DATA MINING & MACHINE LEARNING (I)

Thiago Marzagão

IESB
Centro Universitário

Média

- $\dfrac{\sum x_i}{N}$
- É influenciada por valores extremos.

Moda

- É valor mais freqüente.
- Não é muito informativa quando a distribuição é multimodal.

Mediana

- É valor que divide a distribuição em duas metades.
- Não é muito informativa quando a distribuição é bimodal.

Variância (amostral)

- $s^2 = \dfrac{\sum (x_i - \bar{x})^2}{n-1}$

Desvio-padrão (amostral)

- $s = \sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n - 1}}$
- Vantagem: está na mesma unidade da variável.

como medir a associação entre duas variáveis?

- Como medir o quanto duas variáveis “andam juntas”?

como medir a associação entre duas variáveis?

- Como medir o quanto duas variáveis “andam juntas”?
- Exemplos:

como medir a associação entre duas variáveis?

- Como medir o quanto duas variáveis "andam juntas"?
- Exemplos:
- Anos de estudo X salário.

como medir a associação entre duas variáveis?

- Como medir o quanto duas variáveis "andam juntas"?
- Exemplos:
- Anos de estudo X salário.
- Horas de estudo por semana X nota na disciplina.

como medir a associação entre duas variáveis?

- Como medir o quanto duas variáveis "andam juntas"?
- Exemplos:
- Anos de estudo X salário.
- Horas de estudo por semana X nota na disciplina.
- Emissões de CO_2 X temperatura.

como medir a associação entre duas variáveis?

- Como medir o quanto duas variáveis “andam juntas”?
- Exemplos:
- Anos de estudo X salário.
- Horas de estudo por semana X nota na disciplina.
- Emissões de CO_2 X temperatura.
- Idade X acidentes de carro.

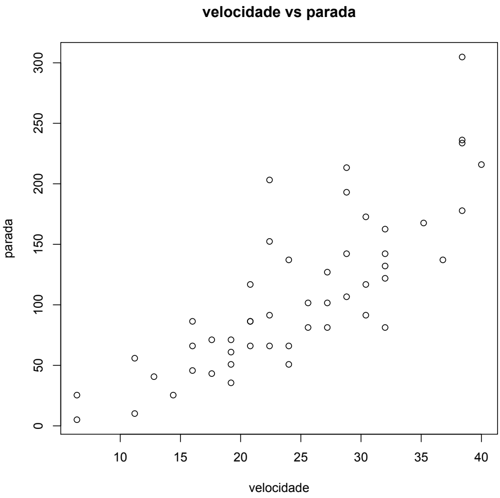
como medir a associação entre duas variáveis?

- Como medir o quanto duas variáveis "andam juntas"?
- Exemplos:
- Anos de estudo X salário.
- Horas de estudo por semana X nota na disciplina.
- Emissões de CO2 X temperatura.
- Idade X acidentes de carro.
- Consumo de carne vermelha X longevidade.

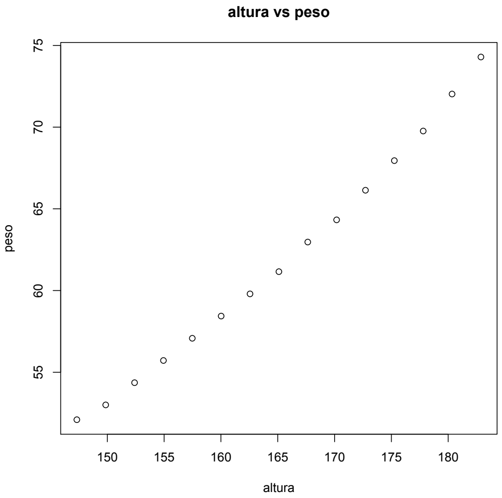
como medir a associação entre duas variáveis?

- Como medir o quanto duas variáveis "andam juntas"?
- Exemplos:
- Anos de estudo X salário.
- Horas de estudo por semana X nota na disciplina.
- Emissões de CO2 X temperatura.
- Idade X acidentes de carro.
- Consumo de carne vermelha X longevidade.
- Horas de academia por semana X peso.

