Machine Learning com Python

Prof. Luciano Galdino

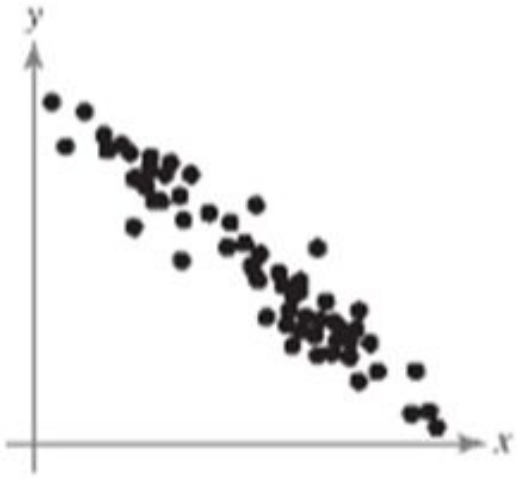
REGRESSÃO LINEAR SIMPLES

Modelo matemático linear capaz de realizar previsões.

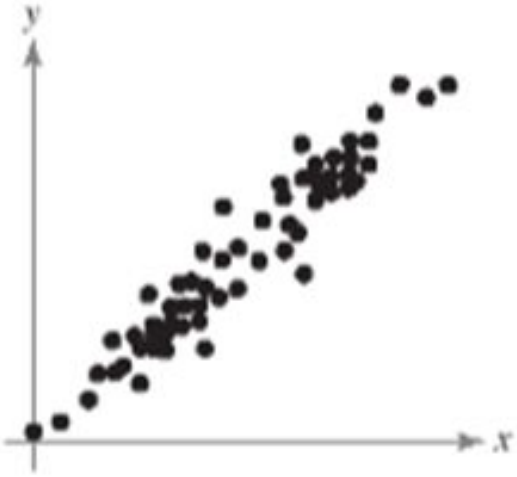
$$y = m.x + b$$

Correlação linear
Determinado através de gráficos de dispersão e do coeficiente de correlação.

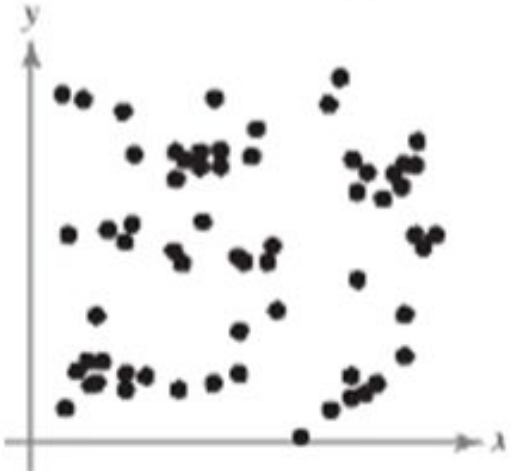
Correlação linear negativa



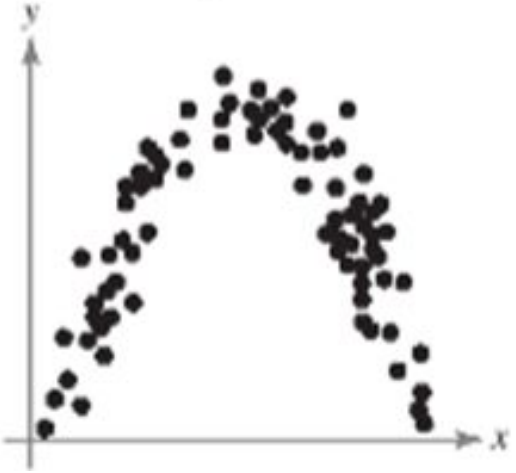
Correlação linear positiva



Não há correlação



Correlação não linear



Coeficiente de Correlação Linear (Coeficiente de Pearson)

- Forma mais precisa de medir a correlação entre duas grandezas.
- Teste Paramétrico (Normalidade).

$$r = \frac{\sum(x_i - \bar{x}).(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 . \sum(y_i - \bar{y})^2}}$$

Coeficiente de correlação (r)	Correlação Positiva	Coeficiente de correlação (r)	Correlação Negativa
r = 1	Perfeita	r = - 1	Perfeita
0,95 ≤ r < 1	Muito forte	- 0,95 ≤ r < -1	Muito forte
0,8 ≤ r < 0,95	Forte	-0,8 ≤ r < -0,95	Forte
0,5 ≤ r < 0,8	Moderada	-0,5 ≤ r < -0,8	Moderada
0 ≤ r < 0,5	Fraca	0 ≤ r < -0,5	Fraca

