THE UNIVERSITY OF MELBOURNE

COMP20008 Elements of Data Processing

Assessing Correlations cont. – Mutual Information

Announcements

- Consultation session with Donia for python programming
  - Monday 3 April (today): 11am-12pm Rm 10.22 Doug McDonell Building (10th floor)

- Students with a workshop scheduled on the Good Friday University holiday (Friday 14th April)
  - This workshop will not go ahead due to the holiday
  - We have scheduled (an abbreviated) replacement workshop to be held Alan Gilbert-111 Monday 10th April 10am-11am. You are welcome to attend this if you are enrolled in the Friday 14th April workshop class.

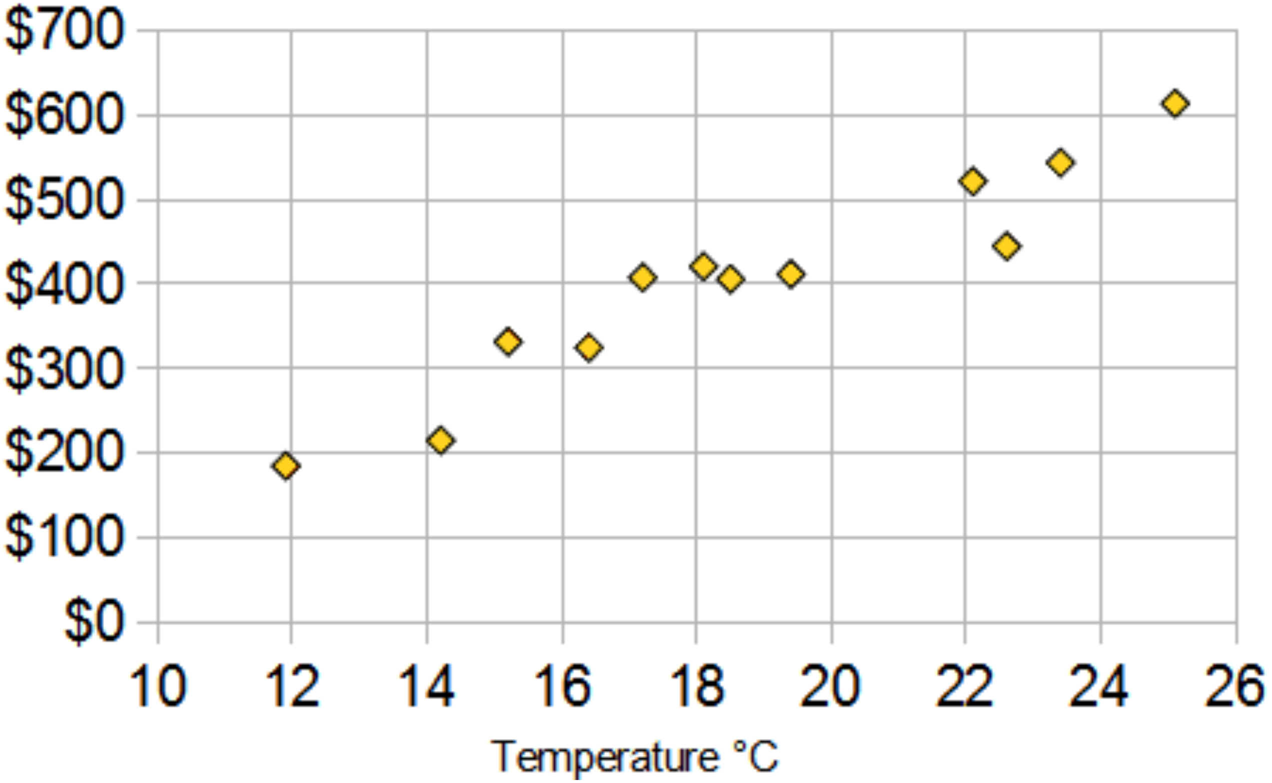
Plan today

- Recap – correlations
  - Pearson correlation

- Another measure for correlation
  - Mutual information
    - Entropy
    - Conditional entropy

