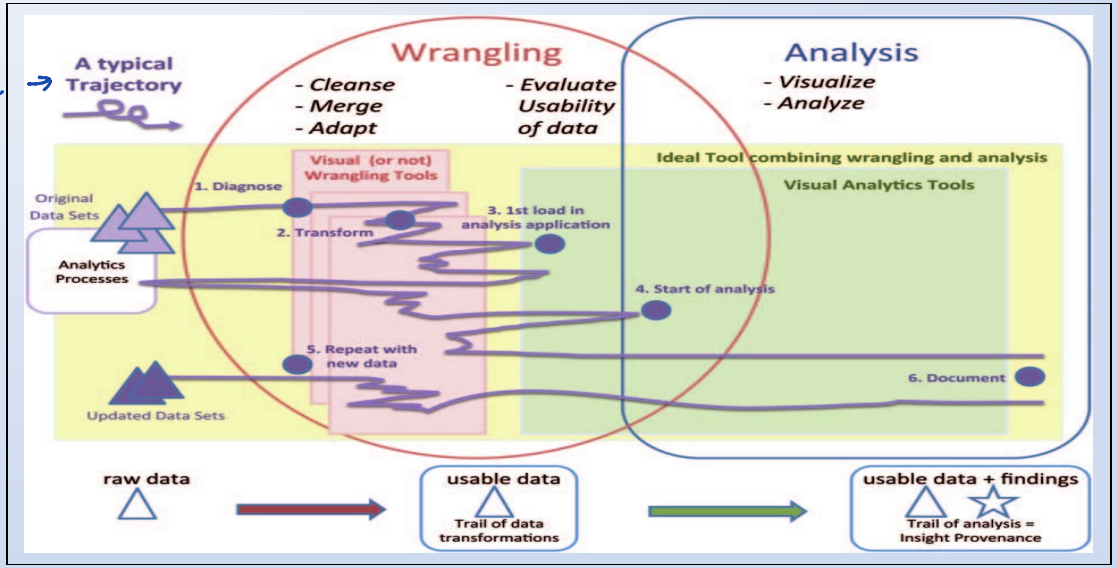
COMP20008 Elements of Data Processing

Correlation

Semester 1, 2020

Contact: uwe.aickelin@unimelb.edu.au

Where are we now?



Plan

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

Correlation

direction + −

strength [0, 1]

What is Correlation?

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

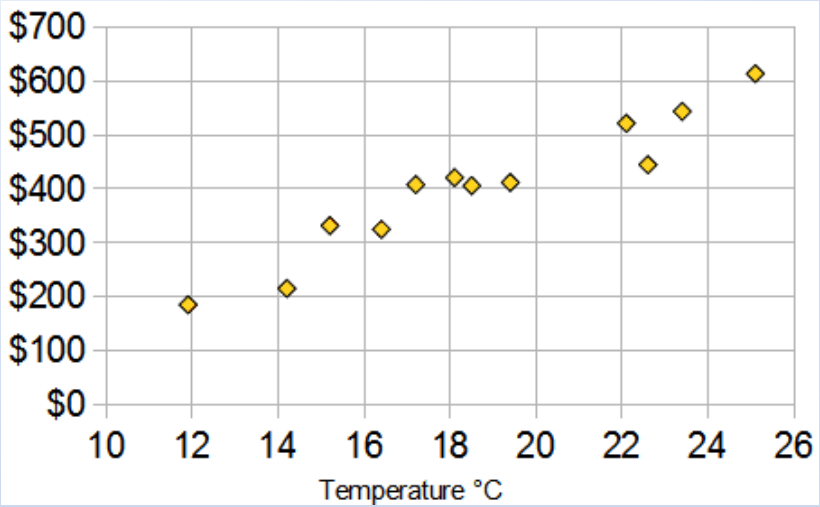
Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

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

What is Correlation?

Visually can be identified via inspecting scatter plots

positive correlation

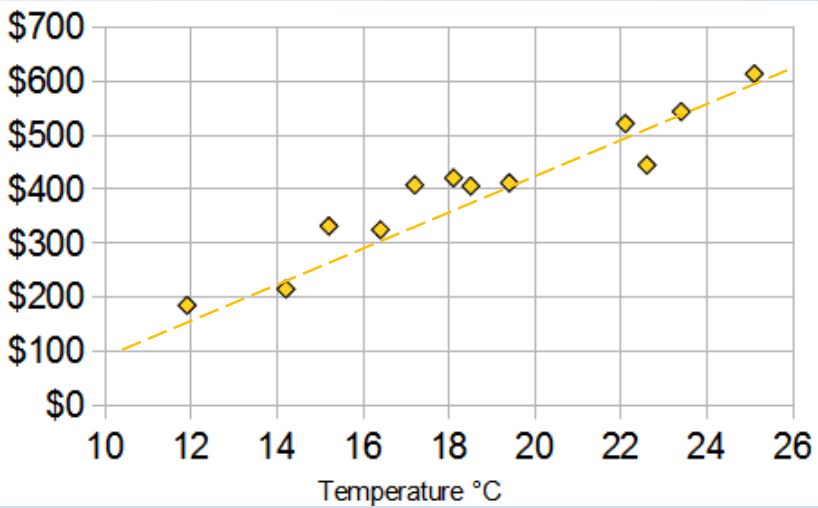


https://www.mathsisfun.com/data/correlation.html

What is Correlation?

