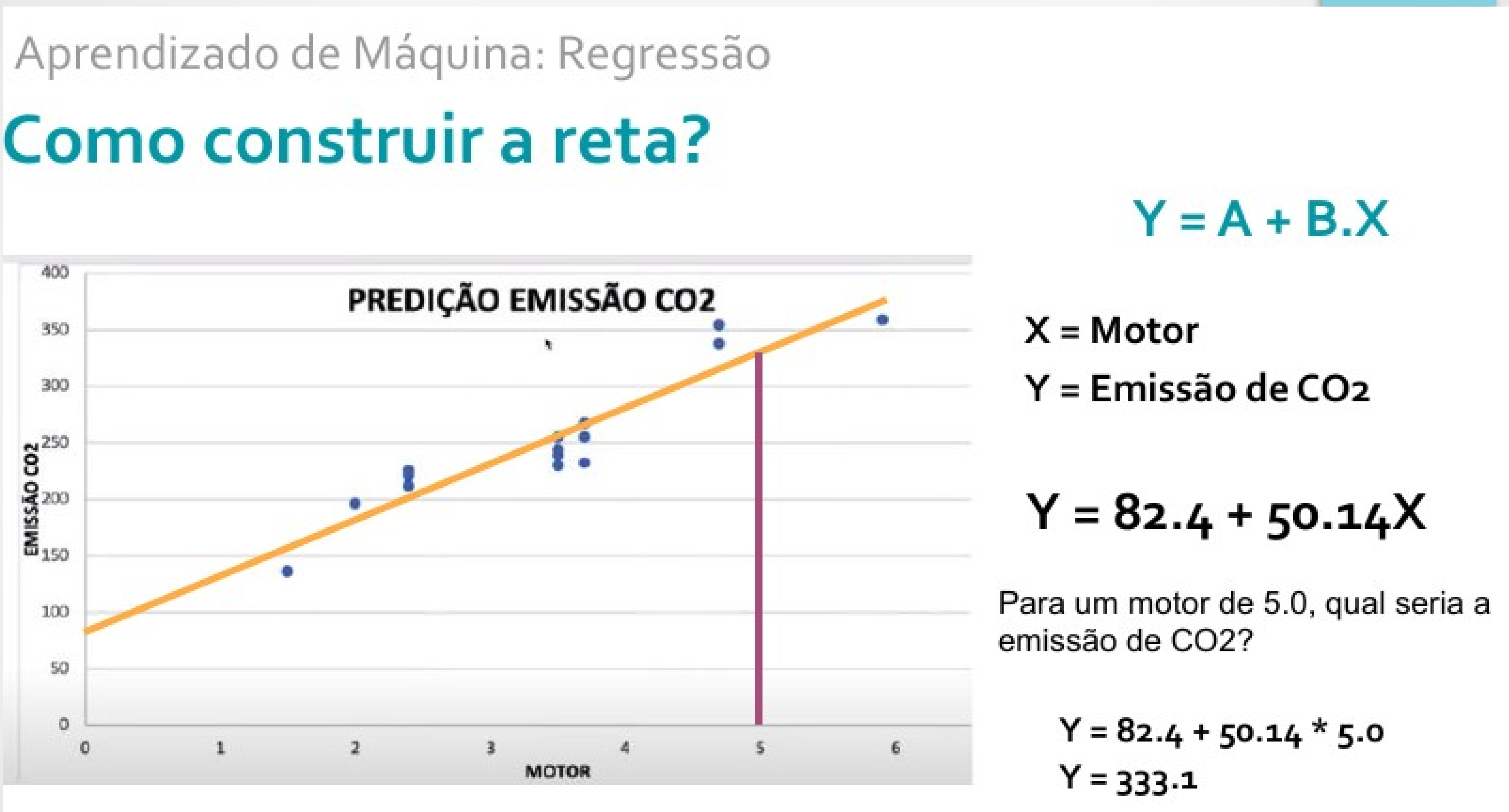
Método dos Mínimos Quadrados

Prof. Marcelo Dib Cruz

Aprendizado de Máquina

- Aprendizado Supervisionado
  - Classificação
    - Árvore de Decisão
  - Regressão
    - **Regressão Linear**
  - Redes Neurais
- Aprendizado nâo supervisionado
  - RNA não supervisionada
  - Clustering ou Agrupamento
    - K-means
    - Automatic Clustering
      - Algoritmo Genético
- Aprendizado por Reforço

