Statistics and Data Analysis

A Brief Introduction to Data Mining

Ling-Chieh Kung

Department of Information Management
National Taiwan University

Data mining

- **Data mining** is about **efficiently** extracting information from data.
- The focus is different from statistics.
  - In statistics, we mainly care about **inference**: Using the information obtained from a sample to infer some hidden facts in a population.
  - In data mining, we mainly care about **computation**: Given a huge data set (maybe representing the population), we do calculations to identify facts.
  - The boundary is of course somewhat vague.
- Three major topics in data mining:
  - Classification.
  - Association.
  - Clustering.

Road map

- **Classification: logistic regression**.
- Association: frequent pattern mining.
- Clustering: the k-means algorithm.

Classification

- A very typical problem is detecting spam mails.
  - Each mail is either a spam mail or not a spam mail.
  - Each mail has some **features**, e.g., the number of times that "money" appears.
  - Given a lot of past mails that **have been classified** as spam or not spam, may we build a model to **classify** the next mail?
- This is a **classification** problem.
- We may consider a classification problem as a regression problem:
  - Each feature is an **independent variable**.
  - The **dependent variable** is the class an observation belongs to.
  - We want to build a formula to do the classification.

Logistic regression

- So far our regression models always have a **quantitative** variable as the **dependent** variable.
  - Some people call this type of regression **ordinary regression**.
- To have a **qualitative** variable as the dependent variable, ordinary regression does not work.
- One popular remedy is to use **logistic regression**.
  - In general, a logistic regression model allows the dependent variable to have multiple levels.
  - We will only consider **binary variables** in this lecture.
- Let's first illustrate why ordinary regression fails when the dependent variable is binary.

Example: survival probability

- 45 persons got trapped in a storm during a mountain hiking. Unfortunately, some of them died due to the storm.[1]
- We want to study how the **survival probability** of a person is affected by her/his **gender** and **age**.

Age	Gender	Survived	Age	Gender	Survived	Age	Gender	Survived
23	Male	No	23	Female	Yes	15	Male	No
40	Female	Yes	28	Male	Yes	50	Female	No
40	Male	Yes	15	Female	Yes	21	Female	Yes
30	Male	No	47	Female	No	25	Male	No
28	Male	No	57	Male	No	46	Male	Yes
40	Male	No	20	Female	Yes	32	Female	Yes
45	Female	No	18	Male	Yes	30	Male	No
62	Male	No	25	Male	No	25	Male	No
65	Male	No	60	Male	No	25	Male	No
45	Female	No	25	Male	Yes	25	Male	No
25	Female	No	20	Male	Yes	30	Male	No
28	Male	Yes	32	Male	Yes	35	Male	No
28	Male	No	32	Female	Yes	23	Male	Yes
23	Male	No	24	Female	Yes	24	Male	No
22	Female	Yes	30	Male	Yes	25	Female	Yes

[1] The data set comes from the textbook *The Statistical Sleuth* by Ramsey and Schafer. The story has been modified.

Descriptive statistics

- Overall survival probability is $\frac{20}{45} = 44.4\%$.
- Survival or not seems to be affected by gender.

Group	Survivals	Group size	Survival probability
Male	10	30	33.3%
Female	10	15	66.7%

- Survival or not seems to be affected by age.

Age class	Survivals	Group size	Survival probability
$[10, 20)$	2	3	66.7%
$[21, 30)$	11	22	50.0%
$[31, 40)$	4	8	50.0%
$[41, 50)$	3	7	42.9%
$[51, 60)$	0	2	0.0%
$[61, 70)$	0	3	0.0%

- May we do better? May we predict one's survival probability?

Ordinary regression is problematic

- Immediately we may want to construct a linear regression model

$$survival_i = \beta_0 + \beta_1 age_i + \beta_2 female_i + \epsilon_i.$$

where *age* is one's age, *gender* is 0 if the person is a male or 1 if female, and *survival* is 1 if the person is survived or 0 if dead.