Linear relations

↑ temperature ↑↑, sales ↑ by a linear proportional amount

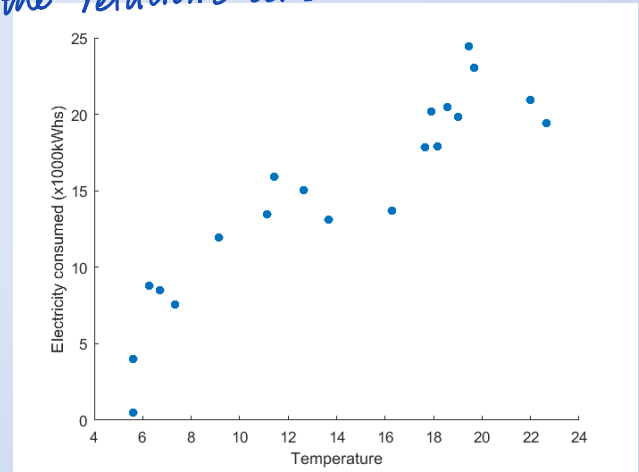


https://www.mathsisfun.com/data/correlation.html

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

[?] by itself it doesn't what's the relation's direction



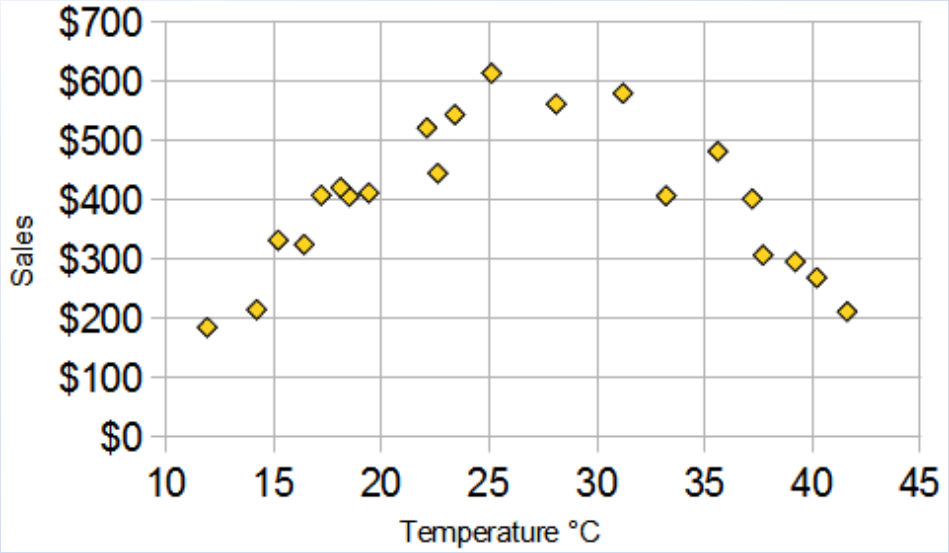
correlation and casuality

相关关系

causation

因果关系

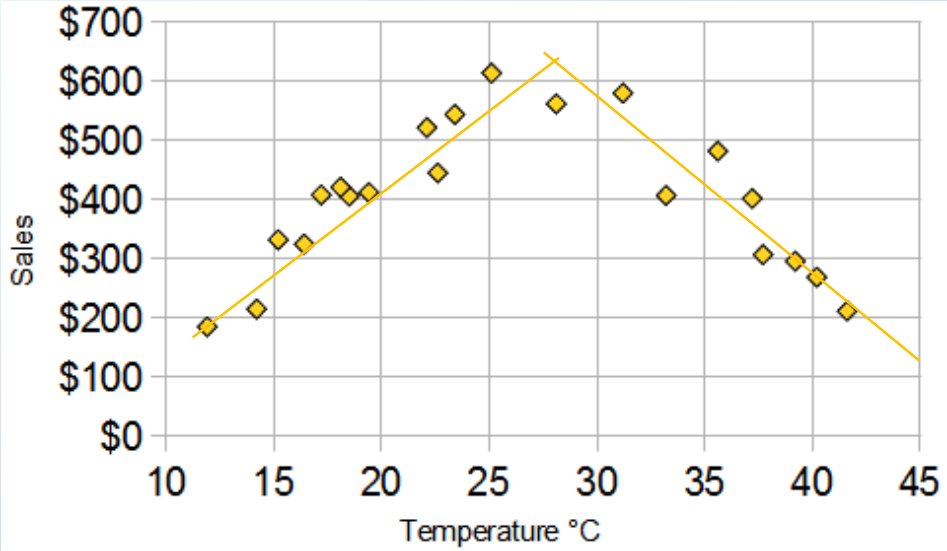
Example of non-linear correlation



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

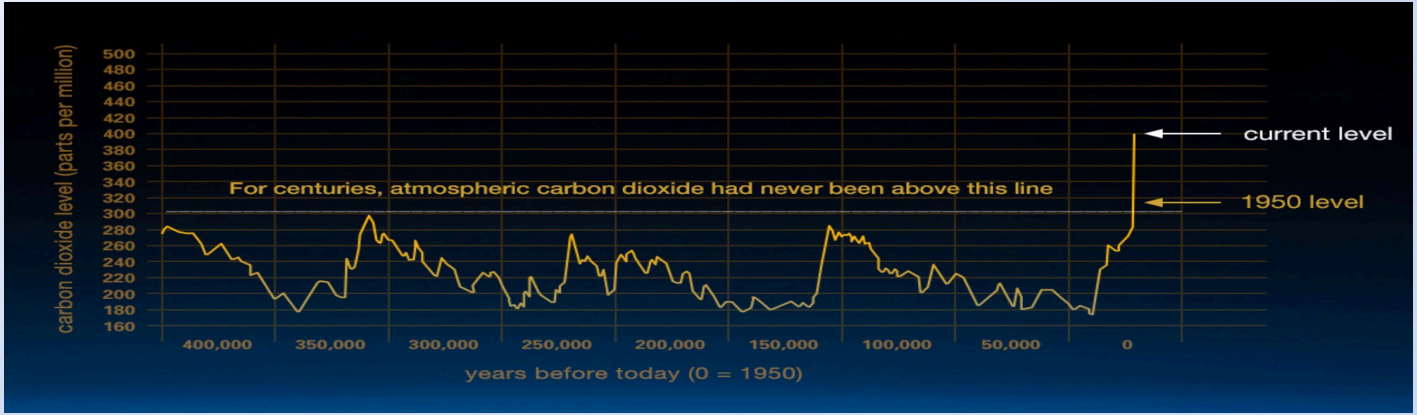


It gets so hot that people aren't going near the shop, and **sales start dropping**

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

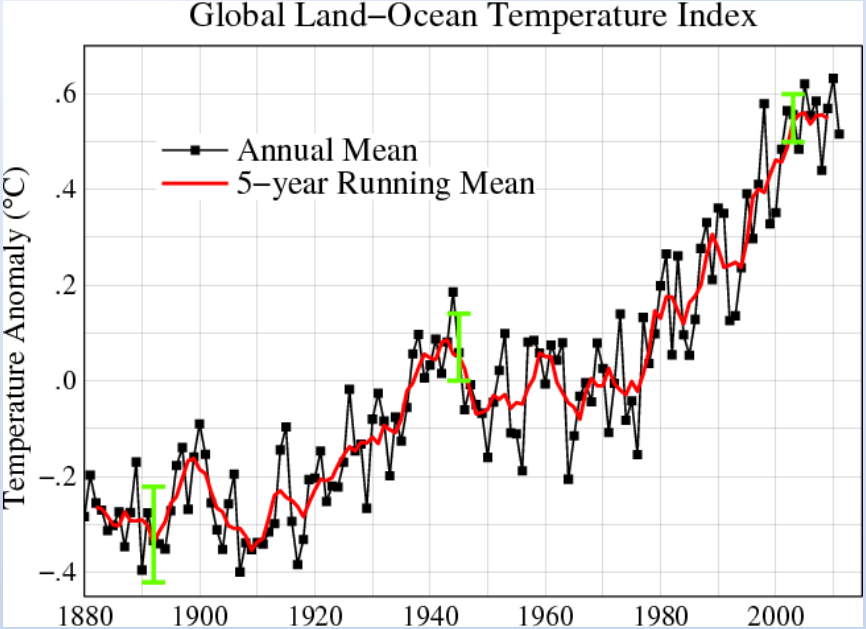
Climate change

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

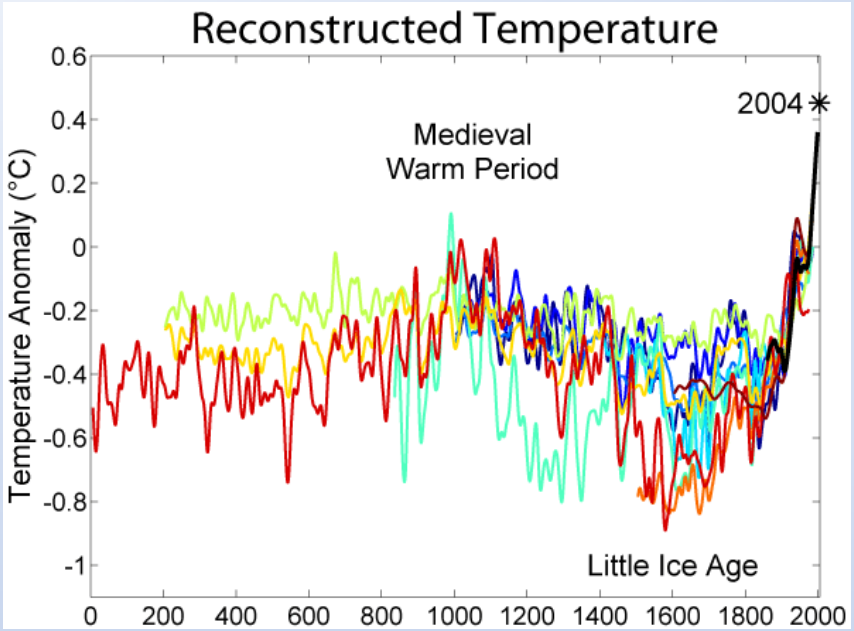


Climate Change?