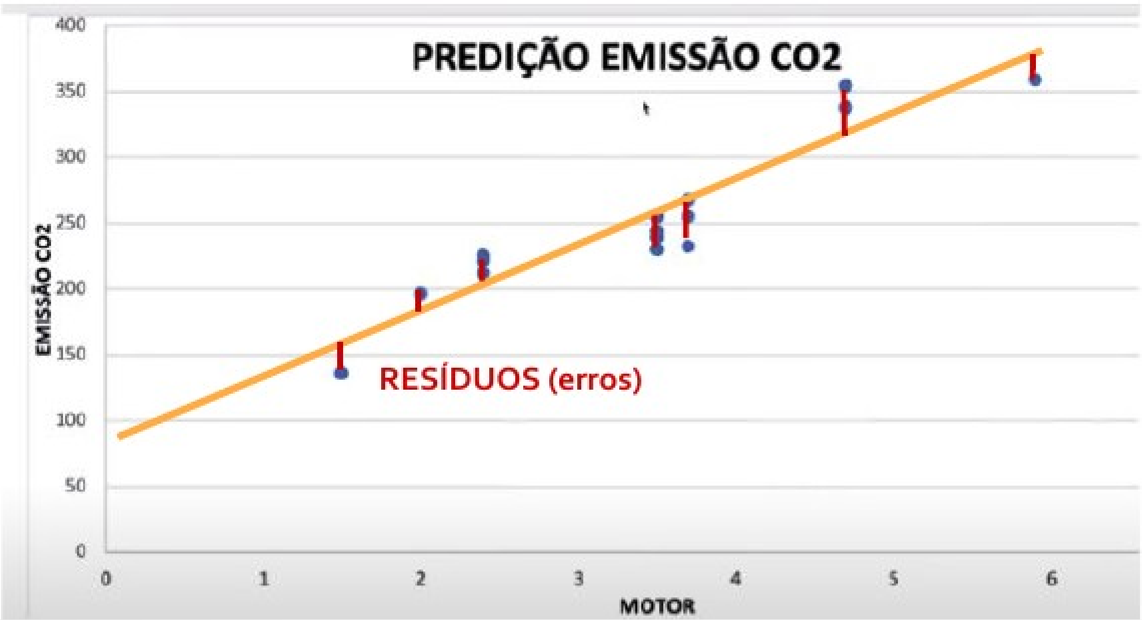
Aprendizado de Máquina



Aprendizado de Máquina

-

Método dos Mínimos Quadrados (MMQ)



Ajusta o modelo de modo que a soma dos quadrados das diferenças dos valores observados e previstos seja **minimizada**

$$min \sum (y_i - \hat{y}_i)^2$$

Aprendizado de Máquina

-

$$q = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{2.1}$$

Os candidatos a ponto de mínimo da função 2.1 são aqueles para os quais são nulos as derivadas parciais de q em relação a cada um de seus parâmetros, isto é:

$$\frac{\partial q}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \tag{2.2}$$

$$\frac{\partial q}{\partial b} = -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0 \tag{2.3}$$

Tendo em vista que:

$$\begin{aligned} \sum_{i=1}^{n} (y_i - a - bx_i) &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - \sum_{i=1}^{n} bx_i \\ &= \sum_{i=1}^{n} y_i - na - \left(\sum_{i=1}^{n} x_i \right) b \end{aligned}$$

e que:

$$\sum_{i=1}^{n} x_i (y_i - a - bx_i) = \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i \right) a - \left(\sum_{i=1}^{n} x_i^2 \right) b$$

obtemos o seguinte sistema de equações, denominado "equações normais" do problema, cujas incógnitas são os parâmetros a e b da equação $y = a + bx$:

Aprendizado de Máquina

-

$$
\begin{cases}
na + \left(\sum_{i=1}^{n} x_i\right) b = \sum_{i=1}^{n} y_i \\
\left(\sum_{i=1}^{n} x_i\right) a + \left(\sum_{i=1}^{n} x_i^2\right) b = \sum_{i=1}^{n} x_i y_i
\end{cases}
\tag{2.4}
$$