- By running

```
d <- read.table("survival.txt", header = TRUE)
fitWrong <- lm(d$survival ~ d$age + d$female)
summary(fitWrong)
```

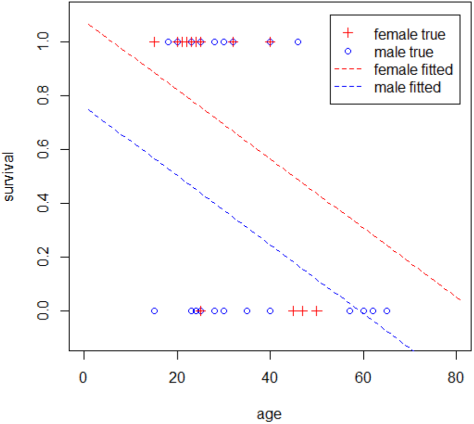
we may obtain the regression line

$$survival = 0.746 - 0.013 age + 0.319 female.$$

Though $R^2 = 0.1642$ is low, both variables are significant.

Ordinary regression is problematic

- The regression model gives us “predicted survival probability.”
  - For a man at 80, the “probability” becomes $0.746-0.013\times 80 = -0.294$, which is **unrealistic**.
- In general, it is very easy for an ordinary regression model to generate predicted “probability” not within 0 and 1.

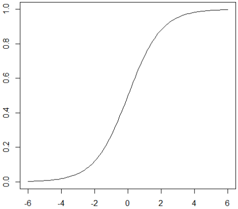


Logistic regression

- The right way to do is to do **logistic regression**.
- Consider the age-survival example.
  - We still believe that the smaller age increases the survival probability.
  - However, not in a linear way.
  - It should be that when one is **young enough**, being younger does not help too much.
  - The **marginal benefit** of being younger should be decreasing.
  - The **marginal loss** of being older should also be decreasing.
- One particular functional form that exhibits this property is

$$y = \frac{e^x}{1+e^x} \quad \Leftrightarrow \quad \log\left(\frac{y}{1-y}\right) = x$$

  - x can be anything in $(-\infty, \infty)$.
  - y is limited in $[0, 1]$.



Logistic regression

- We **hypothesize** that independent variables x_is affect π, the probability for y to be 1, in the following form:[2]

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

- The equation looks scaring. Fortunately, R is powerful.
- In R, all we need to do is to switch from `lm()` to `glm()` with an additional argument `binomial`.
  - `lm` is the abbreviation of "linear model."
  - `glm()` is the abbreviation of "generalized linear model."

[2] The logistic regression model searches for coefficients to make the curve fit the given data points in the best way. The details are far beyond the scope of this course.

Logistic regression in R

- By executing

  ```
  fitRight <- glm(d$survival ~ d$age + d$female, binomial)
  summary(fitRight)
  ```

  we obtain the regression report.

- Some information is new, but the following is familiar:

  ```
  Coefficients:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  1.63312    1.11018   1.471   0.1413
  d$age       -0.07820    0.03728  -2.097   0.0359 *
  d$female     1.59729    0.75547   2.114   0.0345 *
  ---
  Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
  ```

- Both variables are **significant**.

The Logistic regression curve

- The estimated curve is

$$\log\left(\frac{\pi}{1-\pi}\right) = 1.633 - 0.078 age + 1.597 female,$$

or equivalently,

$$\pi = \frac{\exp(1.633 - 0.078 age + 1.597 female)}{1 + \exp(1.633 - 0.078 age + 1.597 female)},$$

where $\exp(z)$ means e^z for all $z \in \mathbb{R}$.

The Logistic regression curve

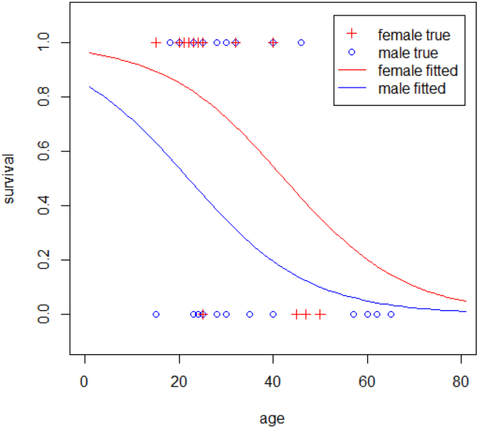
- The curves can be used to do **prediction**.
- For a man at 80, π is