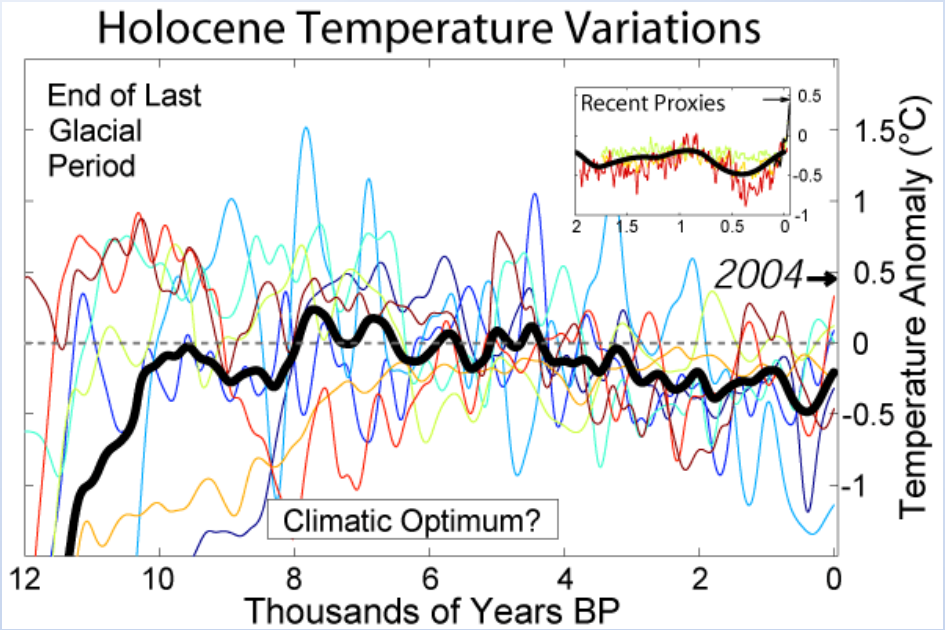
http://en.wikipedia.org/wiki/File:2000_Year_Temperature_Comparison.png



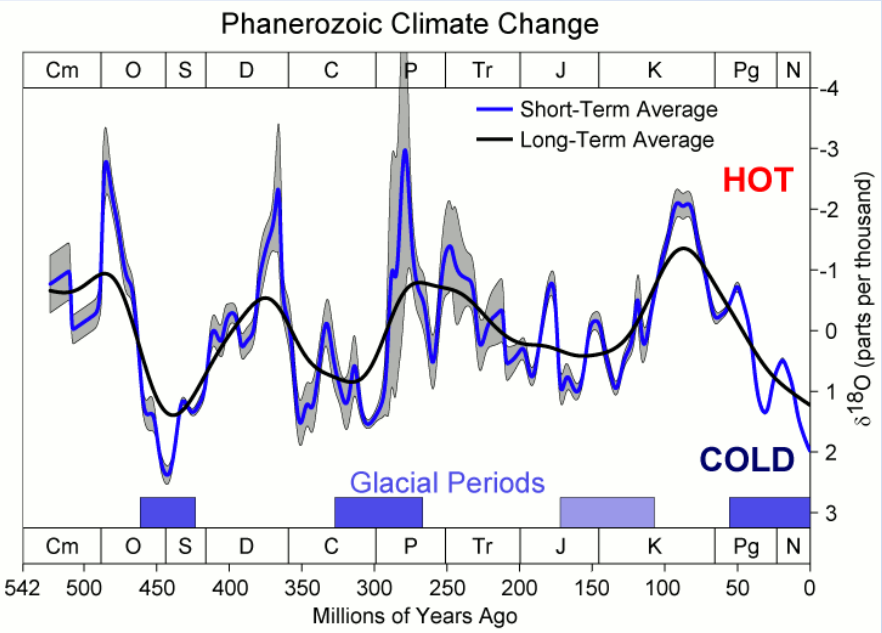
Climate Change?



Climate Change?



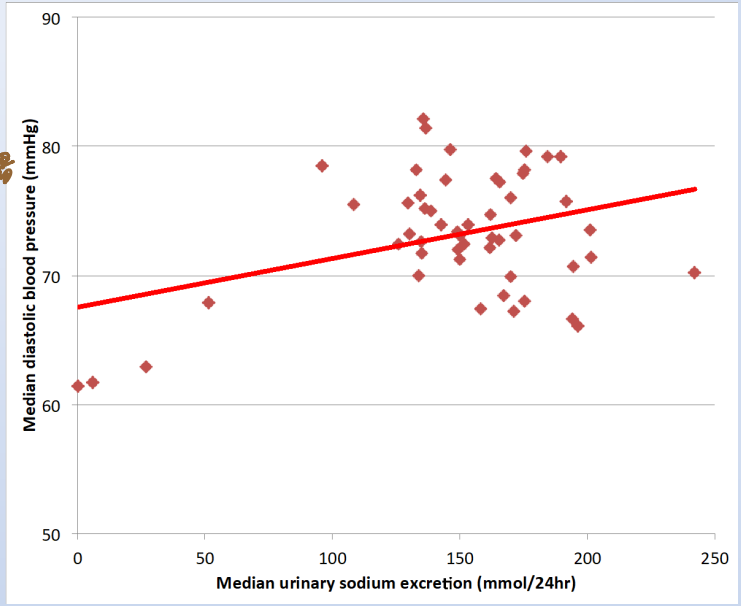
Climate Change?



Salt Causes High Blood Pressure

Intersalt: an international study of electrolyte excretion and blood pressure. Results for 24 hour urinary sodium and potassium excretion.

British Medical Journal; 297: 319-328, 1988.

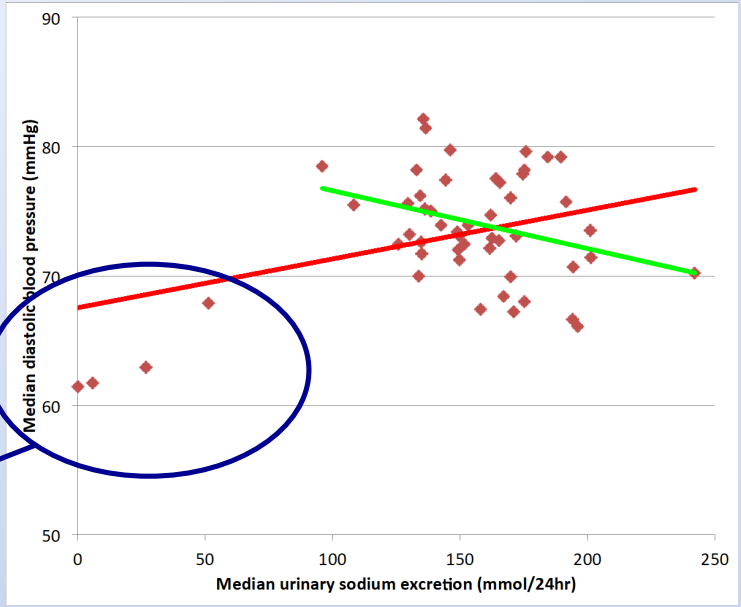


Or Does It!?

Intersalt: an international study of electrolyte excretion and blood pressure. Results for 24 hour urinary sodium and potassium excretion.

British Medical Journal; 297: 319-328, 1988.

If we exclude these four 'outliers' which are non-industrialised countries with non-salt diets, we get a quite different result!

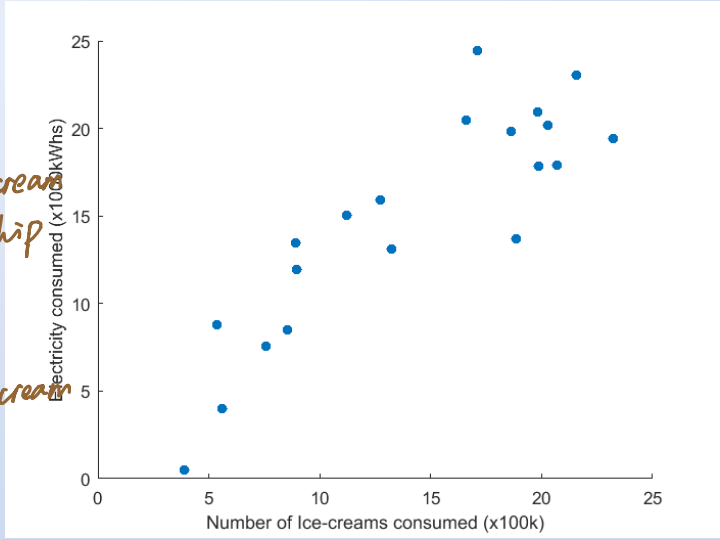


line with few dots may be outliers

for outlier, we can use clustering to identify

Example of Correlated Variables

Correlation does not necessarily imply causality!