Exemplo 1:

Dada a tabela de pontos (x_i, y_i) a seguir, determine pelo Método dos Quadrados Mínimos a equação da reta que melhor se ajusta a esses pontos.

x_i	-1.0	-0.1	0.2	1.0
y_i	1.000	1.099	0.808	1.000

Solução:

Como são $n = 4$ pontos, $\sum_{i=1}^{n} x_i = 0.1$, $\sum_{i=1}^{n} x_i^2 = 2.05$, $\sum_{i=1}^{n} y_i = 3.907$ e $\sum_{i=1}^{n} x_i y_i = 0.0517$, as equações normais do problema são, de acordo com 2.4:

$$
\begin{cases}
4a + 0.10b = 3.9070 \\
0.1a + 2.05b = 0.0517
\end{cases}
$$

A solução deste sistema é $a = 0.9773$ e $b = -0.0224$. Assim, a reta que melhor se ajusta à tabela de pontos dada é:

$$y = 0.9773 - 0.0224x$$

Aprendizado de Máquina

- **Qualidade do Ajuste**
- R2 mede a habilidade do modelo de explicar a variação no resultado, que é um indicador de quão preditivo é o modelo.
  - 0 significa que o modelo não pode explicar nenhuma variabilidade no resultado.
  - 1 significa que o modelo explica todas as variabilidades.
- MAE - Erro absoluto médio. Mede a diferença absoluta entre o valor real e a previsão. Todas as diferenças são ponderadas igualmente nessa média, o que significa que não é tão sensível a desvios quanto MSE.
- MSE (erro quadrático médio, que é média dos quadrados dos erros das previsões do modelo)
- RMSE - Representa a raiz quadrada do MSE .

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Aprendizado de Máquina

- Qualidade do Ajuste

A qualidade de um ajuste linear pode ser verificada em função do coeficiente de determinação r^2, dado por:

$$r^2 = \frac{\sum_{i=1}^{n}(a + bx_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8.13}$$

sendo $\bar{y} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)$. Quanto mais próximo da unidade r^2 estiver, melhor é o ajuste.

Aprendizado de Máquina

- Qualidade do Ajuste

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-a-bx_i)^2+\sum_{i=1}^{n}(a+bx_i-\bar{y})^2$$

a expressão 8.13 pode ser reescrita como:

$$r^2=\frac{\sum\limits_{i=1}^{n}(y_i-\bar{y})^2-\sum\limits_{i=1}^{n}(y_i-a-bx_i)^2}{\sum\limits_{i=1}^{n}(y_i-\bar{y})^2}$$

Como:

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}y_i^2-2\bar{y}\sum_{i=1}^{n}y_i+n\sum_{i=1}^{n}\bar{y}^2\\ &= \sum_{i=1}^{n}y_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2\end{aligned}$$

a expressão para determinação do coeficiente de determinação r^2 pode ser simplificada para:

$$r^2=1-\frac{\sum\limits_{i=1}^{n}(y_i-a-bx_i)^2}{\sum\limits_{i=1}^{n}y_i^2-\frac{1}{n}\left(\sum\limits_{i=1}^{n}y_i\right)^2} \qquad (8.14)$$

Aprendizado de Máquina

- Qualidade do Ajuste

Árvore de Decisão

- Referencias
- Machine Learning. Tom Mitchell. McGraw-Hill.1997.
- https://www.cin.ufpe.br/~if684/EC/aulas/Aula-arvores-decisao-SI.pdf

Foreman, John W. Data Smart: Using Data Science to Transform Information into Insight. Wiley. 1ª Ed., 2013.

- Curso IA para todos do prof. Diogo Cortiz da PUC-SP;

https://www.youtube.com/watch?v=Ze-Q6ZNWpco&list=PLtQM10PgmGogjn0cikgWi8wpQUnV6ERkY

- Livros gratuitos de Ciência de Dados-

https://www.dataquest.io/blog/data-science-books/

- Materiais do prof. Regis da UFC

https://sites.google.com/site/ufcregis/home/data-science-machine-learning