  $$\frac{\exp(1.633-0.078\times 80)}{1+\exp(1.633-0.078\times 80)},$$

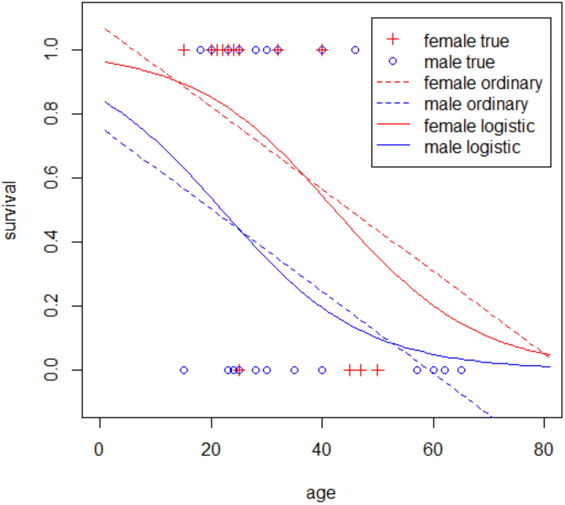
  which is 0.0097.
- For a woman at 60, π is

  $$\frac{\exp(1.633-0.078\times 60+1.597)}{1+\exp(1.633-0.078\times 60+1.597)},$$

  which is 0.1882.
- π is always in $[0,1]$. There is no problem for interpreting π as a probability.



Comparisons



Interpretations

- The estimated curve is

$$\log\left(\frac{\pi}{1-\pi}\right) = 1.633 - 0.078 age + 1.597 female.$$

  Any implication?

  - $-0.078 age$: Younger people will survive more likely.
  - $1.597 female$: Women will survive more likely.
- In general:
  - Use the ***p*-values** to determine the significance of variables.
  - Use the **signs** of coefficients to give qualitative implications.
  - Use the **formula** to make predictions.

Model selection

- Recall that in ordinary regression, we use R^2 and adjusted R^2 to assess the usefulness of a model.
- In logistic regression, we do not have R^2 and adjusted R^2.
- We have **deviance** instead.
  - In a regression report, the **null deviance** can be considered as the total estimation errors without using any independent variable.
  - The **residual deviance** can be considered as the total estimation errors by using the selected independent variables.
  - Ideally, the residual deviance should be **small**.[3]

[3] To be more rigorous, the residual deviance should also be close to its degree of freedom. This is beyond the scope of this course.

Deviances in the regression report

- The null and residual deviances are provided in the regression report.
- For `glm(d$survival ~ d$age + d$female, binomial)`, we have

```
    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 51.256  on 42  degrees of freedom
```

- Let's try some models:

Independent variable(s)	Null deviance	Residual deviance
age	61.827	56.291
female	61.827	57.286
age, *female*	61.827	51.256
age, *female*, *age* × *female*	61.827	47.346

  - Using *age* only is better than using *female* only.
- How to compare models with different numbers of variables?

Deviances in the regression report

- Adding variables will **always reduce** the residual deviance.
- To take the number of variables into consideration, we may use **Akaike Information Criterion** (AIC).
- AIC is also included in the regression report:

Independent variable(s)	Null deviance	Residual deviance	AIC
age	61.827	56.291	60.291
female	61.827	57.286	61.291
age, *female*	61.827	51.256	57.256
age, *female*, *age* × *female*	61.827	47.346	55.346

- AIC is only used to compare **nested** models.
  - Two models are nested if one's variables are form a subset of the other's.
  - Model 4 is better than model 3 (based on their AICs).
  - Model 3 is better than either model 1 or model 2 (based on their AICs).
  - Model 1 and 2 cannot be compared (based on their AICs).

Classification by logistic regression

- Logistic regression helps us identify key factors affecting the outcome.
- What if we really want to classify the next observation?
  - We may use all its features to calculate $\pi \in [0, 1]$.
  - How to determine whether the outcome is "yes" or "no"?
- We choose a **threshold** t to do the classification:
  - If $\pi > t$, classify the observation to class A; otherwise, class B.
  - We may set $t = \frac{1}{2}$ to build a classifier.
  - Optimizing t is beyond the scope of this course.

Road map

- Classification: logistic regression.
- **Association: frequent pattern mining**.
- Clustering: the k-means algorithm.

Frequent pattern mining

- **Frequent pattern mining** is to find the patterns (collection of items) that occur frequently.
  - Market basket analysis: A set of items that are purchased together.
  - A pair of weather condition and sold item that occur together.
  - A set of videos that receive five stars by a Netflix user.
  - A set of Netflix users that give five stars to a movie.
- If some items occurs together frequently, they are **highly associated**.
  - We want to identify these highly associated items.
  - Is that enough?
- Let's consider the following example.