Solução #1: gráfico de dispersão.



Solução #1: gráfico de dispersão.



Solução #2: covariância.

- Breve revisão: variância (amostral).
- $s^2 = \dfrac{\sum (y_i - \bar{y})^2}{n-1}$

Solução #2: covariância.

- Covariância (amostral):
- $s_{xy} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$
- s_{xy} nos diz o quanto as variáveis x e y “andam juntas”
- ... ou seja, o quanto x e y co-variam

Solução #2: covariância.

- altura:

Solução #2: covariância.

- altura:
- 147.32, 149.86, 152.40, 154.94, 157.48, 160.02, 162.56, 165.10, 167.64, 170.18, 172.72, 175.26, 177.80, 180.34, 182.88

Solução #2: covariância.

- altura:
- 147.32, 149.86, 152.40, 154.94, 157.48, 160.02, 162.56, 165.10, 167.64, 170.18, 172.72, 175.26, 177.80, 180.34, 182.88
- altura média: $\bar{x} = 165.1$

Solução #2: covariância.

- altura:
- 147.32, 149.86, 152.40, 154.94, 157.48, 160.02, 162.56, 165.10, 167.64, 170.18, 172.72, 175.26, 177.80, 180.34, 182.88
- altura média: $\bar{x} = 165.1$
- peso:

Solução #2: covariância.

- altura:
- 147.32, 149.86, 152.40, 154.94, 157.48, 160.02, 162.56, 165.10, 167.64, 170.18, 172.72, 175.26, 177.80, 180.34, 182.88
- altura média: $\bar{x} = 165.1$
- peso:
- 52.095, 53.001, 54.360, 55.719, 57.078, 58.437, 59.796, 61.155, 62.967, 64.326, 66.138, 67.950, 69.762, 72.027, 74.292

Solução #2: covariância.

- altura:
- 147.32, 149.86, 152.40, 154.94, 157.48, 160.02, 162.56, 165.10, 167.64, 170.18, 172.72, 175.26, 177.80, 180.34, 182.88
- altura média: $\bar{x} = 165.1$
- peso:
- 52.095, 53.001, 54.360, 55.719, 57.078, 58.437, 59.796, 61.155, 62.967, 64.326, 66.138, 67.950, 69.762, 72.027, 74.292
- peso médio: $\bar{y} = 61.94$

Solução #2: covariância.

- altura:
- 147.32, 149.86, 152.40, 154.94, 157.48, 160.02, 162.56, 165.10, 167.64, 170.18, 172.72, 175.26, 177.80, 180.34, 182.88
- altura média: $\bar{x} = 165.1$
- peso:
- 52.095, 53.001, 54.360, 55.719, 57.078, 58.437, 59.796, 61.155, 62.967, 64.326, 66.138, 67.950, 69.762, 72.027, 74.292
- peso médio: $\bar{y} = 61.94$
- (fazer tabela no quadro: x_i, y_i, $(x_i - \bar{x})$, $(y_i - \bar{y})$, $(x_i - \bar{x})(y_i - \bar{y})$)

Solução #2: covariância.

- Problema: como saber se uma dada covariância é grande ou pequena?
- Covariância depende da escala das duas variáveis.
- Como comparar duas covariâncias quando as escalas são diferentes?

Solução #3: correlação.

- $r_{xy} = \frac{s_{xy}}{s_x s_y}$
- $r_{xy} =$ coeficiente de correlação amostral
- $s_{xy} =$ covariância (amostral)
- $s_x =$ desvio-padrão amostral de x
- $s_y =$ desvio-padrão amostral de y
- r_{xy} varia sempre entre -1 e 1, não importa a escala das duas variáveis
- $r_{xy} = -1$: correlação negativa perfeita
- $r_{xy} = +1$: correlação positiva perfeita

Solução #3: correlação.

- Voltando ao exemplo do peso X altura:

Solução #3: correlação.

- Voltando ao exemplo do peso X altura:
- x: 147.32, 149.86, 152.40, 154.94, 157.48, 160.02, 162.56, 165.10, 167.64, 170.18, 172.72, 175.26, 177.80, 180.34, 182.88

Solução #3: correlação.