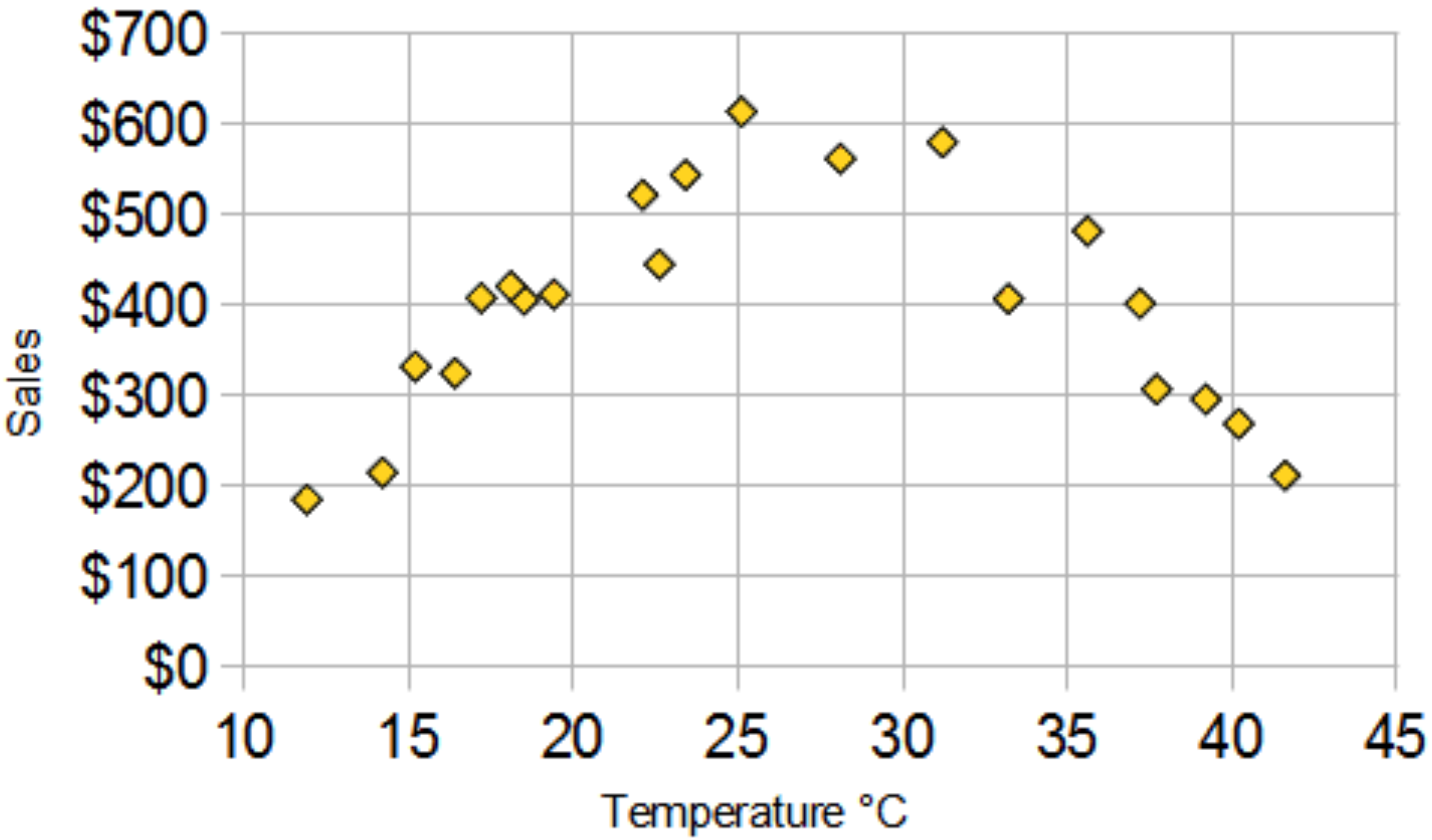
Pearson correlation – assess degree of linear correlation between two features

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



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

How do we measure 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 cprrelation measure that can detect non-linear relationships
  - It operates with discrete features
  - continuous features are first discretised into bins (categories). E.g. small [0,1.4], medium (1.4,1.8), big [1.8,3.0]

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

Variable discretization: Techniques

- Domain knowledge: assign thresholds manually
  - Speed:
    - 0-40: slow
    - 40-60:mid
    - >60: high

- Equal-width 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.