Example

- Ten transactions regarding five products are recorded:
  - (D, E), (A, C, D), (A, D), (A, D), (D, E), (B, C, D), (A, B, E), (A, D), (C, D, E), (C, D).
- To make it easier to read, let's record them in a relational table.
- (C, D) seems to be a frequent pattern.
  - It appears in 40% of transactions.
- However:
  - Given that one purchased C, should we recommend D to her?
  - Given that one purchased D, should we recommend C to her?

A	B	C	D	E
0	0	0	1	1
1	0	1	1	0
1	0	0	1	0
1	0	0	1	0
0	0	0	1	1
0	1	1	1	0
1	1	0	0	1
1	0	0	1	0
0	0	1	1	1
0	0	1	1	0

Example

- The **joint probability** of two items matters.
  - The joint probability that C and D are bought together is 40%.
- The **conditional probability** between two items also matters.
  - Given that D has been bought, the probability of buying C is $\frac{4}{9} = 44.4\%$.
  - Given that C has been bought, the probability of buying D is $\frac{4}{4} = 100\%$.

A	B	C	D	E
0	0	0	1	1
1	0	1	1	0
1	0	0	1	0
1	0	0	1	0
0	0	0	1	1
0	1	1	1	0
1	1	0	0	1
1	0	0	1	0
0	0	1	1	1
0	0	1	1	0

Definition: Sets

- Let $I = \{i_1, i_2, ..., i_m\}$ be the set of **items**.
- Let $T_j \subseteq I$ be a set of items purchased in a transaction T_j.
- Let $T = \{T_1, T_2, ..., T_n\}$ be the set of **transactions**.
- Let $X \subseteq I$ and $Y \subseteq I$ be two sets of items that we are interested in.
- An **association rule** $X \Rightarrow Y$ means "If X occurs, then Y occurs."
  - X is called the antecedent item set.
  - Y is called the consequent item set.
  - We have $X \cap Y = \phi$, i.e., they have no overlapping.

Sets in our example

- $I = \{A, B, C, D, E\}$ is the set of items.
- Let $T = \{T_1, T_2, ..., T_{10}\}$ is the set of transactions.
- $T_1 = \{D, E\}$, $T_2 = \{A, C, D\}$, etc.
- An association rule $C \Rightarrow D$ means "If one purchases C, then she also purchases D."
- Another association rule $\{C, E\} \Rightarrow D$ means "If one purchases C and D, then she also purchases D."
- Let $f(X)$ be the number of transactions containing an item set $X \subseteq I$.
  - $f(A) = 0.5$.
  - $f(A \cup B) = 0.1$.
  - $f(A \cup B \cup C) = 0$.

A	B	C	D	E
0	0	0	1	1
1	0	1	1	0
1	0	0	1	0
1	0	0	1	0
0	0	0	1	1
0	1	1	1	0
1	1	0	0	1
1	0	0	1	0
0	0	1	1	1
0	0	1	1	0

Definition: Association measurements

- Given an association rule $X \Rightarrow Y$, we have three measurements.
- The **support** of the rule is the joint probability

$$\frac{f(X \cup Y)}{n}.$$

- The **confidence** of the rule is the conditional probability

$$\Pr(Y|X) = \frac{f(X \cup Y)}{f(X)}.$$

- The **lift** of the rule is the ratio

$$\frac{\Pr(Y|X)}{\Pr(Y)} = \frac{f(X \cup Y)/f(X)}{f(Y)/n}.$$

Association measurements in our example

- Consider the rule $D \Rightarrow C$.
- We have $f(C) = 4$ and $f(D) = 9$.
- The support is

$$\frac{f(C \cup D)}{10} = 0.4.$$

- The confidence is

$$\Pr(C|D) = \frac{f(C \cup D)}{f(D)} = \frac{4}{9} = 0.44.$$

- The lift is

$$\frac{\Pr(C|D)}{\Pr(C)} = \frac{4/9}{4/10} = 1.11.$$

A	B	C	D	E
0	0	0	1	1
1	0	1	1	0
1	0	0	1	0
1	0	0	1	0
0	0	0	1	1
0	1	1	1	0
1	1	0	0	1
1	0	0	1	0
0	0	1	1	1
0	0	1	1	0

Implications of association measurements

- Basically, we want to find a rule $X \Rightarrow Y$ with a **high confidence**.
  - This means that “once one buys X, with a high chance she will also be willing to buy Y.
- However, we also need a **high support**.
  - If the support is low, the high confidence may be just a coincidence.
- Finally, we need a **higher-than-1 lift**.
  - If X and Y are independent, we can show that the lift of $X \Rightarrow Y$ is

$$\frac{\Pr(Y|X)}{\Pr(Y)} = \frac{f(X \cup Y)/f(X)}{f(Y)/n} = 1.$$

  - The lift must be greater than 1 so that X and Y are positively correlated.
  - Or we may say that using X to predict Y is better than a random guess.

