COMP20008 Elements of Data Processing

Semester 2 2018

Lecture 9: Assessing Correlations

THE UNIVERSITY OF MELBOURNE

Plan today

- Discuss about finding correlations between pairs of features in a dataset
  - Why useful and important
  - Pitfalls
- Review methods for computing correlation
  - Euclidean distance
  - Pearson correlation
- Next lecture
  - Mutual information (another method to compute correlation)

What is Correlation?

- Correlation is used to detect pairs of variables that might have some relationship

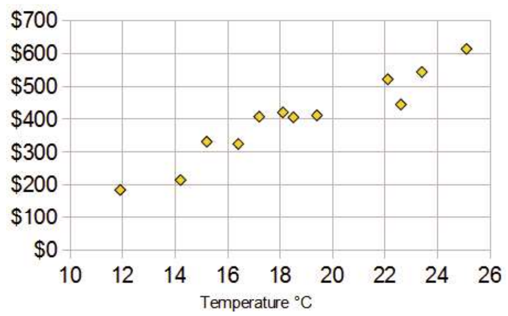
Ice Cream Sales vs Temperature

Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

https://www.mathsisfun.com/data/correlation.html

What is Correlation?

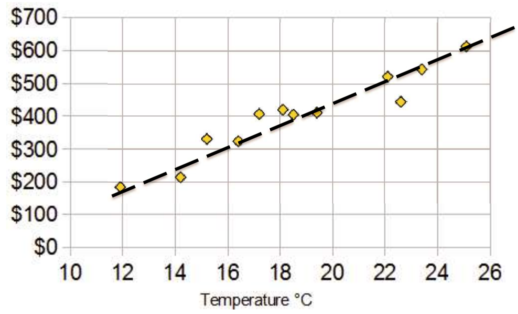
- Visually can be identified via inspecting scatter plots



https://www.mathsisfun.com/data/correlation.html

What is Correlation?

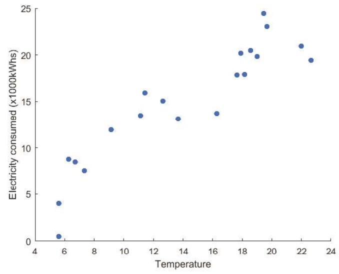
- Linear relations



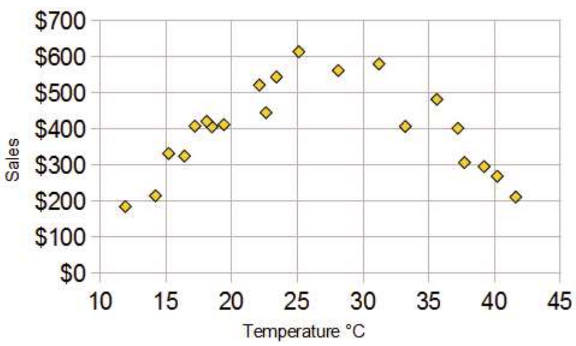
https://www.mathsisfun.com/data/correlation.html

Example of Correlated Variables

- Can hint at potential causal relationships (change in one variable is the result of change in the other)
- Business decision based on correlation: increase electricity production when temperature increases



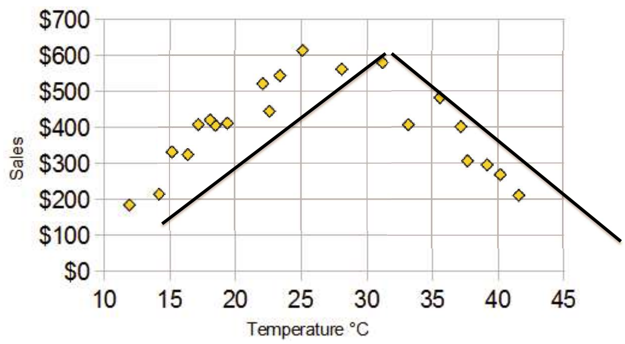
Example of non-linear correlation



It gets so hot that people aren't going near the shop, and **sales start dropping**