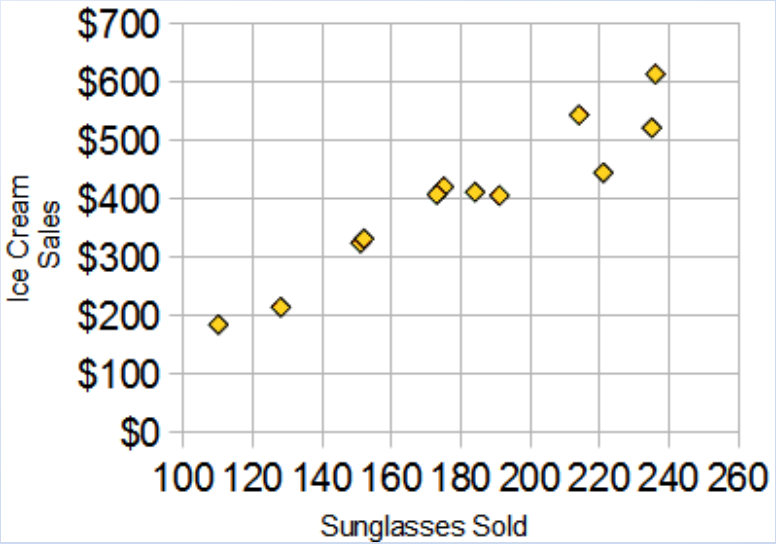


for correlation
electricity and no. of ice-cream has a nice linear relationship

for causality
does that means more ice-cream more electricity? No!
since there is a third observed variable cause them ⇒ T

Example of Correlated Variables

Correlation does not necessarily imply causality!

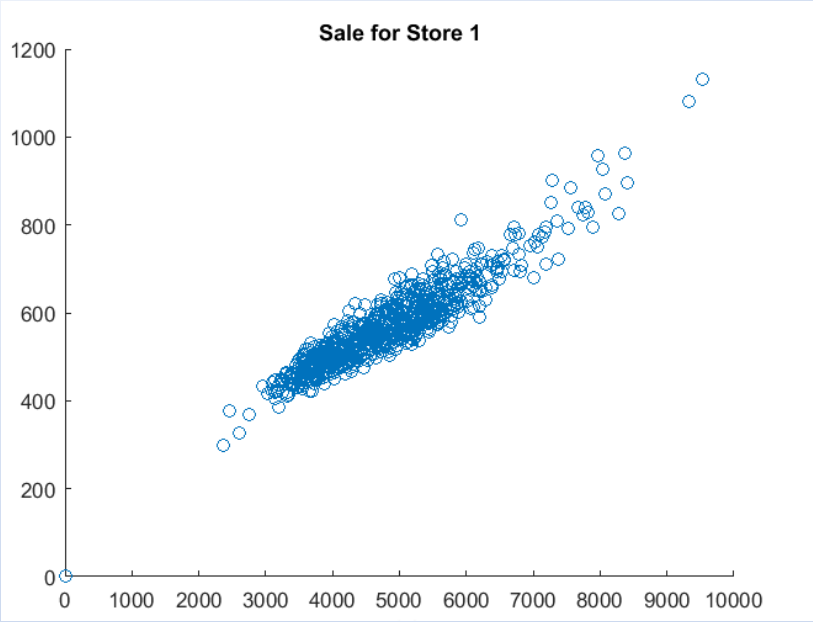


Example: Predicting Sales

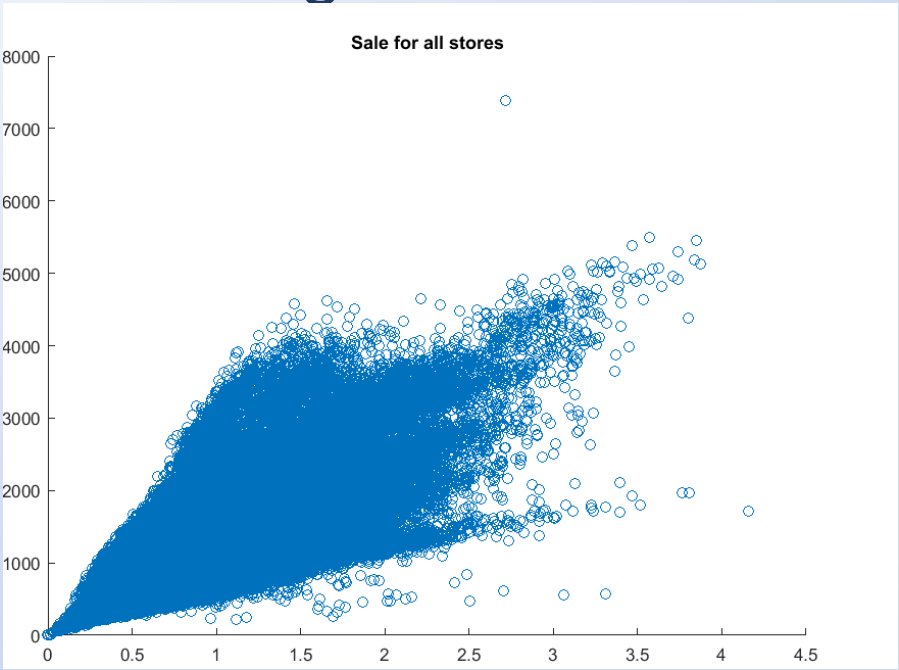
https://www.kaggle.com/c/rossmann-store-sales/data

	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

Factors correlated with human performance

- Bedtime consistency 统一性
- High intensity exercise
- Intermittent fasting 间歇性断食
- Optimism

Why is correlation important?

- Discover relationships

- One step towards discovering causality **A causes B**

Example: Gene A causes lung cancer

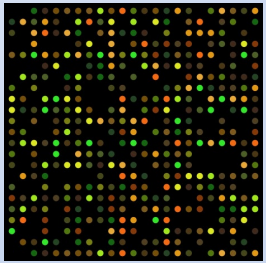
- Feature ranking: select the best features for building better predictive models
  - A good feature to use, is a feature that has high correlation with the outcome we are trying to predict

set of DNA sequence

Microarray data

Each chip contains thousands of tiny probes (调查) 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



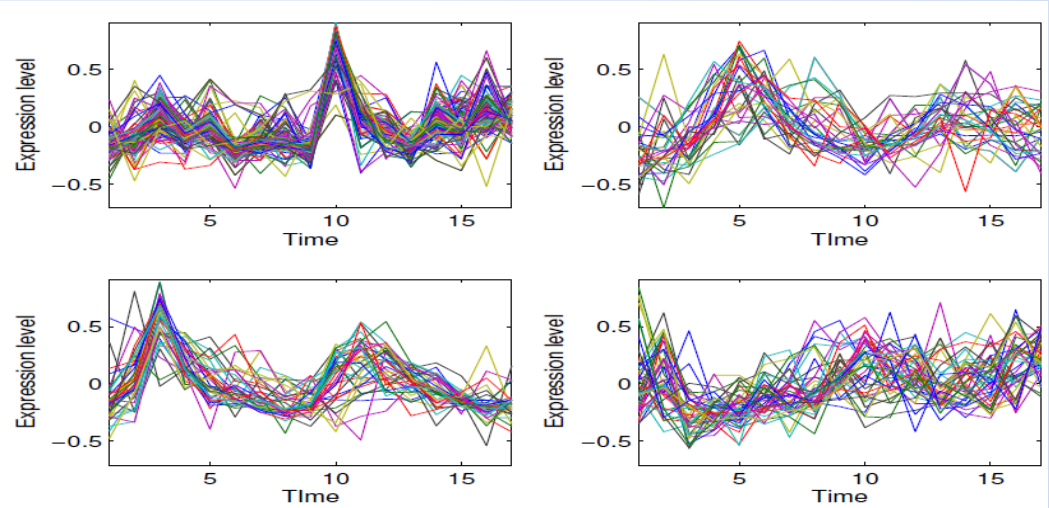
Microarray dataset

	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

Correlation analysis on Microarray data

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



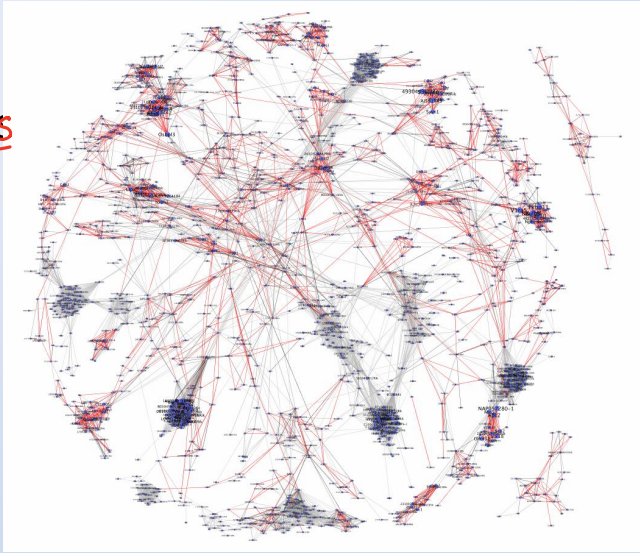
Compare people

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

- Each row represents measurements about a person
- Each column represents levels of a gene