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

Entropy: A basic building block we will need

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

- 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

Entropy

- Given a feature **X**. Then *H(**X**)* is its entropy. Assuming X uses a number of categories(bins)

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

- 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)=?

Entropy of 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)]
- The log can be any base, we will assume base 2

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 bin

What is the entropy of this sample of 9 objects?

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

Properties of the entropy

- $H(X) \geq 0$

- Entropy – when using log base 2 – measures uncertainty of the outcome in bits. 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 H(Y|X)

- Measures how much information 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

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

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

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

Conditional entropy

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

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

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

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

Conditional entropy example

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

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

H(Y|X)=0.965
H(Y)=1.379

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

Mutual information definition

$$\begin{aligned} MI(X,Y) &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \end{aligned}$$

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

- MI 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

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)

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

Mutual information: properties and normalisation

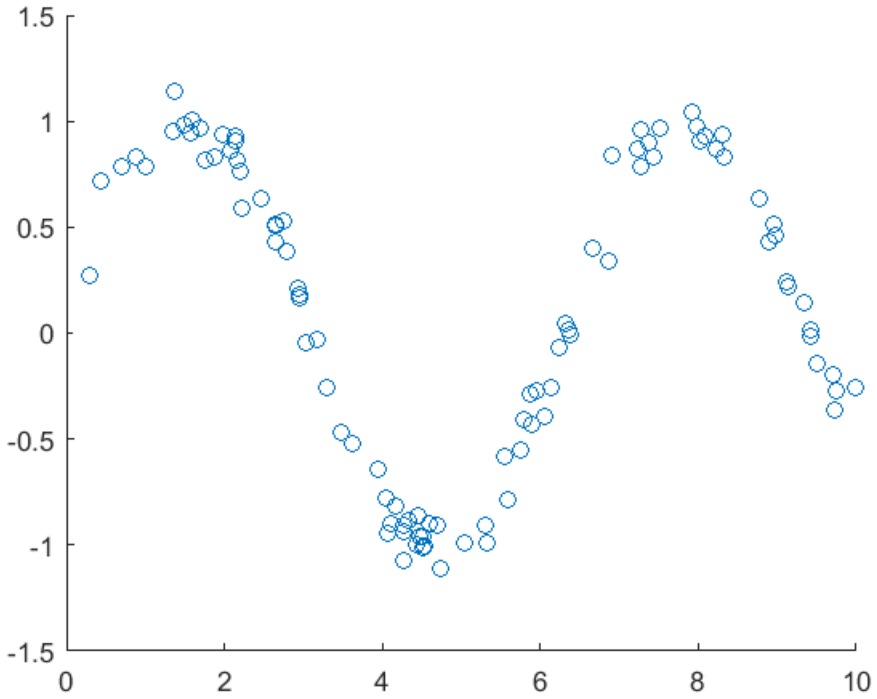
- MI(X,Y) is always at least zero, may be larger than 1

- In fact, one can show it is true that
  - $0 \leq MI(X,Y) \leq \min(H(X),H(Y))$