https://www.mathsisfun.com/data/correlation.html

Example of non-linear correlation



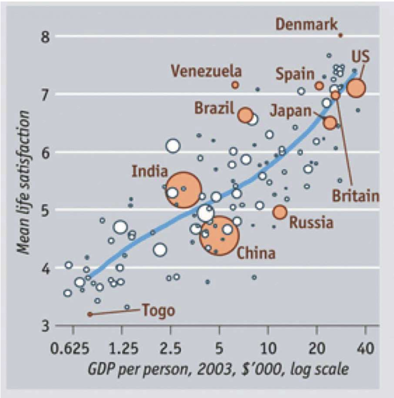
It gets so hot that people aren't going near the shop, and **sales start dropping**

https://www.mathsisfun.com/data/correlation.html

THE UNIVERSITY OF MELBOURNE

Wealth and happiness

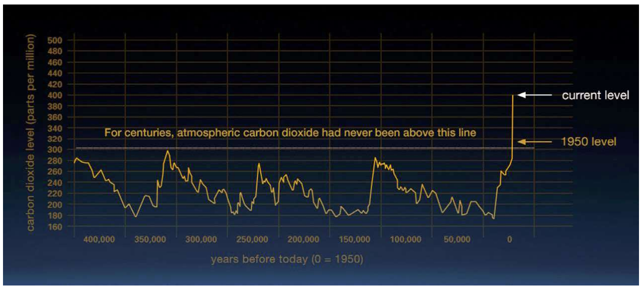
[from https://www.economist.com/blogs/dailychart/2010/11/daily_chart_1]



THE UNIVERSITY OF MELBOURNE

Climate change

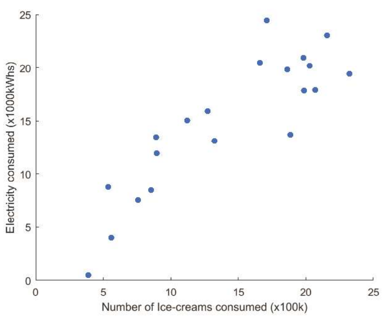
[https://climate.nasa.gov/evidence/]



THE UNIVERSITY OF MELBOURNE

Example of Correlated Variables

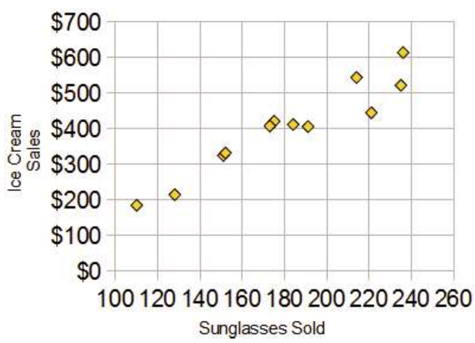
- Correlation does not necessarily imply causality!



THE UNIVERSITY OF MELBOURNE

Example of Correlated Variables

- Correlation does not necessarily imply causality!

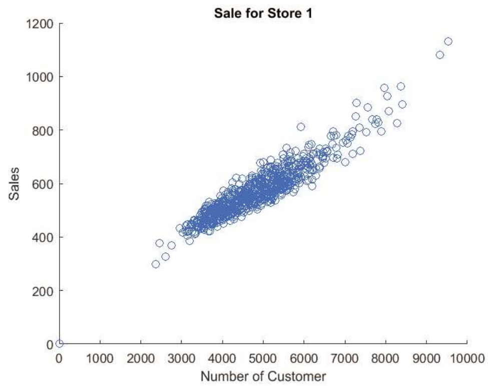


Example: Predicting Sales

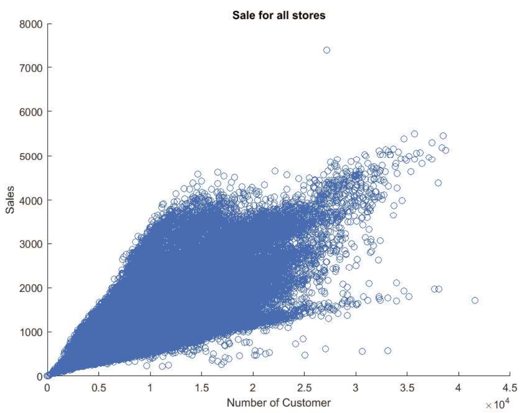
- https://www.kaggle.com/c/rossmann-store-sales/data

