



Métricas de avaliação II: MSE

≡ Ciclo	Ciclo 03: Aprendizado supervisionado - Regressão
# Aula	23
🕒 Created	@January 30, 2023 7:58 AM
☑ Done	☑
☑ Ready	☑

Objetivo da Aula:

- ☐ A reta de regressão
- ☐ O erro MSE
- ☐ Os problemas do MSE
- ☐ MSE na prática
- ☐ Resumo
- ☐ Próxima aula

Conteúdo:

▼ 1. A reta de regressão

▼ 2. O erro MSE

MSE (Mean Square Error) calcula a média do quadrado das diferenças entre os valores reais e preditos.

O MSE é usado como um função de perda (métrica de performance) e representa quão bem o model ajustou aos dados de treinamento. Quanto menor o MSE, melhor o ajuste do modelo aos dados.

▼ 2.1 Fórmula

O erro médio quadrático, conhecido como MSE, é calculado da seguinte forma:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

▼ 2.2 Exemplo:

Actual Value (y)	Predicted Value (\hat{y})	Error (y - \hat{y})	Squared Error (Error ²)
10	18	-8	64
120	10	110	12100
15	13	2	4
18	25	-7	49

20	19	1	1
25	26	-1	1
28	27	1	1
30	29	1	1
33	31	2	4
35	38	-3	9
38	37	1	1
40	39	1	1
43	42	1	1
45	44	1	1
			142
			$MSE = (12100 / 14) = 10.14$
			$= 864,00$

▼ 2.3 Como interpretar o MSE

O MSE é um único número que nos diz o quão próximo as previsões estão dos valores reais, na média.

▼ 3. Os problemas do MSE

1. **Sensível na presença de outliers:** Um único grande erro pode impactar significativamente o valor final do MSE.
2. **Fora de escala:** O MSE é o erro quadrático, que não está na mesma escala da variável resposta. Por exemplo, o MSE da previsão de metro é metro ao quadrado. Não é possível comparar se o erro está bom ou ruim em relação a escala original.

▼ 4. MSE na prática

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn import metrics as mt

# 1.0 Load dataset
df = pd.read_csv( '../dataset/train.csv' )

# 2.0 Seleção de Features
features = ['idade', 'divida_atual', 'renda_anual', 'valor_em_investimentos',
            'taxa_utilizacao_credito', 'num_emprestimos', 'num_contas_bancarias', 'num_cartoes_credito',
            'dias_atraso_dt_venc', 'num_pgto_atrasados', 'num_consultas_credito', 'taxa_juros']

label = ['saldo_atual']

x_train = df.loc[:, features]
y_train = df.loc[:, label]

# 3.0 Model Training
lr_model = LinearRegression()
lr_model.fit( x_train, y_train )

y_pred = lr_model.predict( x_train )

df1 = df.loc[:, ['id_cliente', 'saldo_atual']]
df1['predicted'] = y_pred

# 4.0 Model Performance
## 4.1 R squared
r2_squared = np.round( 100*mt.r2_score( y_train, y_pred ), 2 )
print( 'R2 square: {}'.format( r2_squared ) )
```

```
## 4.2 MSE
mse = np.round( mt.mean_squared_error( y_train, y_pred ) , 2 )
print( 'A cada previsão, o erro médio é de: U${}'.format( mse ) )

# 5.0 Conclusão
print( '{}% da variação da variável alvo y é reduzida, levando em consideração o preditor'.format( r2_squared ) )
print( '{}% da variação da variável alvo y é "explicada pela variação do preditor x'.format( r2_squared ) )
```

▼ 5. Resumo

1. O erro MSE mede o erro médio ao realizar cada previsão.
2. O principal problema do MSE é a alta importância aos outliers.

▼ 6. Próxima aula

Métricas de avaliação III: RMSE