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

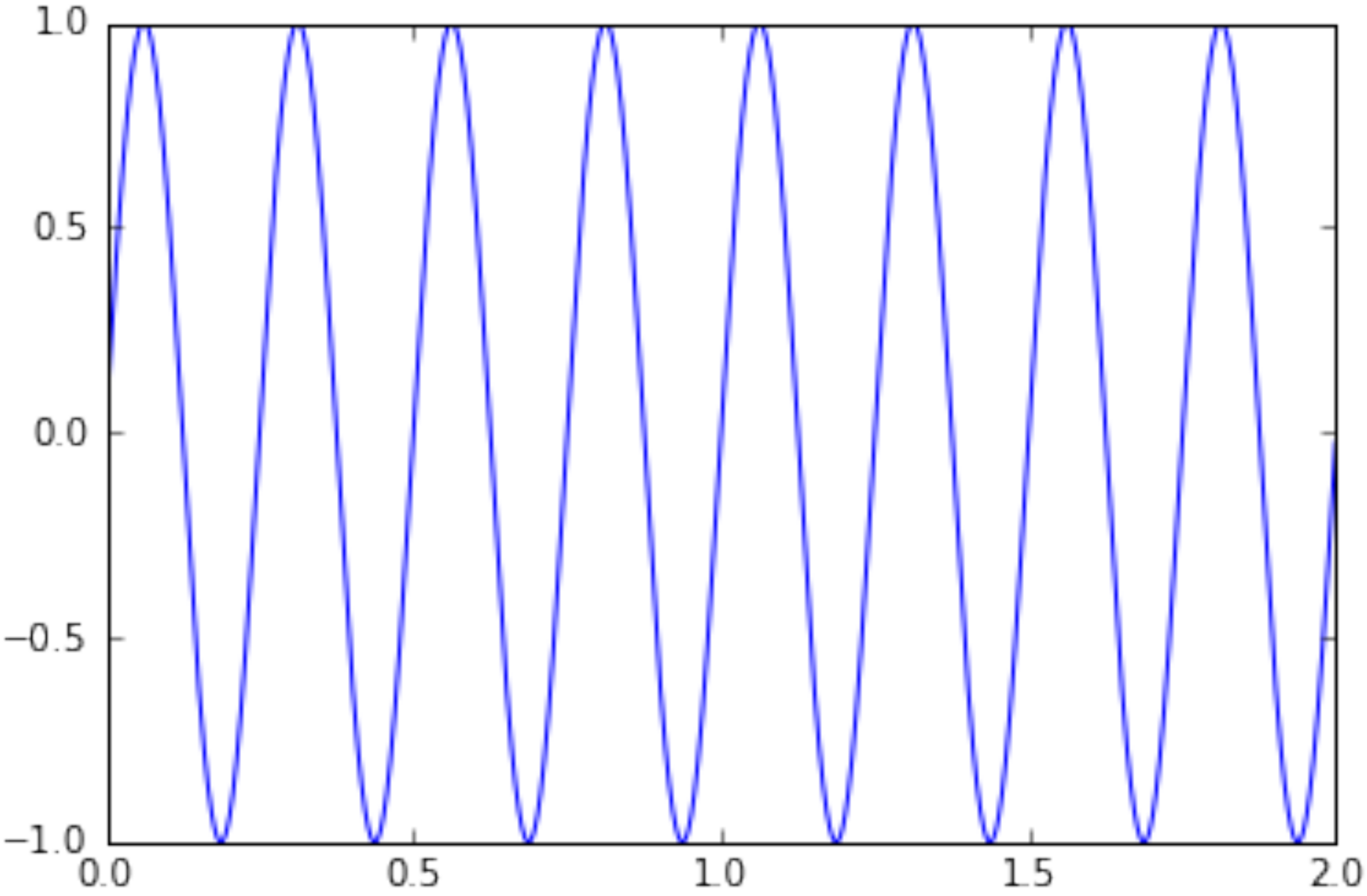
- NMI(X,Y)
  - large: X and Y are highly correlated (more dependent)
  - small: X and Y have low correlation (more independent)