1	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHolida
2	1	5	31/07/2015	5263	555	1	1	0	1
3	2	5	31/07/2015	6064	625	1	1	0	1
4	3	5	31/07/2015	8314	821	1	1	0	1
5	4	5	31/07/2015	13995	1498	1	1	0	1
6	5	5	31/07/2015	4822	559	1	1	0	1
7	6	5	31/07/2015	5651	589	1	1	0	1
8	7	5	31/07/2015	15344	1414	1	1	0	1
9	8	5	31/07/2015	8492	833	1	1	0	1
10	9	5	31/07/2015	8565	687	1	1	0	1
11	10	5	31/07/2015	7185	681	1	1	0	1
12	11	5	31/07/2015	10457	1236	1	1	0	1
13	12	5	31/07/2015	8959	962	1	1	0	1
14	13	5	31/07/2015	8821	568	1	1	0	0
15	14	5	31/07/2015	6544	710	1	1	0	1
16	15	5	31/07/2015	9191	766	1	1	0	1
17	16	5	31/07/2015	10231	979	1	1	0	1
18	17	5	31/07/2015	8430	946	1	1	0	1
19	18	5	31/07/2015	10071	936	1	1	0	1
20	19	5	31/07/2015	8234	718	1	1	0	1
21	20	5	31/07/2015	9593	974	1	1	0	0
22	21	5	31/07/2015	9515	682	1	1	0	1
23	22	5	31/07/2015	6566	633	1	1	0	0
24	23	5	31/07/2015	7273	560	1	1	0	1
25	24	5	31/07/2015	14190	1082	1	1	0	1
26	25	5	31/07/2015	14180	1586	1	1	0	1

Example: Predicting Sales



Example: Predicting Sales



Example: Predicting Sales

- Other correlations
  - Sales vs. holiday
  - Sales vs. day of the week
  - Sales vs. distance to competitors
  - Sales vs. average income in area

THE UNIVERSITY OF MELBOURNE

Example rank correlation

- *"If a university has a higher-ranked football team, then is it likely to have a higher-ranked basketball team?"*

Football ranking	University team
1	Melbourne
2	Monash
3	Sydney
4	New South Wales
5	Adelaide
6	Perth

Basketball ranking	University team
1	Sydney
2	Melbourne
3	Monash
4	New South Wales
5	Perth
6	Adelaide

THE UNIVERSITY OF MELBOURNE

Why is correlation important?

- Discover relationships
- One step towards discovering causality

A causes B

Examples:

Gene A causes lung cancer

- Feature ranking: select the best features for building better predictive models
  - A good feature to use, is a feature that has high correlation with the outcome one is trying to predict

THE UNIVERSITY OF MELBOURNE

Case study: Microarray data

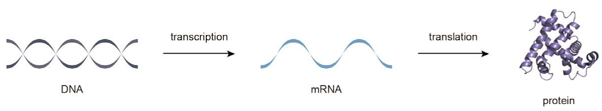
- DNA Microarrays (Gene Chips)
- Measure genes' level of activity

https://en.wikipedia.org/wiki/Bio-MEMS

THE UNIVERSITY OF MELBOURNE

The Central Dogma of Molecular Biology

- DNA makes RNA makes proteins



- DNA contains multiple genes containing information to produce different types of proteins
- To much or too little proteins of certain type can cause diseases
- Gene chips can measure the amount or mRNA (a proxy for protein level) – activity level (expression level)

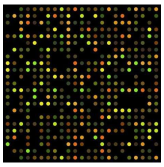
http://www.atdbio.com/content/14/Transcription-Translation-and-Replication

THE UNIVERSITY OF MELBOURNE

Microarray data

- Each chip contains thousands of tiny probes corresponding to the genes (20k - 30k genes in humans). Each probe measures the activity (expression) level of a gene