Association measurements in our example

- For $D \Rightarrow B$:
  - The confidence $\Pr(B|D) = 0.11$ is small.
- For $B \Rightarrow A$:
  - The confidence $\Pr(A|B) = 0.5$ is high.
  - The support $\frac{f(A \cup B)}{n} = 0.1$ is small.
- For $E \Rightarrow A$:
  - The lift $\frac{f(A \cup E)/f(A)}{f(E)/n} = \frac{1/5}{4/10} = 0.5 < 1$.

A	B	C	D	E
0	0	0	1	1
1	0	1	1	0
1	0	0	1	0
1	0	0	1	0
0	0	0	1	1
0	1	1	1	0
1	1	0	0	1
1	0	0	1	0
0	0	1	1	1
0	0	1	1	0

Remarks

- Given a set of transactions T, we look for association rules that have high confidences, high supports, and greater-than-1 lifts.
  - What is "high"?
- There is no general rule to define "high enough."
  - People choose their own **minimum confidence** and **minimum support** for filtering association rules.
  - The requirement for lift is always 1.
- If many rules satisfy the given criterion, we may increase the cutoffs.
  - Otherwise, we may decrease the cutoffs.
- A rule may also have multiple antecedent items.
  - It is easier for the confidence to be high.
  - It is quite likely that the support is low.

Shopping data set

- A data set records 786 transactions made by different customers for ten different goods.

ID	Ready made	Frozen foods	Alcohol	Fresh Vegetables	Milk	Bakery goods	Fresh meat	Toiletries
1	1	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	1
3	1	0	0	0	0	0	0	1
4	1	0	0	0	1	1	0	0
5	1	0	0	0	0	0	0	0

ID	Snacks	Tinned Goods	Gender	Age	Marital	Children	Working
1	1	0	Female	18 to 30	Widowed	No	Yes
2	0	0	Female	18 to 30	Separated	No	Yes
3	1	0	Male	18 to 30	Single	No	Yes
4	0	0	Female	18 to 30	Widowed	No	Yes
5	0	0	Female	18 to 30	Separated	No	Yes

Recommendations

- Goal: Given one's items in her shopping cart, make recommendations.
- If a rule $X \Rightarrow Y$ is significant, we may use it to recommend Y if X is in the cart.
- Let's ignore demographic information and focus on the cart.

Association rules

- Let's set the minimum support and minimum confidence to be 0.1 and 0.6, respectively.
- 8842 rules are found.

Association rules

- The top 5 association rules (ranked by confidence):

Antecedent set	Consequent set	Support	Confidence	Lift
Ready made = 0 Tinned Goods = 1	Fresh meat = 0	0.239	1	1.030
Ready made = 0 Snacks = 0	Fresh meat = 0	0.277	1	1.030
Ready made = 0 Alcohol = 1 Bakery goods = 0	Fresh meat = 0	0.113	1	1.030

Ready made = 0
Alcohol = 1
Toiletries = 0
	Fresh meat = 0	0.157	1	1.030

Alcohol = 1 Bakery goods = 0 Tinned goods = 0	Fresh vegetables = 0	0.129	1	1.090

Association rules for fresh vegetables

- Let’s focus on rules whose consequent sets contain a purchasing action.
- Let’s try **fresh vegetables**, because we want to promote them.
  - With the minimum support 0.1 and minimum confidence 0.6, no rule!
  - With the minimum support 0.1 and minimum confidence 0.1, no rule!
  - Fresh vegetables are **seldom sold**, so no rule can have a high support with fresh vegetables.
- With the minimum support 0.05 and minimum confidence 0.1, we find seven rules.
- What are them?

Association rules for fresh vegetables

- The top 5 association rules for fresh vegetables (ranked by confidence):

Antecedent set	Consequent set	Support	Confidence	Lift
Tinned goods = 1	Fresh vegetables = 1	0.069	0.151	1.824
Fresh meat = 0 Tinned goods = 1	Fresh vegetables = 1	0.062	0.145	1.748
Bakery goods = 1	Fresh vegetables = 1	0.058	0.136	1.651
Toiletries = 0 Tinned goods = 1	Fresh vegetables = 1	0.052	0.129	1.559
Bakery goods = 1 Fresh meat = 0	Fresh vegetables = 1	0.052	0.127	1.540

Short association rules

- It may be too hard to check too many items in the cart in a short time.
- Let's good at association rules whose **length** is 2.
  - The length of an association rule is the total number of items in the antecedent and consequent item sets.
  - A length-2 association rule is from one item to one item.
- With the minimum support 0.1 and minimum confidence 0.6, we find 99 rules.
- What are them?