Genetic network

Connect genes with high correlation – understand behaviour of groups of genes



http://www.john.ranola.org/?page_id=116

Euclidian Distance

Problems 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

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

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

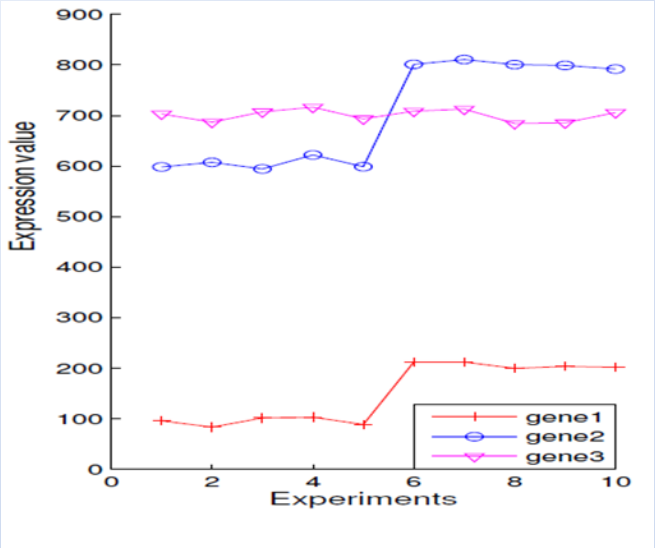
- Euclidean distance: does not give a clear intuition about how well variables are correlated

Problems of Euclidean distance

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

red and blue are highly correlated

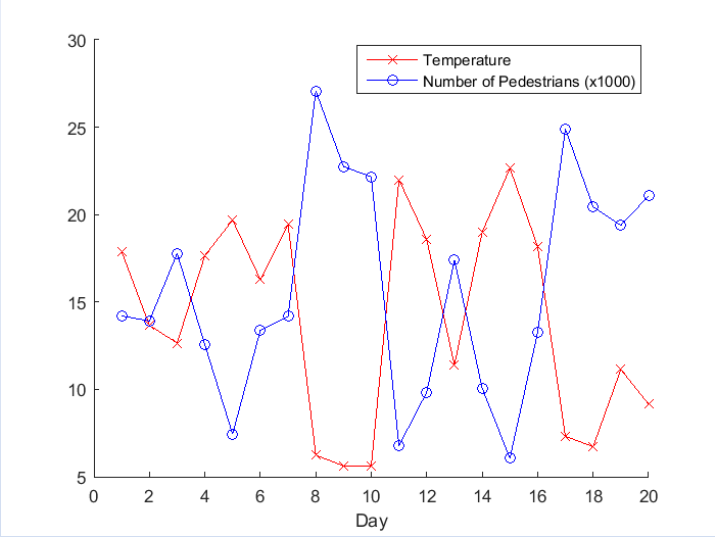
but since they are in different scales, we cannot see it



Problems of Euclidean distance

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

distance won't help



Pearson

plot data

see linear relationship → Pearson

else → MI

if you get a Pearson correlation of +1, it means the data has high (perfect) correlation.

It should mean a strong relationship, but some extreme example where the line is monotonic, but flat

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

Pearson's correlation coefficient (r)

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

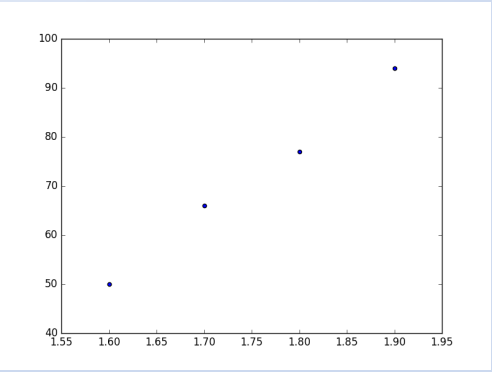
normalize

- x and y are the two attributes in your dataset
- Sample means $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

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

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

① slope = strength
how strong the correlation
② how good the line fits to the dot is
how confident you are
③ how close the dots are to each other is not important

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

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

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

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

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

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

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) → you get a parallel line of same slope
  - Adding a constant K to one feature's values makes no difference
- Can only detect linear relationships

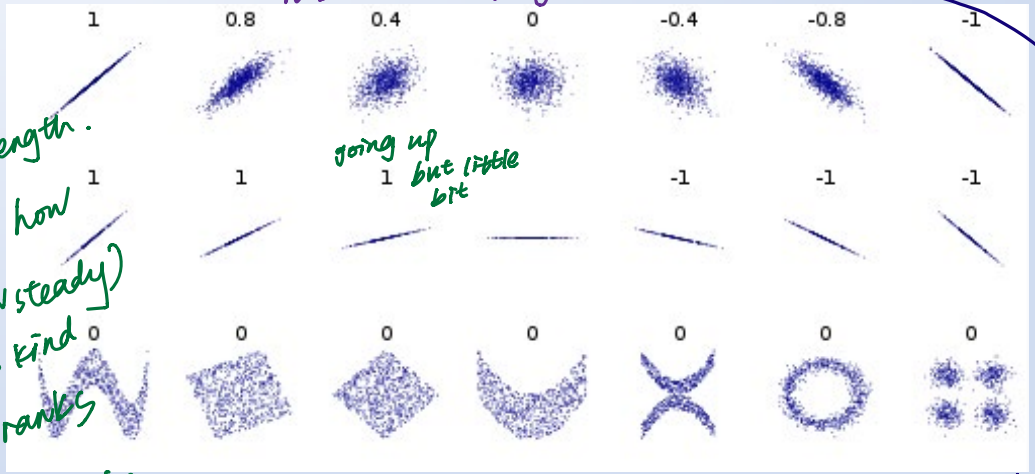
  $y = a.x + b + noise$
- Cannot detect non-linear relationships

  $y = x^3 + noise$

Examples

pearson → will give both the direction and also the strength

MI → only give strength



going up but little bit

? There are examples where slope is not the same as strength.

Pearson measures how monotonic (eg how steady) the slope, it is kind of looking at ranks

So the measure can be misleading.

but in general slope = strength

however

① at least 0, no upper bound no direction

② higher MI higher correlation

NMI can compare parameter

but for MI, NMI, if you do the histgram, you may

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

also get direction

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

→ categorical feature
can't do person correlation

linear feature → require numerical

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

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

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.

can only measure

limitation: ① linear relationship

maybe actually non-linear relationship

② 0.85 will be / or — or —

maybe very small difference

in which case time ↑ doesn't change grade

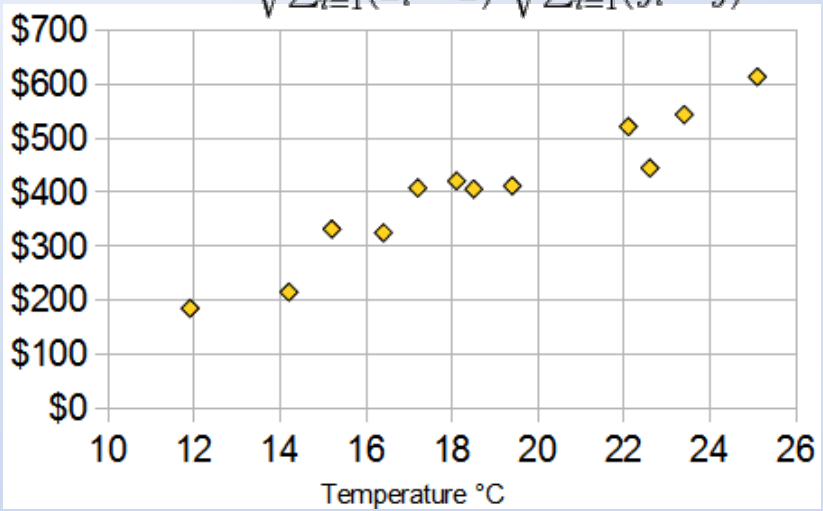
maybe
data based on some outliers.

you may have few points which are highly correlated and the rest is not correlated