	Gene 1 expression	Gene 2 expression	…	Gene 20K expression
	0.3	1.2	…	3.1



THE UNIVERSITY OF MELBOURNE

Microarray dataset

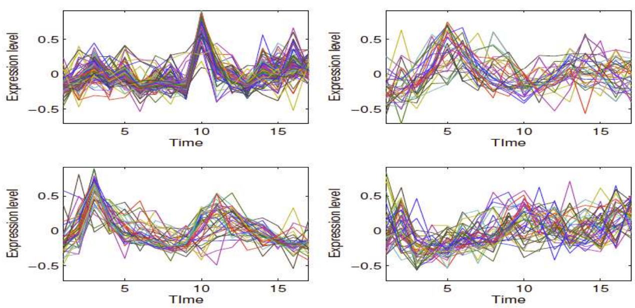
	Gene 1	Gene 2	Gene 3	…	Gene n
Time 1	2.3	1.1	0.3	…	2.1
Time 2	3.2	0.2	1.2	…	1.1
Time 3	1.9	3.8	2.7	…	0.2
…	…	…	…	…	…
Time m	2.8	3.1	2.5	…	3.4

- Each row represents measurements at some time
- Each column represents levels of a gene

THE UNIVERSITY OF MELBOURNE

Correlation analysis on Microarray data

- Can reveal genes that exhibit similar patterns ⇨ similar or related functions ⇨ Discover functions of unknown genes



THE UNIVERSITY OF MELBOURNE

Problem of Euclidean distance

- Objects can be represented with different measure scales

	Day 1	Day 2	Day 3	…	Day m
Temperature	20	22	16	…	33
#Ice-creams	50223	55223	45098	…	78008
#Electricity	102034	105332	88900	…	154008

d(temp,ice-cr)= 540324
d(temp,elect)= 12309388

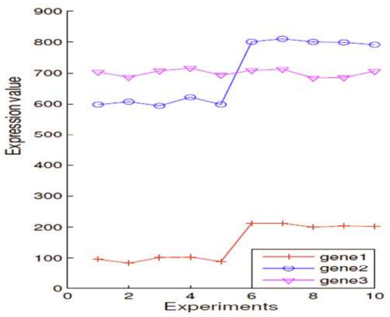
- Euclidean distance: does not give a clear intuition about how well variables are correlated

$$\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_n-y_n)^2}$$

$$\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

THE UNIVERSITY OF MELBOURNE

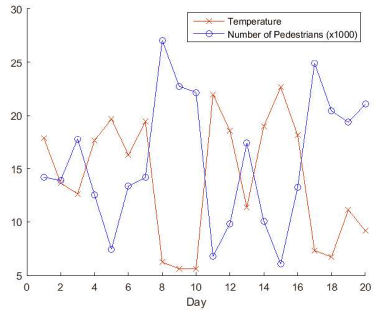
Problem of Euclidean distance



- Cannot discover variables with similar behaviours/dynamics but at different scale

THE UNIVERSITY OF MELBOURNE

Problem of Euclidean distance



- Cannot discover variables with similar behaviours/dynamics but in the opposite direction (negative correlation)

THE UNIVERSITY OF MELBOURNE

Assessing linear correlation – Pearson correlation

- We will define a correlation measure r_{xy}, assessing samples from two features x and y
  - Assess how close their scatter plot is to a straight line (a linear relationship)

- Range of r_{xy} lies within [-1,1]:
  - 1 for perfect positive linear correlation
  - -1 for perfect negative linear correlation
  - 0 means no correlation
  - Absolute value |r| indicates strength of linear correlation

- http://www.bc.edu/research/intasc/library/correlation.shtml

THE UNIVERSITY OF MELBOURNE

Pearson's correlation coefficient (r)

$$r = \mathrm{r}_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right).\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}}$$

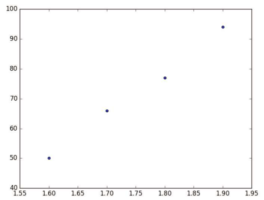
- x and y are the two attributes in your dataset

- Sample means $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Pearson coefficient example

Height (x)	Weight (y)
1.6	50
1.7	66
1.8	77
1.9	94

- How do the values of x and y move (vary) together?
- Big values of x with big values of y?
- Small values of x with small values of y?