Examples

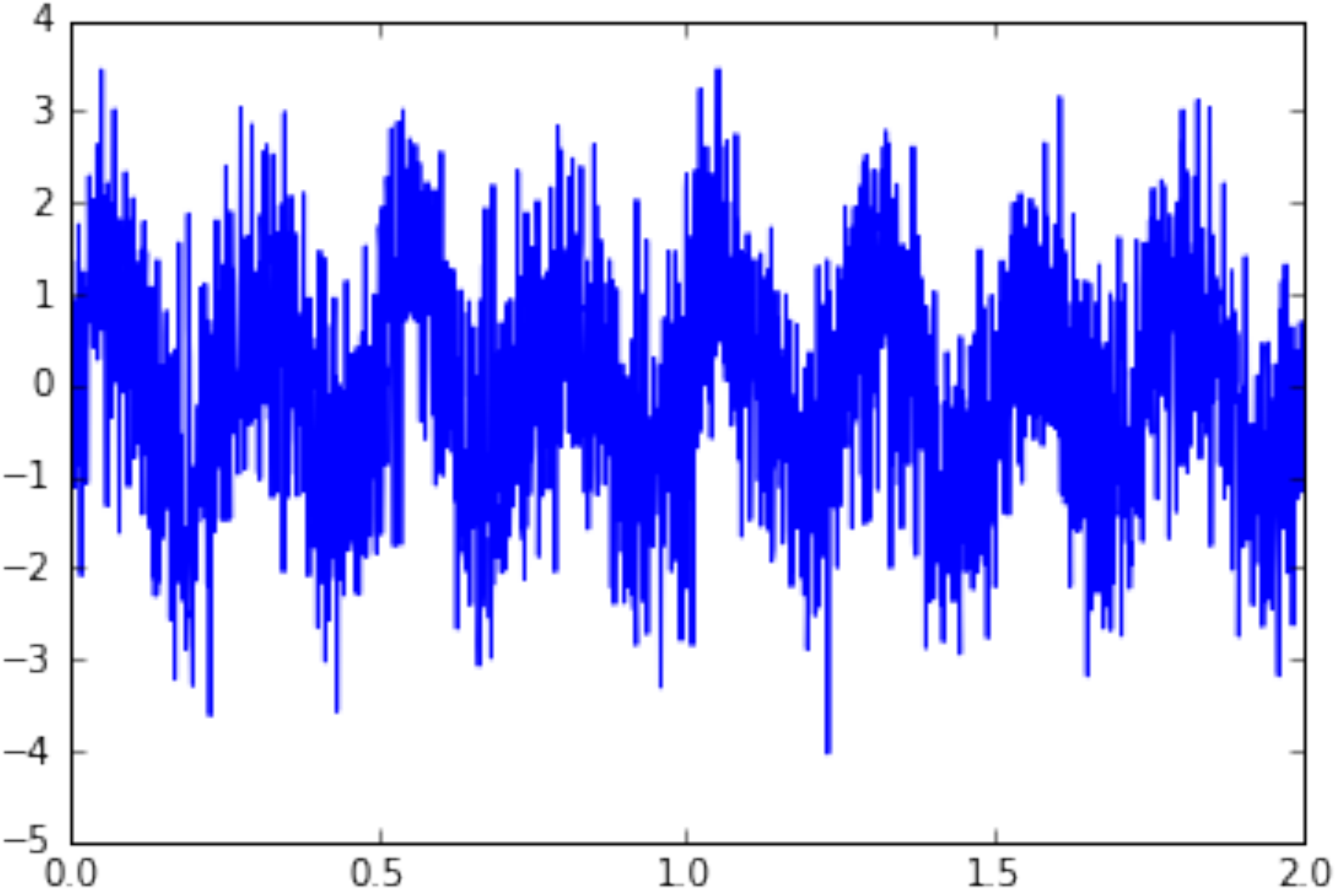


- Pearson: -0.0864
- NMI: 0.43 (3-bin equal frequency discretization)

