Workshop 7

COMP20008

Elements of Data Processing
Zijie Xu

Agenda

- A2 Team forming
- Correlation
  - Pearson correlation
  - Mutual Information

A2 Team forming

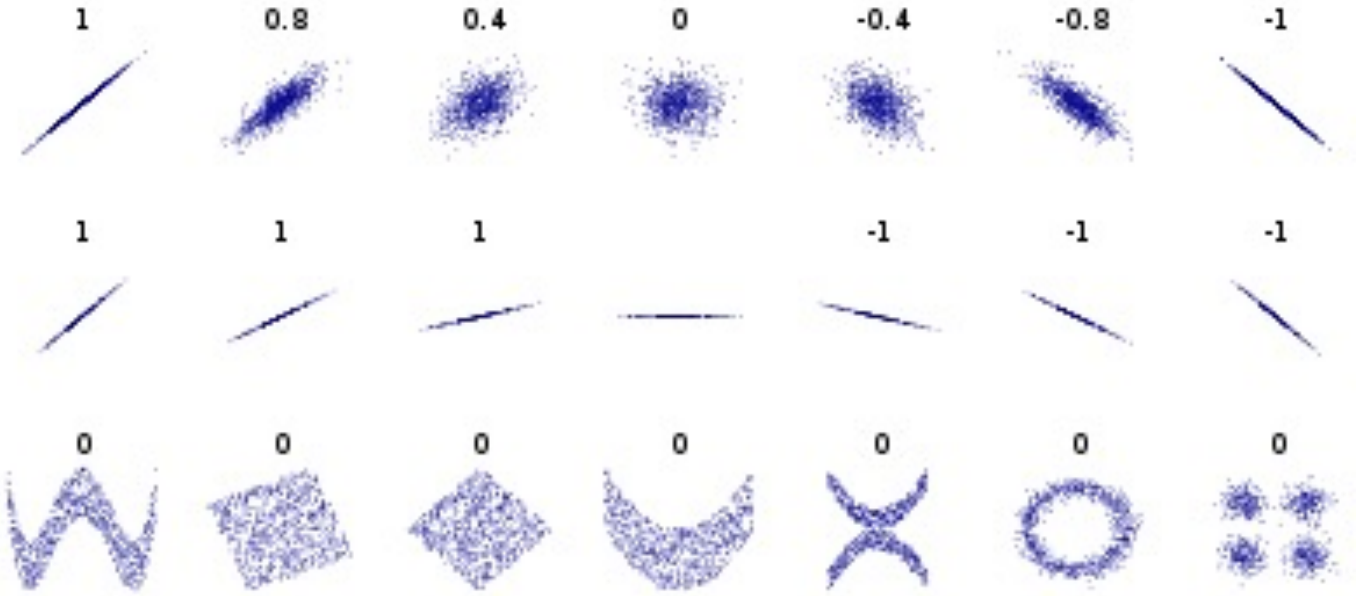
- Assignment 2 has been/will be released today!
- Teams of 3-4 students
- All team members need to be from same workshop
- Register your team on Google sheet at the end of today's workshop

Correlation

- Correlation is the mutual relationship between two random variables
- Correlation does not imply causality
- Two correlation metrics to discuss in particular
  - Pearson correlation coefficient r_{xy}
  - Mutual information

Pearson correlation coefficient r_{xy}

- Pearson correlation coefficient r_{xy} measures the strength and direction of a linear relationship between two random variables
- Assumptions: continuous, linear relationship, no spurious outliers, normally distributed



Pearson correlation coefficient r_{xy}

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- Sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - Estimate $\mu_x = \mathrm{E}[X]$
  - Arithmetic mean is the average value of a variable
  - n Is the sample size
  - x_i are the individual data points with index i

Pearson correlation coefficient r_{xy}

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- Sample covariance

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

  - Estimate $\mathrm{Cov}(X, Y) = \sigma_{xy} = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$
  - Covariance measures the linear relationship of how much two variables change together
  - Technically, need to devide by $n - 1$ to remove bias

Pearson correlation coefficient r_{xy}

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- Sample variance

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  - Estimate $\sigma_x^2 = \mathrm{E}[(X - \mathrm{E}[X])^2]$
  - Variance is special case of covariance
  - Variance measures how far the spread of data is around its average
  - Take squre root to obtain sample standard deviation $s_x = \sqrt{s_x^2}$
  - Technically, need to devide by $n - 1$ to remove bias

Pearson correlation coefficient r_{xy}

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- Putting altogether…

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

  - Estimate $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$
  - Ranges from -1 to +1
  - -1 and +1 stand for perfect/strong linear relationships
  - Values near 0 implies a weak linear relationship

Binning

- Binning is used to transform continuous variables into a discrete form
- Data points fall into set of intervals that span across the range
- Two strategies to discuss in particular: Equal width & Equal frequency

- Example: 70, 73, 75, 78, 80, 85, 89, 91, 97

Equal width		Equal frequency	
[70,80)	70, 73, 75, 78	[70,78)	70, 73, 75
[80,90)	80, 85, 89	[80,89)	78, 80, 85
[90,100]	91, 97	[89,100]	89, 91, 97

Mutual information

- Mutual information measures the reduction in uncertainty for one variable given a known value of the other variable
- Works only on discrete random variables
- Can detect non-linear dependencies

$$\mathrm{MI}(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Entropy and conditional entropy

- Entropy is the average "uncertainty" inherent to a random variable

$$H(X) = \mathrm{E}[-\log_2(\Pr(X))] = -\sum_i \Pr(X = x_i) \log_2(\Pr(X = x_i))$$

- Conditional entropy measure the “uncertainty” of a random variable given that another random variable has occurred
  - Each $H(Y|X = x_i)$ is calculated using conditional probability $\Pr(Y = y_i|X = x_i)$ with the entropy formula

$$H(Y|X) = \sum_{x_i \in X} \Pr(X = x_i) \, H(Y|X = x_i)$$

Thank you

More Resources: Canvas