Pearson coefficient example

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2).(\sum_{i=1}^{n}(y_i - \bar{y})^2)}}$$

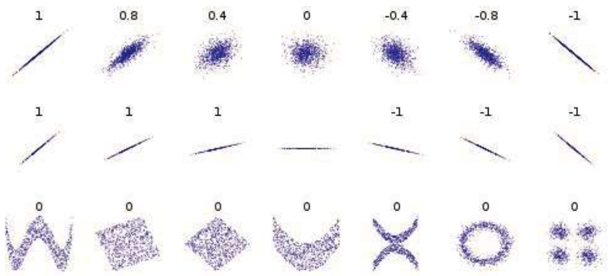
x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1.6	50	-0.15	-21.75	3.2625	0.0225	473.0625
1.7	66	-0.05	-5.75	0.2875	0.0025	33.0625
1.8	77	0.05	5.25	0.2625	0.0025	27.5625
1.9	94	0.15	22.25	3.3375	0.0225	495.0625

$\bar{x}$ = 1.75 | $\bar{y}$ = 71.75

$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ = 7.15 | $\sum_{i=1}^{n}(x_i - \bar{x})^2$ = 0.05 | $\sum_{i=1}^{n}(y_i - \bar{y})^2$ = 1028.75

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2).(\sum_{i=1}^{n}(y_i - \bar{y})^2)}} = \frac{(7.15)}{\sqrt{0.05 \text{ x } 1028.75}} = 0.996933$$

Examples



https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

Interpreting Pearson correlation values

- In general it depends on your domain of application. Jacob Cohen has suggested
  - 0.5 is large
  - 0.3-0.5 is moderate
  - 0.1-0.3 is small
  - less than 0.1 is trivial

THE UNIVERSITY OF MELBOURNE

Properties of Pearson's correlation

- Range within [-1,1]
- Scale invariant: r(x,y)= r(x, Ky)
  - Multiplying a feature's values by a constant K makes no difference
- Location invariant: r(x,y)= r(x, K+y)
  - Adding a constant K to one feature's values makes no difference
- Can only detect linear relationships

  y = a.x + b + noise

Cannot detect non-linear relationship

y = x^3 + noise

THE UNIVERSITY OF MELBOURNE

2017 Exam question 3a

Instance ID	*Predicted class*	*Actual class*
1	X	X
2	X	Y
3	Y	Y
4	X	X
5	X	Y
6	Y	X
7	X	X
8	Y	Y
9	Y	X
10	Y	Y

a) (1 mark) Would Pearson correlation be suitable to compute the correlation between the *Predicted class* and *Actual class*? Why or why not?

THE UNIVERSITY OF MELBOURNE

2016 exam question 2a)

2. a) Richard is a data wrangler. He does a survey and constructs a dataset recording average time/day spent studying and average grade for a population of 1000 students:

Student Name	*Average time per day studying*	*Average Grade*
...	...	...

i) (3 marks) Richard computes the Pearson correlation coefficient between *Average time per day studying* and *Average grade* and obtains a value of 0.85. He concludes that more time spent studying causes a student's grade to increase. Explain the limitations with this reasoning and suggest two alternative explanations for the 0.85 result.

THE UNIVERSITY OF MELBOURNE

Notes

- Interactive correlation calculator
  - http://www.bc.edu/research/intasc/library/correlation.shtml
- Correlation <> Causality

http://tylervigen.com/spurious-correlations

- Google trends correlation

Points you should know from today

- be able to explain why identifying correlations is useful for data wrangling/analysis
- understand what is correlation between a pair of features
- understand how correlation can be identified using visualisation
- understand the concept of a linear relation, versus a non linear relation for a pair of features
- understand why the concept of correlation is important, where it is used and understand why correlation is not the same as causation
- understand the use of Euclidean distance for computing correlation between two features and its advantages/disadvantages

Point to know – cont.

- understand the use of Pearson correlation coefficient for computing correlation between two features and its advantages/disadvantages
- understand the meaning of the variables in the Pearson correlation coefficient formula and how they can be calculated. Be able to compute this coefficient on a simple pair of features. The formula for this coefficient will be provided on the exam.
- be able to interpret the meaning of a computed Pearson correlation coefficient
- understand the advantages and disadvantages of using the Pearson correlation coefficient for assessing the degree of relationship between two features