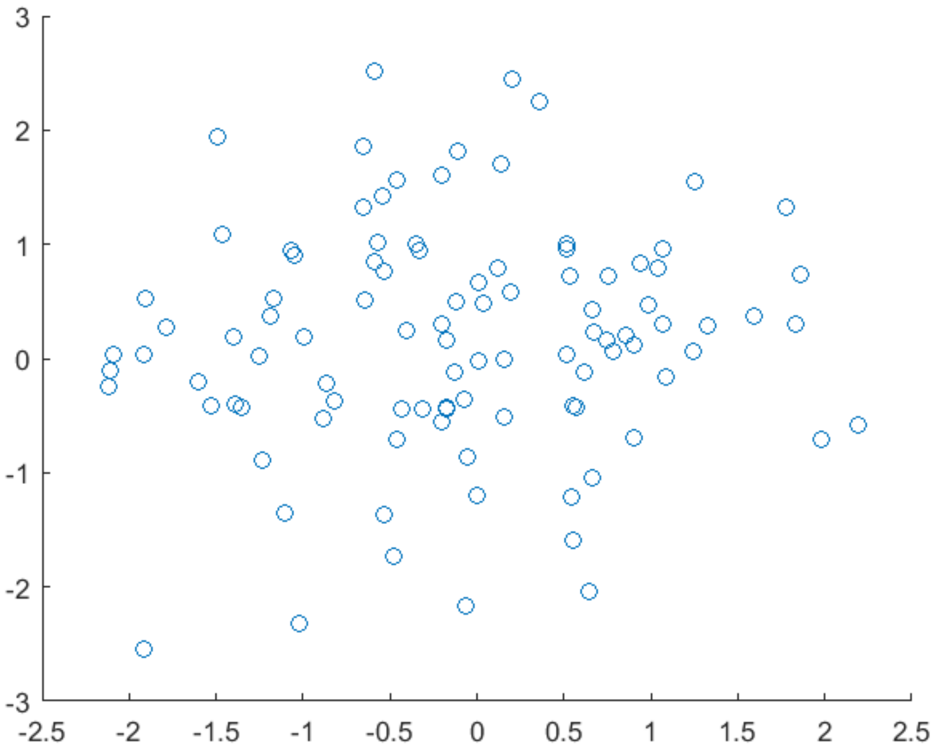
Pearson correlation=-0.1
Normalised mutual information (NMI)=0.84



Pearson correlation=-0.05
Normalised mutual information (NMI)=0.35

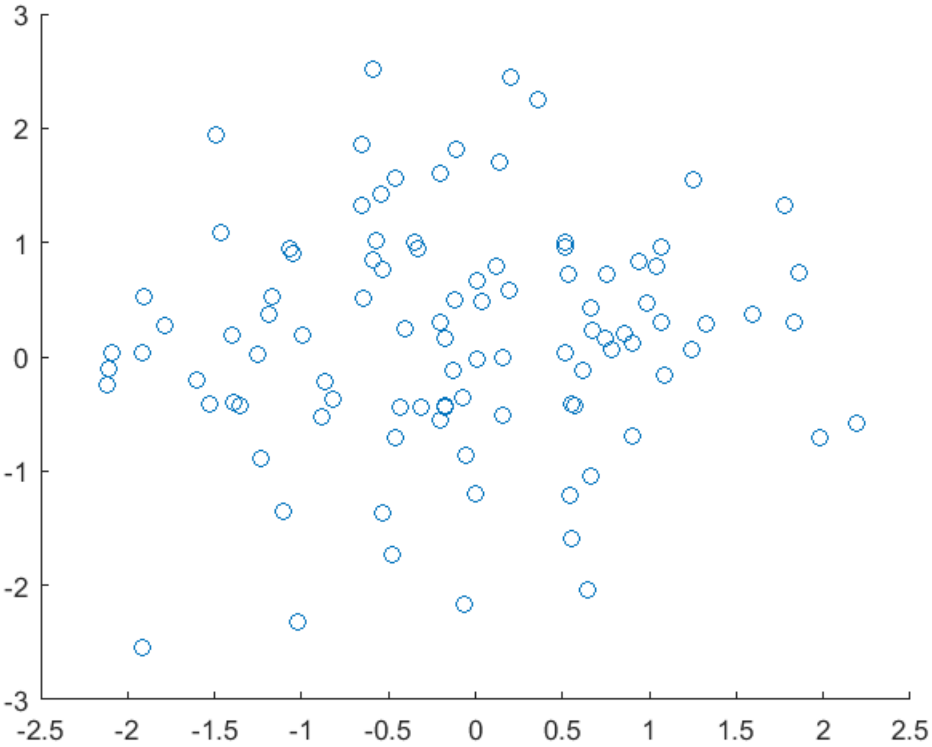


Examples



- Pearson?
- NMI?