Short association rules

- The top 5 length-2 association rules regarding a purchase (ranked by confidence):

Antecedent set	Consequent set	Support	Confidence	Lift
Milk = 1	Bakery goods = 1	0.140	0.743	1.733
Milk = 1	Ready made = 1	0.134	0.709	1.441
Milk = 1	Tinned goods = 1	0.127	0.676	1.483
Milk = 1	Snacks = 1	0.124	0.662	1.395
Milk = 1	Alcohol = 1	0.115	0.608	1.542

Considering demographic information

- May demographic information help us?
- Let’s focus on fresh vegetables again:

Antecedent set	Consequent set	Support	Confidence	Lift

Tinned.Goods = 1
Working = Yes
	Fresh vegetables = 1	0.059	0.163	1.973

Fresh meat = 0 Tinned goods = 1 Working = Yes	Fresh vegetables = 1	0.052	0.155	1.878
Tinned.Goods=1	Fresh vegetables = 1	0.069	0.151	1.824
Fresh meat = 0 Tinned goods = 1	Fresh vegetables = 1	0.062	0.145	1.748
Bakery goods = 1	Fresh vegetables = 1	0.058	0.136	1.651

- Adding demographic information generates the top 2 rules.

Road map

- Classification: logistic regression.
- Association: frequent pattern mining.
- **Clustering: the k-means algorithm**.

Introduction

- Recall the wholesale data set:

Channel	Label	Fresh	Milk	Grocery	Frozen	D. & P.	Deli.
1	1	30624	7209	4897	18711	763	2876
1	1	11686	2154	6824	3527	592	697
				⋮			
2	3	14531	15488	30243	437	14841	1867

- The wholesaler records the annual amount each customer spends on six product categories:
  - Fresh, milk, grocery, frozen, detergents and paper, and delicatessen.
  - Amounts have been scaled to be based on “monetary unit.”
- Channel: hotel/restaurant/café = 1, retailer = 2.
- Region: Lisbon = 1, Oporto = 2, others = 3.

Dividing customers into groups

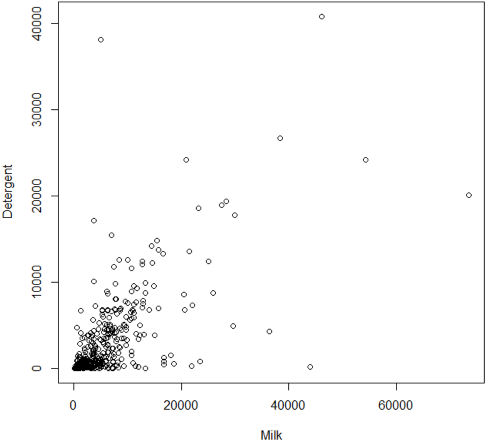
- In many cases, we would like to **customize** the advertising, service, and selling plans for different customers.
  - E.g., the price for milk may be different from customer to customer.
  - E.g., we may assign special agents for big customers.
- While there are 440 customers, we do not want to have 440 ways.
  - We want to **divide** customers to **groups**.
  - According to channel, region, a kind of sales, or what?
- This task is called **clustering**.

Clustering vs. classification

- Both **clustering** and **classification** are grouping data points (e.g., customers) into groups.
- However, they are different.
- Classification: Group information is **known** for existing data points.
    - Each existing data point is known to be in a group,
    - E.g., survival or death of a person, purchasing or not of a customer.
    - We use existing data points to identify critical factors leading to the grouping outcomes.
    - For future data whose groups are unknown, we classify them into groups.
- Clustering: Group information is **unknown** for existing data points.
    - We divide data points to clusters to make points within a class as **similar** as possible.
    - A future data point is put into the cluster that is "closest" to it.

Example

- How to create 6 clusters based on the milk and Detergent sales?



