpful in different scenarios!

# Visualizing Text Data

Sometimes a **Word Cloud** can be helpful for visualizing text data.

### Example

Let's see how to create a Word Cloud in Tableau using the "OneWordReviews" file and see when it might be helpful.