Teste de Hipótese para o coeficiente de Correlação

Teste t

$$t = \frac{r}{\sigma_r} = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

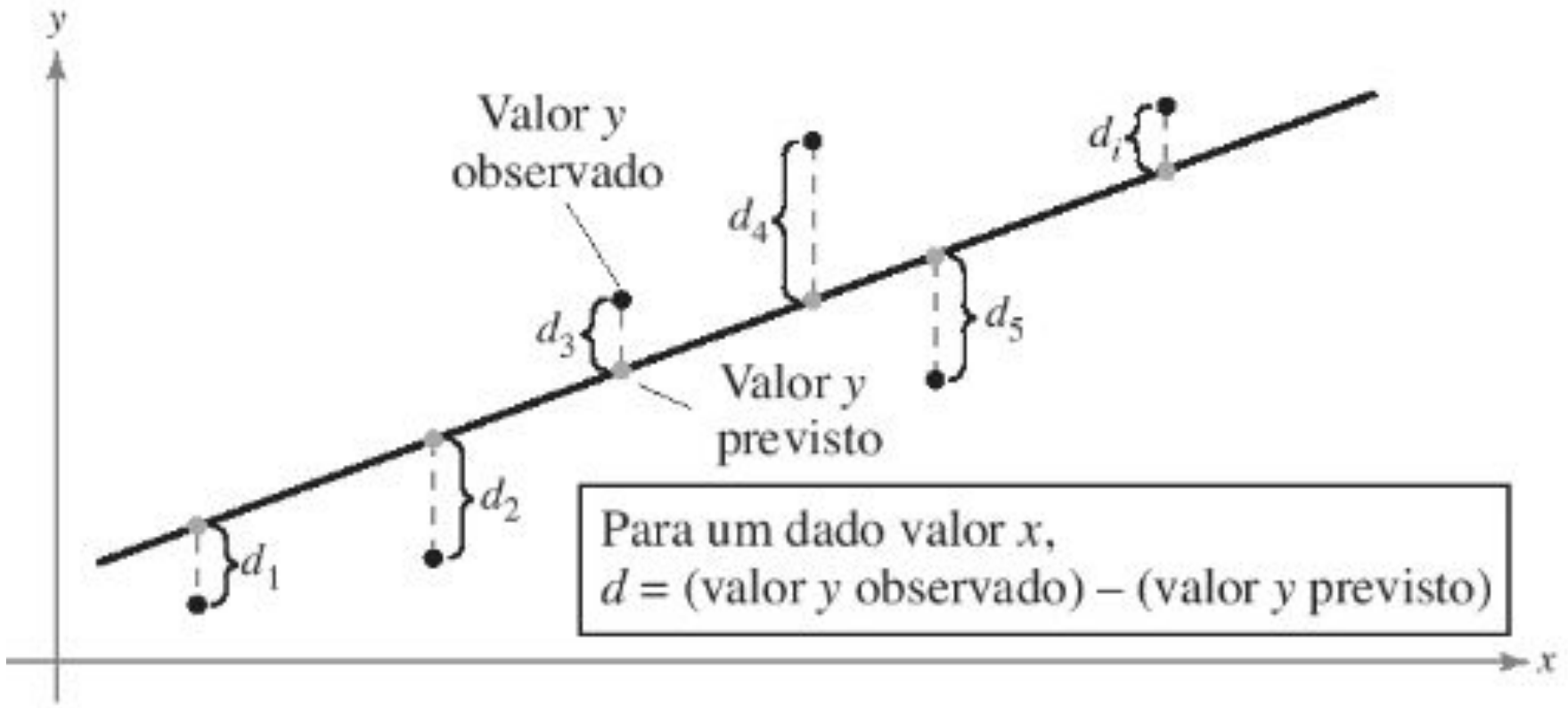
Graus de liberdade

$$gl = n - 2$$

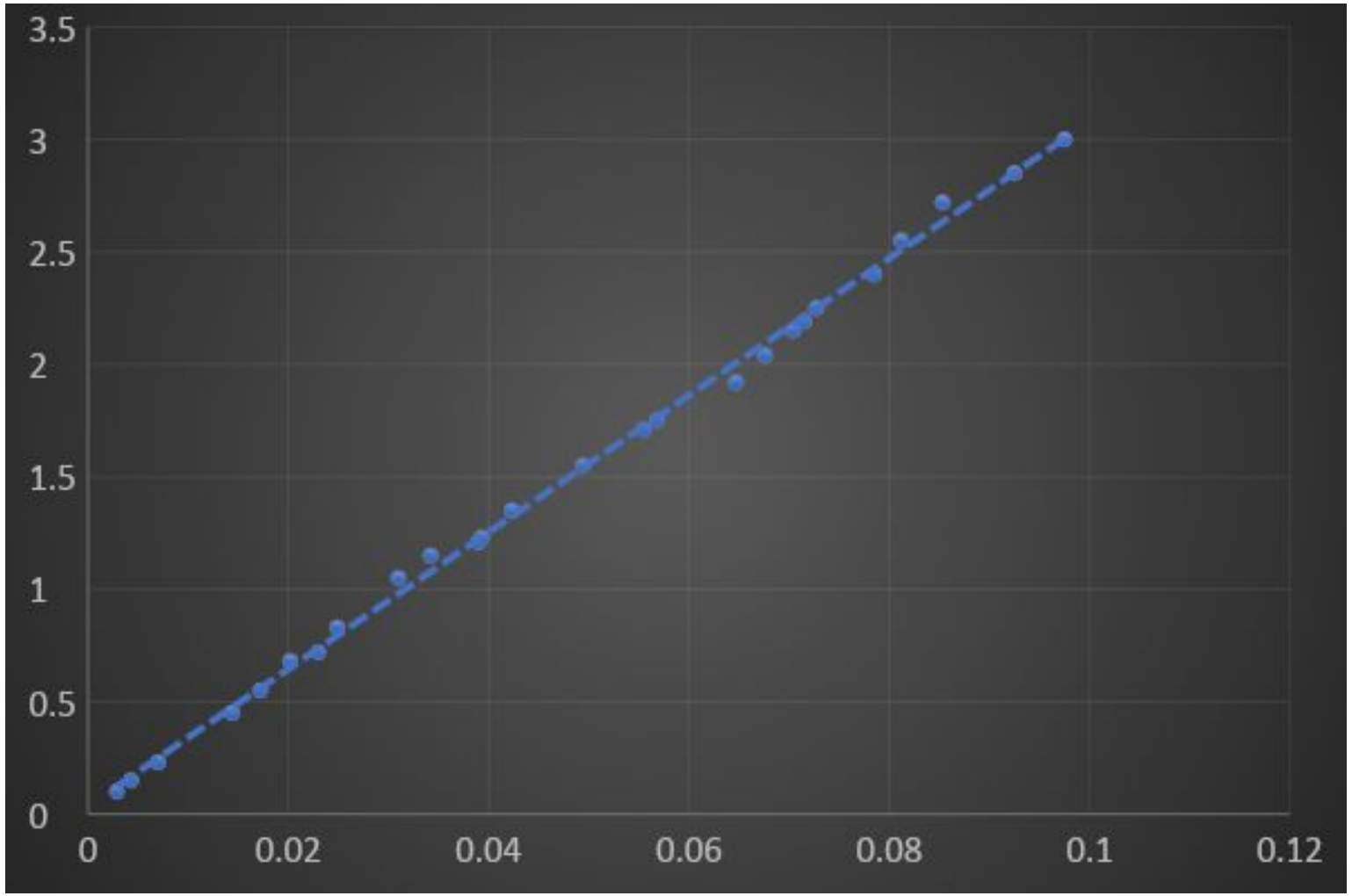
	Nível de confiança, c	0,50	0,80	0,90	0,95	0,98	0,99
	Unicaudal, α	0,25	0,10	0,05	0,025	0,01	0,005
g.l.	Bicaudal, α	0,50	0,20	0,10	0,05	0,02	0,01
1		1,000	3,078	6,314	12,706	31,821	63,657
2		0,816	1,886	2,920	4,303	6,965	9,925
3		0,765	1,638	2,353	3,182	4,541	5,841
4		0,741	1,533	2,132	2,776	3,747	4,604
5		0,727	1,476	2,015	2,571	3,365	4,032
6		0,718	1,440	1,943	2,447	3,143	3,707
7		0,711	1,415	1,895	2,365	2,998	3,499
8		0,706	1,397	1,860	2,306	2,896	3,355
9		0,703	1,383	1,833	2,262	2,821	3,250
10		0,700	1,372	1,812	2,228	2,764	3,169
11		0,697	1,363	1,796	2,201	2,718	3,106
12		0,695	1,356	1,782	2,179	2,681	3,055
13		0,694	1,350	1,771	2,160	2,650	3,012
14		0,692	1,345	1,761	2,145	2,624	2,977
15		0,691	1,341	1,753	2,131	2,602	2,947
16		0,690	1,337	1,746	2,120	2,583	2,921
17		0,689	1,333	1,740	2,110	2,567	2,898
18		0,688	1,330	1,734	2,101	2,552	2,878
19		0,688	1,328	1,729	2,093	2,539	2,861
20		0,687	1,325	1,725	2,086	2,528	2,845
21		0,686	1,323	1,721	2,080	2,518	2,831
22		0,686	1,321	1,717	2,074	2,508	2,819
23		0,685	1,319	1,714	2,069	2,500	2,807
24		0,685	1,318	1,711	2,064	2,492	2,797
25		0,684	1,316	1,708	2,060	2,485	2,787