Mutual Information

Recap: Pearson correlation – assess linear correlation between two features

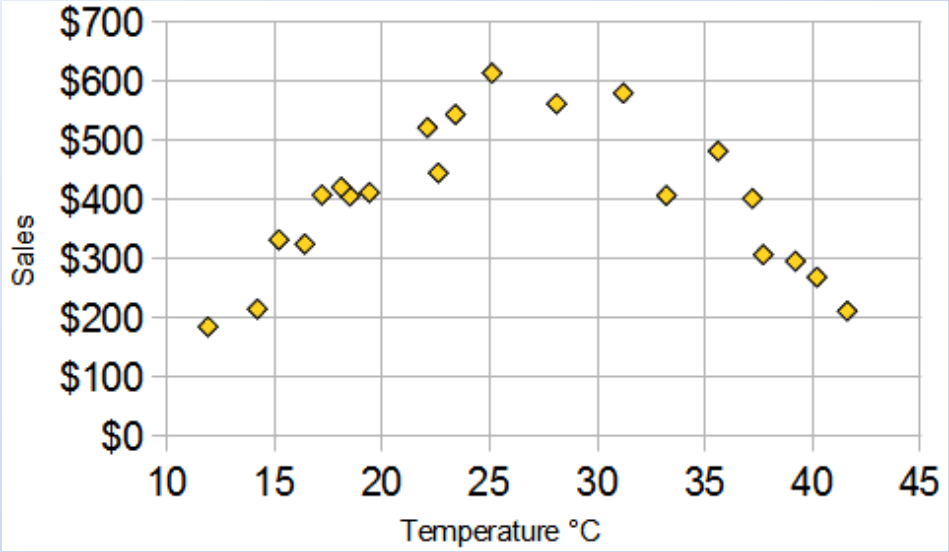
$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



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

What about non-linear correlation?

Pearson correlation is not suitable for this scenario (value less than 0.1)



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

Introduction to Mutual Information

A correlation measure that can detect non-linear relationships

- It operates with discrete features
- **Pre-processing**: continuous features are first discretised into bins (categories). E.g. small [0,1.4], medium (1.4,1.8), large [1.8,3.0]

1) more calculation
2) less intuitive
3) downsides: before using, need to discritise

Object	Height	Discretised Height
1	2.03	large
2	1.85	large
3	1.23	small
4	1.31	small
5	1.72	medium
6	1.38	small
7	0.94	small

numeric value → categories
bin values

Variable discretisation: Techniques

- Domain knowledge: assign thresholds manually
  - Car speed:
    - 0-40: slow
    - 40-60: medium
    - >60: high

bin → not same size → since we have domain knowledge

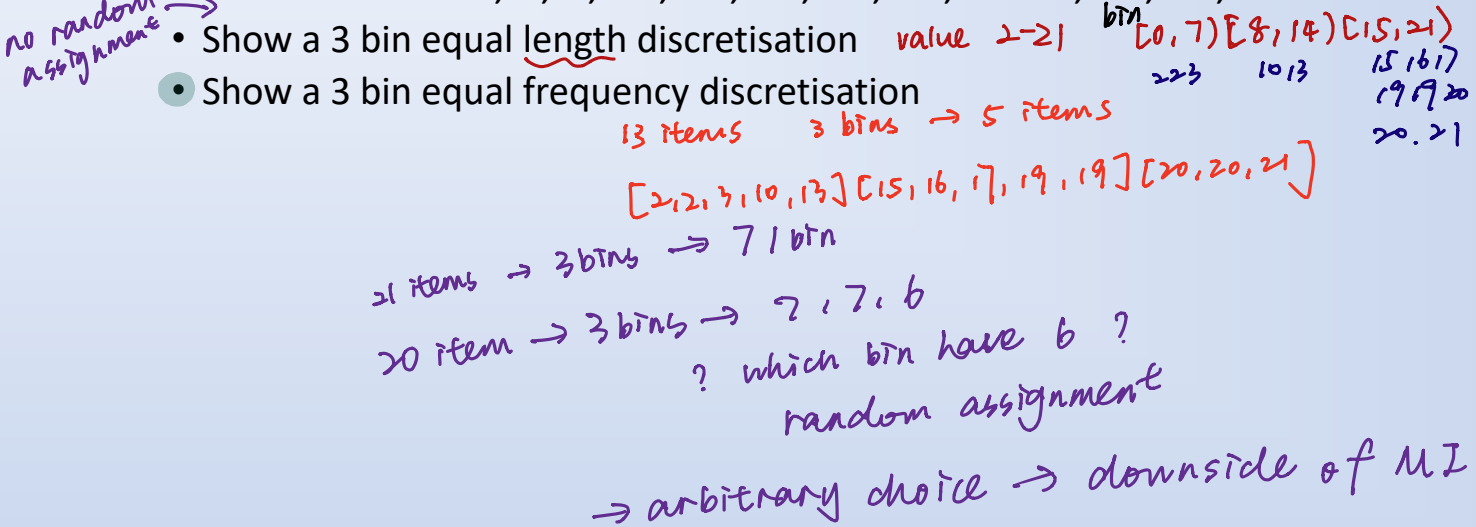
- Equal-length bin
  - Divide the range of the continuous feature into equal length intervals (bins). If speed ranges from 0-100, then the 10 bins are [0,10), [10,20), [20,30), …[90,100]

- Equal frequency bin
  - Divide range of continuous feature into equal frequency intervals (bins). Sort the values and divide so that each bin has same number of objects

how many items do we need to categorize eg. 50 item, each bin for 5 → 10 bins

Discretisation example

- Given the values 2, 2, 3, 10, 13, 15, 16, 17, 19 19, 20, 20, 21
- Show a 3 bin equal length discretisation
- Show a 3 bin equal frequency discretisation



A recap on logarithms (to the base 2)

- $y = \log_2 x$ (y is the solution to the question "To what power do I need to raise 2, in order to get x?'')
- $2*2*2*2=16$ which means $\log_2 16 = 4$ (16 is 2 to the power 4)
- $\log_2 32 = 5$
- $\log_2 30 = 4.9$
- $\log_2 1.2 = 0.26$
- $\log_2 0.5 = -1$
- In what follows, we'll write log instead of $\log_2$

Entropy

- Entropy is a measure used to assess the amount of uncertainty in an outcome

- E.g. Randomly select an element from {1,1,1,1,1,1,1,2}, versus randomly select an element from {1,2,2,3,3,4,5}

① entropy → low

② entropy → high

  - In which case is the value selected more "predictable"? Why?

① the content in first dataset is coherent, homogenous

Entropy → how mixed up of data

- E.g. Randomly select an element from {1,1,1,1,1,1,1,2}, versus randomly select an element from {1,2,2,3,3,4,5}
  - In which case is the value selected more "predictable"? Why?
  - The former case more certain => low entropy
  - The latter case is less certain => higher entropy
  - Entropy is used to quantify this degree of uncertainty (surprisingness)

Another example

- Consider the sample of all people in this lecture theatre. Each person is labelled young (<30 years) or old (>30 years)

- Randomly select a person and inspect whether they are young or old.
  - How surprised am I likely to be by the outcome?

- Suppose I repeat the experiment using a random sample of people catching the train to the city in peak hour?
  - How surprised am I likely to be by the outcome?

Entropy

① if everything is equally distributed among all bins → maximum entropy

② everything in one bin, other empty → minimum entropy

- Given a feature **X**. Then *H(**X**)* is its entropy. Assuming **X** uses a number of categories (bins)
- May sometimes write p(i) instead of p

$$H(\mathbf{X}) = -\sum_{i=1}^{\#bins} p_i \log p_i$$

- p_i: proportion of points in the i-th bin

Entropy of a Random Variable

$$H(\mathbf{X}) = -\sum_{i=1}^{\#bins} p_i \log p_i$$

- p_i: proportion of points in the i-th bin

- E.g. Suppose there are 3 bins, each bins contains exactly one third of the objects (points)
- H(X)=- [0.33 x log(0.33) + 0.33 x log(0.33) + 0.33 x log(0.33)]

Another Entropy example

A	B	B	A	C	C	C	C	A

We have 3 categories/bins (A,B,C) for a feature **X**

9 objects, each in exactly one of the 3 bins

What is the entropy of this sample of 9 objects?

$-3 \times \frac{1}{3} \log_2 \frac{1}{3}$

Answer: H(**X**)=1.53

$-[$

How would you compute the entropy for the “Likes to sleep” feature?