Cluster centers and distances

- Let $x^i = (x^i_1, x^i_2)$ be data point i, $i = 1, ..., 440$, where x^i_1 and x^i_2 are its milk and detergent sales, respectively.
- We want to create 6 clusters.
  - Let C_j be the set of points in cluster j, $j = 1, ..., 6$.
  - For cluster j, there is a **cluster center** $c^j = (c^j_1, c^j_2)$, $j = 1, ..., 6$.
  - If a point is in cluster j (i.e., $x^i \in C_j$), its **distance** to cluster center c^j is no longer than that to cluster c^k for all $k \neq j$.
  - The (Euclidean) distance between two points x^i and c^j is

$$d(x^i, c_j) = \sqrt{(x^i_1 - c^i_1)^2 + (x^i_2 - c^i_2)^2}.$$

- Therefore, the task of making 6 clusters is equivalent to choosing 6 points to be cluster centers.
  - A cluster center needs not to be an existing data point.

Quality of a set of clusters

- How to measure the quality of a set of 6 clusters?
- In cluster j, we want

$$\sum_{i \in C_j} d(x^i, c_j)^2 = \sum_{i \in C_j} \left[(x_1^i - c_1^i)^2 + (x_2^i - c_2^i)^2 \right]$$

  to be small, i.e., the points in the cluster are close to the center.
- We want to find 6 centers to minimize the **within-cluster sum of squared errors**

$$\text{WSSE} = \sum_{j=1}^{6} \sum_{i \in C_j} d(x^i, c_j)^2 = \sum_{j=1}^{6} \sum_{i \in C_j} \left[(x_1^i - c_1^i)^2 + (x_2^i - c_2^i)^2 \right].$$

Quality of a set of clusters

- If we only have one cluster, the within-cluster sum of squared errors can be minimized by setting the cluster center at $\bar{x}$, where

$$\bar{x}_p = \frac{\sum_{i=1}^{440} x_p^i}{440}.$$

- Let

$$\text{TSSE} = \sum_{i=1}^{440} d(x^i, \bar{x})^2 = \sum_{i=1}^{440} \left[(x_1^i - \bar{x}_1)^2 + (x_2^i - \bar{x}_2)^2 \right],$$

- Hopefully the fraction $\frac{WSSE}{TSSE}$ is small.

Finding cluster centers

- To find cluster centers, we may use the R function `kmeans()`.

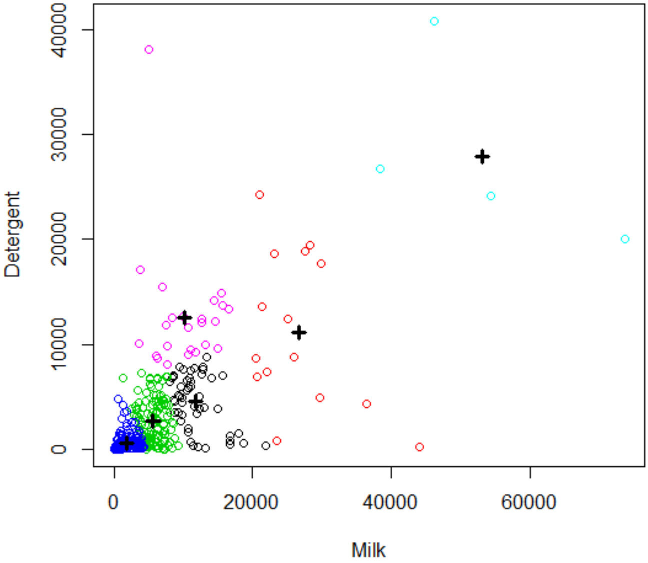
  ```
  W <- read.table("wholesale.txt", header = TRUE)
  w <- W[, c(4, 7)]
  km <- kmeans(w, centers = 6)
  ```

- The object `km` contains information about clusters.
  - `km$cluster` indicates the cluster each point belongs to.
  - `km$center` contains the coordinates of the cluster centers.
  - `km$totss` is TSSE.
  - `km$withinss` is WSSE.

Finding cluster centers

- Let's visualize the clustering outcome.

```
plot(w[, ], xlab = "Milk", ylab = "Detergent")
for(i in 1:6)
  points(w[which(km$cluster == i), ], col = i)
points(km$centers, col = 9, lwd = 3, pch = 3)
```


Detergent
0
10000
20000
30000
40000
Milk
0
20000
40000
60000

Five remaining questions

- The scales of milk and detergent sales are different.
- How to decide the number of clusters to build?
- May we use more than two variables?
- May we use categorical variables?
- How to choose variables for the clustering process to be based on?