Examples



- Pearson: 0.08
- NMI: 0.009

THE UNIVERSITY OF MELBOURNE

Computing correlation with a (discrete) class feature

- Identifying features that are highly correlated with a class feature

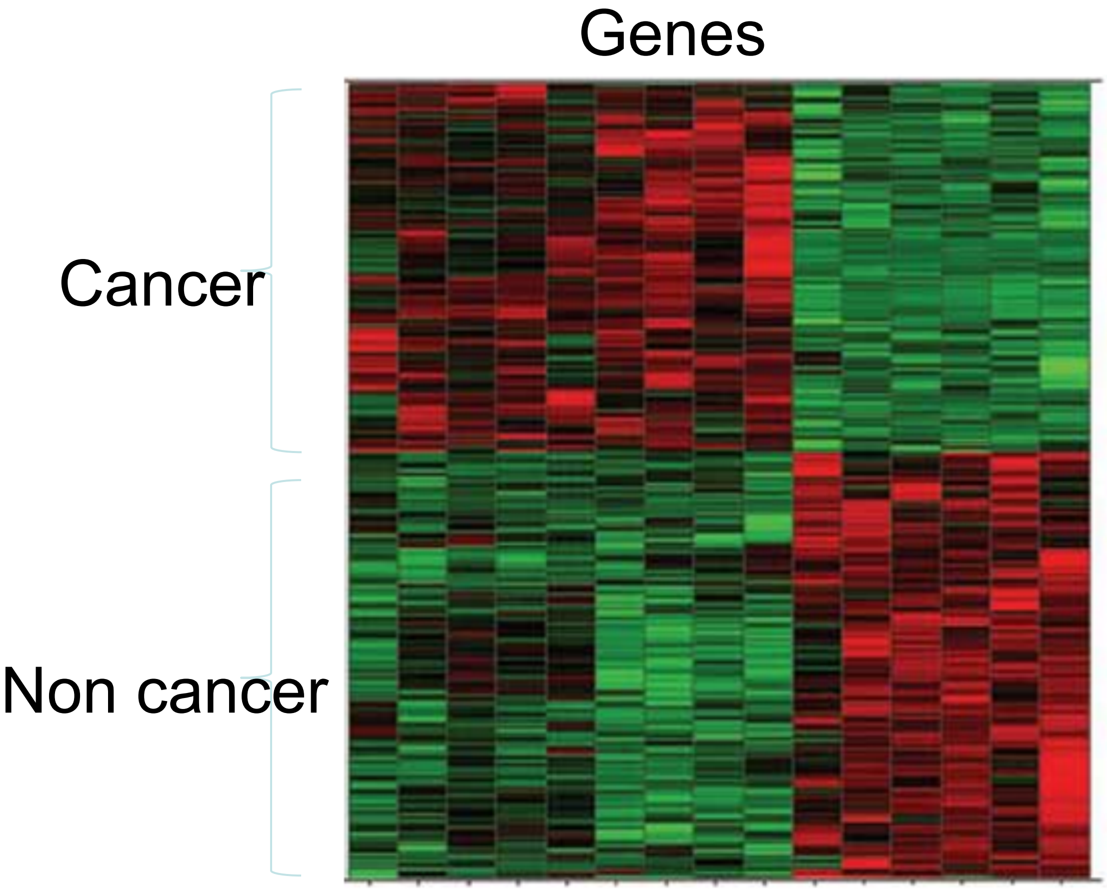
HoursSleep	HoursExercise	HairColour	HoursStudy	Happy (class feature)
12	20	Brown	low	Yes
11	18	Black	low	Yes
10	10	Red	medium	Yes
10	9	Black	medium	Yes
10	10	Red	high	No
7	11	Red	high	No
6	15	Brown	high	No
2	13	Brown	high	No

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

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(HairColour,Happy)=0.27 (NMI=0.28)
- MI(HoursStudy,Happy)=0.70 (NMI=0.74)
- ….
- Can rank features according to their predictiveness –then focus further on just these

Which genes correlate with cancer?



Feature (gene) selection

- Cancer = f(gene1, gene2, … , gene n)
- Use correlation to reduce the number of variables

	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

Cancer
1
1
0
…
0

- User relevant genes only: improving accuracy & performance

Advantages and disadvantage of mutual information

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

- Disadvantage
  - 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

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

- 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.

Points to take away

- 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 in the 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 on the exam.
- understand the role of data discretization 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

Points to take away (continued)

- 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