Person	Likes to sleep
1	Yes
2	No
3	Maybe
4	Never
5	Yes
6	Yes
7	Never
8	Yes

4 bins

yes bin 4

No bin 1

maybe bin 1

Never bin 2

Properties of entropy

- $H(X) \geq 0$ H(X)=0 perfectly sorted
- Entropy – when using log base 2 – measures uncertainty of the outcome in bits. 0 or 1 This can be viewed as the information associated with learning the outcome
- Entropy is maximized for uniform distribution (highly uncertain what value a randomly selected object will have)

Conditional entropy - intuition

Suppose I randomly sample a person. I check if they wear glasses – how surprised am I by their age?

Person	WearGlasses(X)	Age (Y)
1	No	young
2	No	young
3	No	young
4	No	young
5	Yes	old
6	Yes	old
7	Yes	old

conditional entropy → 0

if I know age, perfectly predict wearGlasses

if we have highly correlated features, maybe just

remove all uncertainty

want to use
→ we cannot get addit
information

Conditional entropy H(Y|X)

Measures how much information is needed to describe outcome Y, given that outcome X is known. Suppose X is Height and Y is Weight.

Object	Height (X)	Weight (Y)
1	big	light
2	big	heavy
3	small	light
4	small	light
5	small	light
6	small	light
7	small	heavy

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$$

Conditional entropy

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$$

Object	Height (X)	Weight (Y)
1	big	light
2	big	heavy
3	small	light
4	small	light
5	small	light
6	small	light
7	small	heavy

big small

H(Y|X) = 2/7 * H(Y|X=big) + 5/7 * h(Y|X=small)

= 2/7(-0.5log 0.5 -0.5 log 0.5) + 5/7(-0.8log 0.8-0.2log 0.2)

= 0.801

Mutual information definition

$$MI(X,Y) = H(Y) - H(Y|X)$$
$$= H(X) - H(X|Y)$$

- Where X and Y are features (columns) in a dataset

- MI (mutual information) is a measure of correlation
  - The amount of information about X we gain by knowing Y, or
  - The amount of information about Y we gain by knowing X

Mutual information example

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$$

Object	Height (X)	Weight (Y)
1	small	light
2	big	heavy
3	small	light
4	small	light
5	small	light
6	small	light
7	small	heavy

H(Y) = 0.8631205
H(Y|X) = 0.5572

H(Y)-H(Y|X) = 0.306 (how much information about Y is gained by knowing X)

Mutual information example 2

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$$

Object	Height (X)	Weight (Y)
1	big	light
2	big	heavy
3	small	light
4	small	jumbo
5	medium	light
6	medium	light
7	small	heavy

H(Y) = 1.379 → more messing
H(Y|X) = 0.965

H(Y)-H(Y|X) = 0.414

Properties of Mutual Information

- The amount of information shared between two variables X and Y

- MI(X,Y)
  - large: X and Y are highly correlated (more dependent)
  - small: X and Y have low correlation (more independent)

- $0 \leq$ MI(X,Y)

- Sometimes also referred to as 'Information Gain'

Mutual information: normalisation

- MI(X,Y) is always at least zero, may be larger than 1

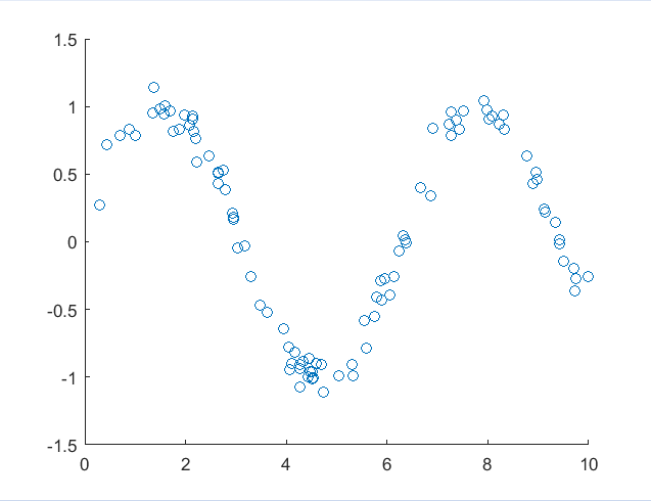
- In fact, one can show it is true that
  - $0 \leq MI(X,Y) \leq \min(H(X),H(Y))$
  - (where min(a,b) indicates the minimum of a and b)

- Thus if want a measure in the interval [0,1], we can define normalised mutual information (NMI)
  - $NMI(X,Y) = MI(X,Y) / \min(H(X),H(Y))$

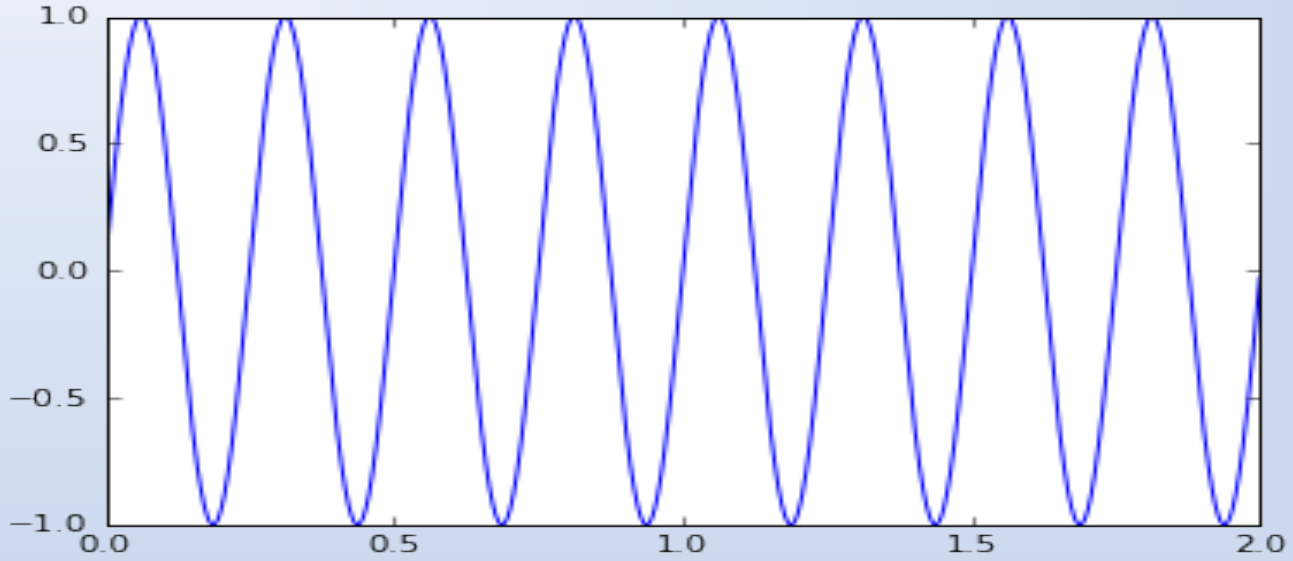
- NMI(X,Y)
  - large: X and Y are highly correlated (more dependent)
  - small: X and Y have low correlation (more independent)

MI is how much we learn about data
→ the most we can learn about it is the entropy in the dataset
→ move all uncertainty we can most learn total entropy of X or total entropy of Y (the min of that)

Pearson correlation=-: -0.0864
Normalised mutual information (NMI)= 0.43
(3-bin equal spread bins)