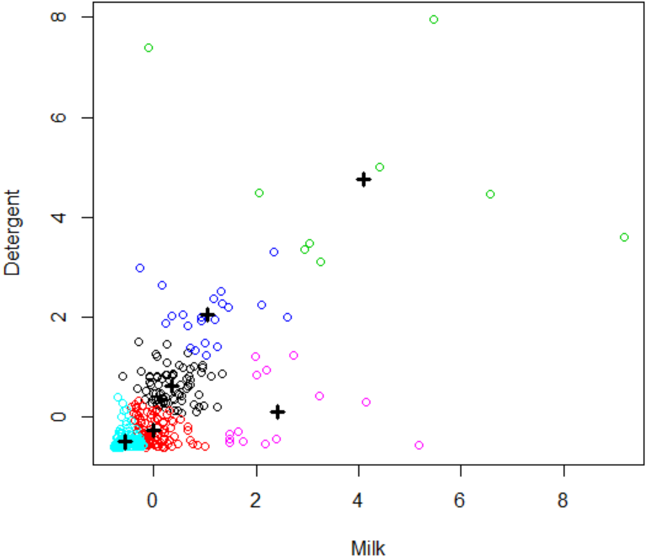
Scaling variables before clustering

- The scales of milk and detergent sales are different.
- In this case, we may scale them first.
- The most common way is to **standardize** each of them into ***z*-scores**:

$$z_p^i = \frac{x_p^i - \bar{x}_p}{s_p}, \text{ where } s_p = \sqrt{\frac{\sum_{i=1}^{440}(x_p^i - \bar{x}_p)^2}{440}}.$$

- In R:

```
w[, 1] <- (w[, 1] - mean(w[, 1])) / sd(w[, 1])
w[, 2] <- (w[, 2] - mean(w[, 2])) / sd(w[, 2])
```

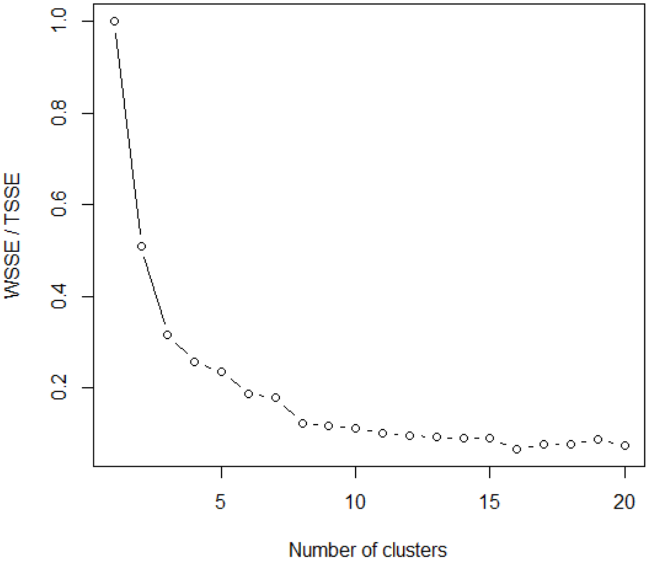


Detergent
8
6
4
2
0
0
2
4
6
8
Milk

Number of clusters

- The more clusters, the smaller WSSE.
  - However, each cluster also becomes less informative.
- We typically stop increasing the number of clusters when the **marginal improvement on WSSE** becomes too small.
- In R:

```
z <- rep(0, 20)
for(k in 1:20)
{
  km <- kmeans(w, centers = k)
  z[k] <- km$tot.withinss / km$totss
}
plot(z, type = "b", xlab = "Number of clusters",
     ylab = "WSSE / TSSE")
```


WSSE / TSSE
0.2
0.4
0.6
0.8
1.0
5
10
15
20
Number of clusters

Using more than two variables

- We may include as many variables as we want.
  - As long as they are **quantitative**.
- In R:

```
w <- W[, 3:8]
for(i in 1:6)
  w[, i] <- (w[, i] - mean(w[, i])) / sd(w[, i])
km <- kmeans(w, centers = 6)
```

Categorical variables

- May we include a categorical variable in the clustering process?
- Unfortunately, no!
  - Because there is no way to calculate distances.

How to choose variables?

- How to choose variables for the clustering process to be based on?
  - Milk and detergent?
  - Milk, fresh food, and detergent?
  - All variables?
- It depends on what you want to do.
  - The decision maker makes her own judgment.
  - Some other methods (e.g., regression) can be applied.