- Voltando ao exemplo do peso X altura:
- x: 147.32, 149.86, 152.40, 154.94, 157.48, 160.02, 162.56, 165.10, 167.64, 170.18, 172.72, 175.26, 177.80, 180.34, 182.88
- y: 52.095, 53.001, 54.360, 55.719, 57.078, 58.437, 59.796, 61.155, 62.967, 64.326, 66.138, 67.950, 69.762, 72.027, 74.292

Solução #3: correlação.

- Voltando ao exemplo do peso X altura:
- x: 147.32, 149.86, 152.40, 154.94, 157.48, 160.02, 162.56, 165.10, 167.64, 170.18, 172.72, 175.26, 177.80, 180.34, 182.88
- y: 52.095, 53.001, 54.360, 55.719, 57.078, 58.437, 59.796, 61.155, 62.967, 64.326, 66.138, 67.950, 69.762, 72.027, 74.292
- Já calculamos a covariância: 79.39

Solução #3: correlação.

- Voltando ao exemplo do peso X altura:
- x: 147.32, 149.86, 152.40, 154.94, 157.48, 160.02, 162.56, 165.10, 167.64, 170.18, 172.72, 175.26, 177.80, 180.34, 182.88
- y: 52.095, 53.001, 54.360, 55.719, 57.078, 58.437, 59.796, 61.155, 62.967, 64.326, 66.138, 67.950, 69.762, 72.027, 74.292
- Já calculamos a covariância: 79.39
- Falta calcular s_x e s_y

Solução #3: correlação.

- Voltando ao exemplo do peso X altura:
- x: 147.32, 149.86, 152.40, 154.94, 157.48, 160.02, 162.56, 165.10, 167.64, 170.18, 172.72, 175.26, 177.80, 180.34, 182.88
- y: 52.095, 53.001, 54.360, 55.719, 57.078, 58.437, 59.796, 61.155, 62.967, 64.326, 66.138, 67.950, 69.762, 72.027, 74.292
- Já calculamos a covariância: 79.39
- Falta calcular s_x e s_y
- $s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

Solução #3: correlação.

- Voltando ao exemplo do peso X altura:
- x: 147.32, 149.86, 152.40, 154.94, 157.48, 160.02, 162.56, 165.10, 167.64, 170.18, 172.72, 175.26, 177.80, 180.34, 182.88
- y: 52.095, 53.001, 54.360, 55.719, 57.078, 58.437, 59.796, 61.155, 62.967, 64.326, 66.138, 67.950, 69.762, 72.027, 74.292
- Já calculamos a covariância: 79.39
- Falta calcular s_x e s_y
- $s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$
- $s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$

Solução #3: correlação.

- Voltando ao exemplo do peso X altura:
- x: 147.32, 149.86, 152.40, 154.94, 157.48, 160.02, 162.56, 165.10, 167.64, 170.18, 172.72, 175.26, 177.80, 180.34, 182.88
- y: 52.095, 53.001, 54.360, 55.719, 57.078, 58.437, 59.796, 61.155, 62.967, 64.326, 66.138, 67.950, 69.762, 72.027, 74.292
- Já calculamos a covariância: 79.39
- Falta calcular s_x e s_y
- $s_x = \sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n-1}}$
- $s_y = \sqrt{\dfrac{\sum (y_i - \bar{y})^2}{n-1}}$
- $r_{xy} = \dfrac{s_{xy}}{s_x s_y} = \dfrac{79.39}{(11.35)(7.02)} = 0.99$

Como assim “amostral”?

- Se os dados são da população e não de uma amostra, é só substituir $n-1$ por N nas fórmulas da covariância e do desvio-padrão.
- $\sigma_{xy} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N}$
- $\sigma_x = \sqrt{\dfrac{\sum (x_i - \bar{x})^2}{N}}$
- $\sigma_y = \sqrt{\dfrac{\sum (y_i - \bar{y})^2}{N}}$

Cuidado!

- Correlação != causação.
- Correlação pode não ser linear.