Linhas de regressão

Linha que melhor se ajusta aos dados plotados, onde a soma dos quadrados dos resíduos seja mínima.



Modelo Matemático (Equação da reta)



$$y = m.x + b$$

Coeficientes

$$m = \frac{\sum(x_i - \bar{x}).(y_i - \bar{y})}{\sum(y_i - \bar{y})^2}$$

$$b = \bar{y} - m\bar{x}$$

Coeficiente de determinação

Porcentagem da variação de y que pode ser explicada pela relação de x e y.

$$r^2 = \frac{Variação\ encontrada}{Variação\ total}$$

$$r^2 = \rho^2$$

Coeficiente de correlação de postos de Spearman

Teste Não paramétrico.

Medida da força da relação entre duas variáveis. Utiliza os postos de entradas de amostras de dados pareados.

Pode ser utilizado na relação de dados lineares e também não lineares, assim como também para dados no nível ordinal.

Cálculo do Coeficiente de Spearman

$$r_R = 1 - \frac{6\Sigma_i {d_i}^2}{n(n^2-1)}$$

n = número amostras.

di = diferença de alcance de cada elemento.

Coeficiente de correlação (r_R)	Correlação Positiva	Coeficiente de correlação (r_R)	Correlação Negativa
$r_R = 1$	Perfeita	$r_R = -1$	Perfeita
$0{,}95 \leq r_R < 1$	Muito forte	$-0{,}95 \leq r_R < -1$	Muito forte
$0{,}8 \leq r_R < 0{,}95$	Forte	$-0{,}8 \leq r_R < -0{,}95$	Forte
$0{,}5 \leq r_R < 0{,}8$	Moderada	$-0{,}5 \leq r_R < -0{,}8$	Moderada
$0 \leq r_R < 0{,}5$	Fraca	$0 \leq r_R < -0{,}5$	Fraca

Coeficiente de correlação de Kendall

Teste não paramétrico indicado para número pequeno de amostras.

Ou para populações com grandes quantidades de empates (valores repetidos).

Pode ser utilizado juntamente com o Spearman para comparação.

É mais conservador que o teste de Spearman.

Cálculo do Coeficiente de Kendall

$$\tau = \frac{\overbrace{x_i > x_j \text{ e } y_i > y_j \text{ ou se } x_i < x_j \text{ e } y_i < y_j}^{\text{(quantidade de pares concordantes)}} - \overbrace{x_i > x_j \text{ e } y_i < y_j \text{ ou se } x_i < x_j \text{ e } y_i > y_j}^{\text{(quantidade de pares discordantes)}}}{n(n-1)/2}$$

Coeficiente de correlação (T)	Correlação Positiva	Coeficiente de correlação (T)	Correlação Negativa
T = 1	Perfeita	**T** = - 1	Perfeita
0,95 ≤ **T** < 1	Muito forte	- 0,95 ≤ **T** < -1	Muito forte
0,8 ≤ **T** < 0,95	Forte	-0,8 ≤ **T** < -0,95	Forte
0,5 ≤ **T** < 0,8	Moderada	-0,5 ≤ **T** < -0,8	Moderada
0 ≤ **T** < 0,5	Fraca	0 ≤ **T** < -0,5	Fraca