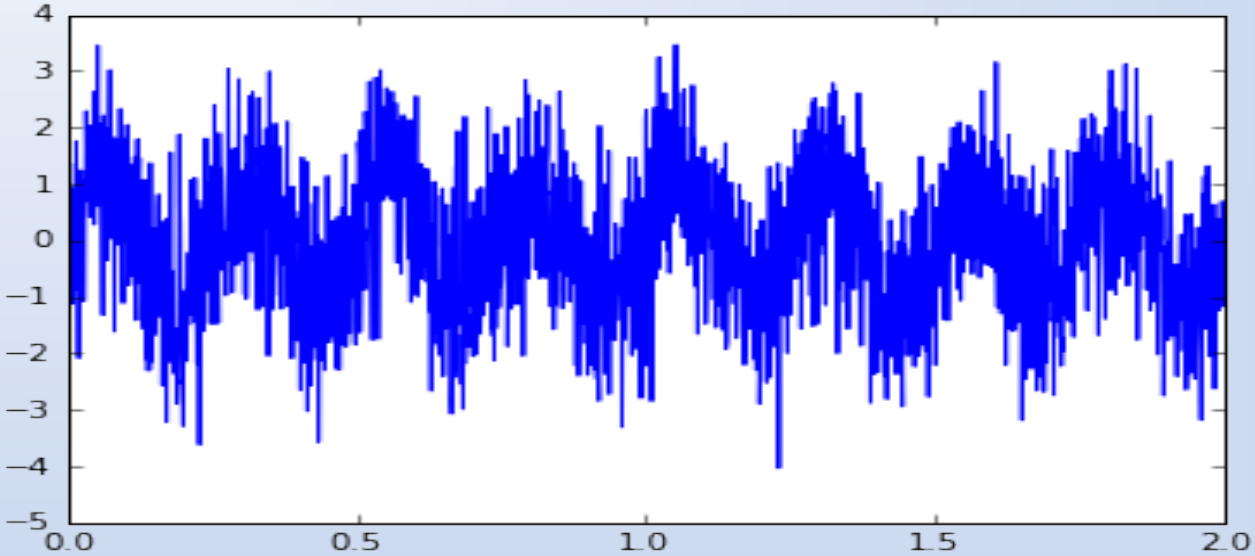
Pearson correlation=-0.1
Normalised mutual information (NMI)=0.84



Pearson correlation=-0.05 → fail

Normalised mutual information (NMI)=0.35

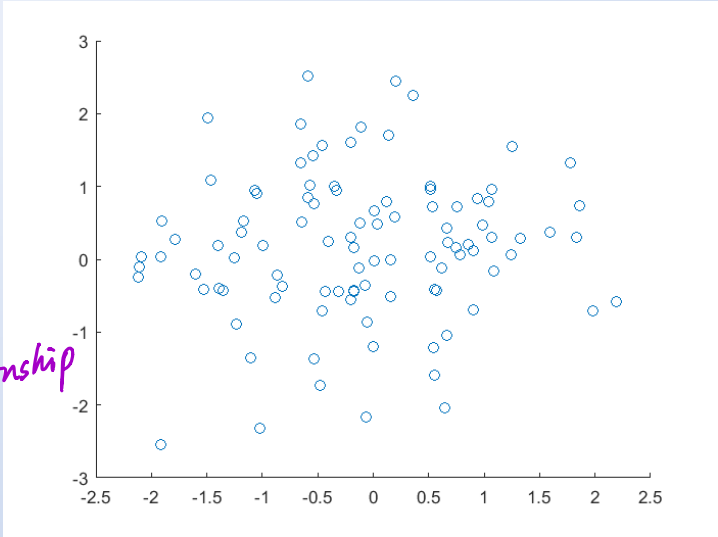
NMI → better for non-linear relationship



Examples

- Pearson?
- NMI?

for Pearson
→ no linear relationship

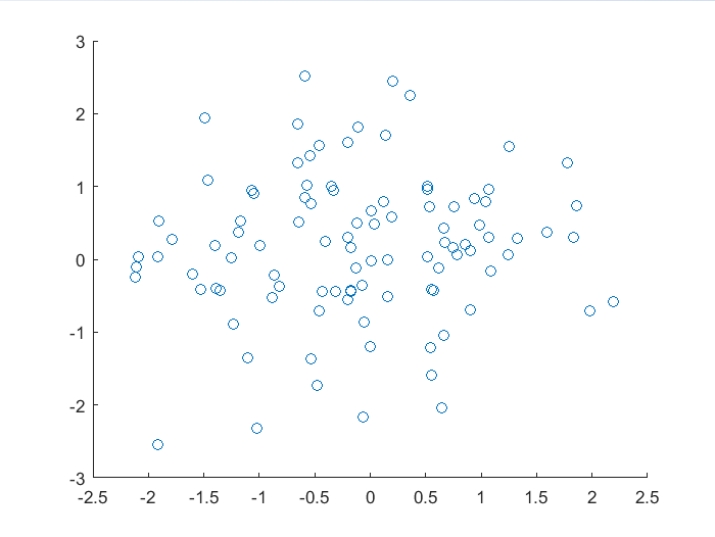


double entropy
≠ double uncertainty

we cannot say that
an entropy of 0.5
is half as uncertain
as an entropy of 1

Examples

- Pearson: 0.08
- NMI: 0.009



Computing MI with class features

Identifying features that are highly correlated with a class feature

HoursSleep	HoursExercise	HairColour	HoursStudy	Happy (class feature)
12	20	Brown	low	Yes
11	18	Black	low	Yes
10	10	Red	medium	Yes
10	9	Black	medium	Yes
10	10	Red	high	No
7	11	Red	high	No
6	15	Brown	high	No
2	13	Brown	high	No

Compute MI(HoursSleep, Happy), MI(HoursExercise, Happy),
and MI(HoursStudy, Happy), MI(HairColour, Happy). Retain most predictive feature(s)

Computing MI with class features

HoursSleep	HoursExercise	HairColour	HoursStudy	Happy (class feature)
12	20	Brown	low	Yes
11	18	Black	low	Yes
10	10	Red	medium	No
10	9	Black	medium	Yes
10	10	Red	high	No
7	11	Black	high	No
6	15	Brown	high	No
2	13	Brown	high	No

- MI(HoursStudy, Happy)=0.70 (NMI=0.74)
- ….
- Can rank features according to their predictiveness –then focus further on just these

Advantages and disadvantages of MI

- Advantages
  - Can detect both linear and non-linear dependencies (unlike Pearson)
  - Applicable and very effective for use with discrete features (unlike Pearson correlation)

- Disadvantages
  - If feature is continuous, it first must be discretised to compute mutual information. This involves making choices about what bins to use.
    - This may not be obvious. Different bin choices will lead to different estimations of mutual information.

Choose the statement that is not true

- Mutual information can't detect the existence of linear relationships in the data, but Pearson correlation can
- Mutual information can be used to assess how predictive a feature is of an outcome
- Mutual information can detect the existence of non-linear relationships in the data
- Mutual information doesn't indicate the "direction" of a relationship, just its strength

one ↑ one ↓ → same MI
one ↑ one ↑

Question 2ai) from 2016 exam

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 spent studying	Average Grade
…	….	….

i) 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.

Question 2aii) from 2016 exam

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 spent studying	Average Grade
…	….	….

ii) Richard separately discretises the two features *Average time per day spent studying* and *Average grade*, each into 2 bins. He then computes the normalised mutual information between these two features and obtains a value of 0.1, which seems surprisingly low to him. Suggest two reasons that might explain the mismatch between the normalised mutual information value of 0.1 and the Pearson Correlation coefficient of 0.85. Explain any assumptions made.

① slope is very low

may cause

positive relationship

but actually MI low

② most students don't have a relationship but for few, strong relationship → high Pearson

③ way we discretise is wrong (bin is not correct)

Points to remember (1)

- 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 nonlinear 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

Points to remember (2)

- 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

linear measure → not good for non-linear

not tell you strength of relationship, tell the direction of relationship

Points to remember (3)

- understand the advantages and disadvantages of using mutual information for computing correlation between a pair of features. Understand the main differences between this and Pearson correlation.
- understand the meaning of the variables for mutual information and how they can be calculated. Be able to compute this measure on a simple pair of features. The formula for mutual information will be provided in the exam.
- understand the role of data discretisation in computing mutual information
- understand the meaning of the entropy of a random variable and how to interpret an entropy value. Understand its extension to conditional entropy
- be able to interpret the meaning of the mutual information between two features

no direction → just strength

Points to remember (4)

- understand the use of mutual information for computing correlation of some feature with a class feature and why this is useful. Understand how this provides a ranking of features, according to their predictiveness of the class

- understand that normalised mutual information can be used to provide a more interpretable measure of correlation than mutual information. The formula for normalised mutual information will be provided on the exam

Acknowledgements

- Materials are partially adopted from …
  - Previous COMP2008 slides including material produced by James Bailey, Pauline Lin, Chris Ewin, Uwe Aickelin and others
  - Interactive correlation calculator: http://www.bc.edu/research/intasc/library/correlation.shtml
  - Correlation <> Causality: http://tylervigen.com/spurious-correlations
  - Google trends correlation