INF0615 – Aprendizado de Máquina Supervisionado

Trabalho 1 - Regressão Linear

Nicole Nogueira Silva Rodolfo Dalla Costa

Introdução

O Monóxido de carbono (CO) é um gás incolor e inflamável produzido com base na queima incompleta de material combustível rico em carbono. Apesar de suas aplicações na indústria, é um gás asfixiante muito tóxico para os seres humanos. Nesse contexto, avaliar a concentração de CO é fundamental para mensurar a qualidade do ar de uma determinada região. Dessa forma, o objetivo desse trabalho é desenvolver modelos de regressão linear para predizer a concentração de CO no ar usando um conjunto de dados coletados com diversas informações sobre a característica do ar.

Banco de dados

Tabela 1: Estatísticas sumárias do banco de dados

year	month	day	hour	PM2.5	PM10	SO2	NO2
Min. :2013	Min. : 1.00	Min. : 1.0	Min. : 0.0	Min. : 2	Min. : 2	Min. : 0	Min. : 2.0
1st Qu.:2014	1st Qu.: 3.00	1st Qu.: 8.0	1st Qu.: 6.0	1st Qu.: 20	1st Qu.: 36	1st Qu.: 2	1st Qu.: 23.0
Median :2015	Median : 6.00	Median :16.0	Median :12.0	Median : 55	Median : 82	Median : 7	Median : 43.0
Mean :2015	Mean : 6.47	Mean :15.7	Mean :11.6	Mean : 79	Mean :105	Mean : 16	Mean : 50.6
3rd Qu.:2016	3rd Qu.: 9.00	3rd Qu.:23.0	3rd Qu.:18.0	3rd Qu.:111	3rd Qu.:145	3rd Qu.: 19	3rd Qu.: 71.0
Max. :2017	Max. :12.00	Max. :31.0	Max. :23.0	Max. :844	Max. :999	Max. :500	Max. :290.0

O3	TEMP	PRES	DEWP	RAIN	wd	WSPM	target
Min. : 0	Min. :-19.9	Min. : 983	Min. :-36.0	Min. : 0.0	NE : 25096	Min. : 0.00	Min. : 100
1st Qu.: 10	1st Qu.: 3.1	1st Qu.:1002	1st Qu.: -9.0	1st Qu.: 0.0	ENE : 20029	1st Qu.: 0.90	1st Qu.: 500
Median : 45	Median : 14.4	Median :1010	Median : 2.9	Median : 0.0	NW : 19211	Median : 1.40	Median : 900
Mean : 57	Mean : 13.5	Mean :1011	Mean : 2.4	Mean : 0.1	N : 17987	Mean : 1.74	Mean : 1229
3rd Qu.: 82	3rd Qu.: 23.2	3rd Qu.:1019	3rd Qu.: 15.0	3rd Qu.: 0.0	E : 17371	3rd Qu.: 2.20	3rd Qu.: 1500
Max. :1071	Max. : 41.6	Max. :1043	Max. : 29.1	Max. :52.1	SW : 16929	Max. :12.90	Max. :10000
NA	NA	NA	NA	NA	(Other):127959	NA	NA

O banco de dados foi dividido em 3 blocos, um conjunto de treinamento que será utilizado para treinar os modelos, um conjunto de validação para mensurar o desempenho dos modelos e um conjunto de teste. A base de treino possui 244582 linhas e 17 colunas enquanto a base de validação possui 61147 linhas e as mesmas colunas do conjunto de treino. Nota-se que o banco de dados não possui dados faltantes, se houvesse, o ideal é avaliar se a informação faltante é erro de preenchimento no momento da coleta ou se realmente traz uma informação relevante.

A partir do summary apresentado na Tabela 1 é possível notar que a variável "wd", que indica a direção do vento no momento da coleta, é a única variável categorizada da base. Para lidar com isso, tranformamos a variável utilizando a técnica de One-Hot-encoding.

Análise descritiva

Para entender as relações entre as variáveis e a concentração de monóxido de carbono (nosso Target), vamos analisar a distribuição da correlação entre as colunas da base. A Figura 1 apresenta um panorama geral da correlação com todas as features. Nota-se que a variável resposta tem correlação considerável com as variáveis PM25, PM10 e NO2.

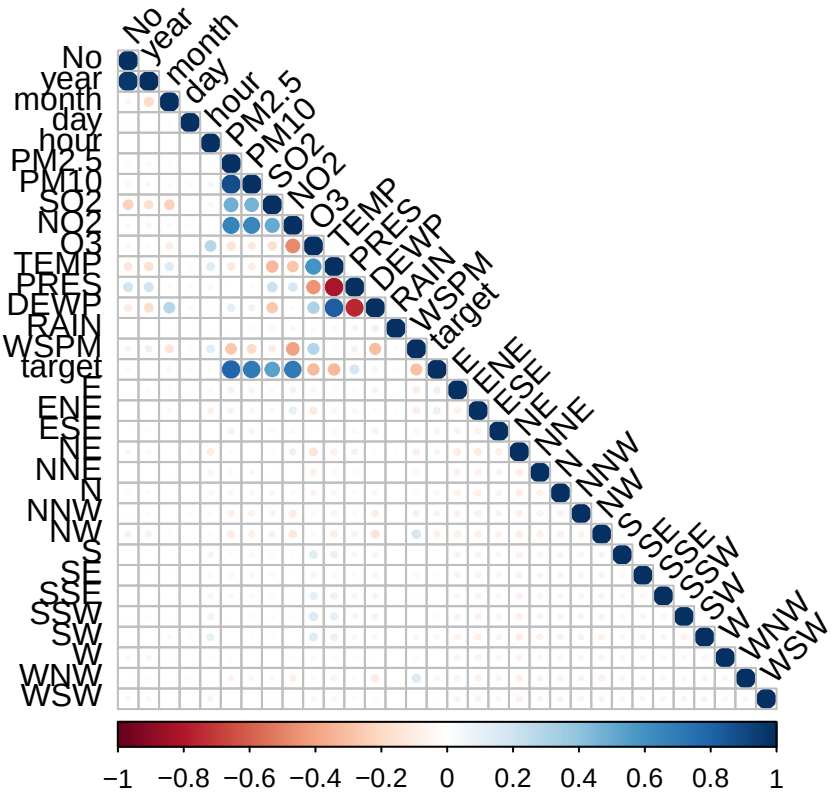

Figura 1: Correlações 2 a 2 das variáveis.

O gráfico da Figura 2 apresenta a distribuição da concentração de CO em relação a concentração de PM2.5 no ar. É possível observar que existe uma relação linear entre as variáveis já que a concentração de CO aumenta conforme o PM2.5 aumenta também, o que indica que ajustar um modelo de regrssão linear pode ser uma estratégia adequada.

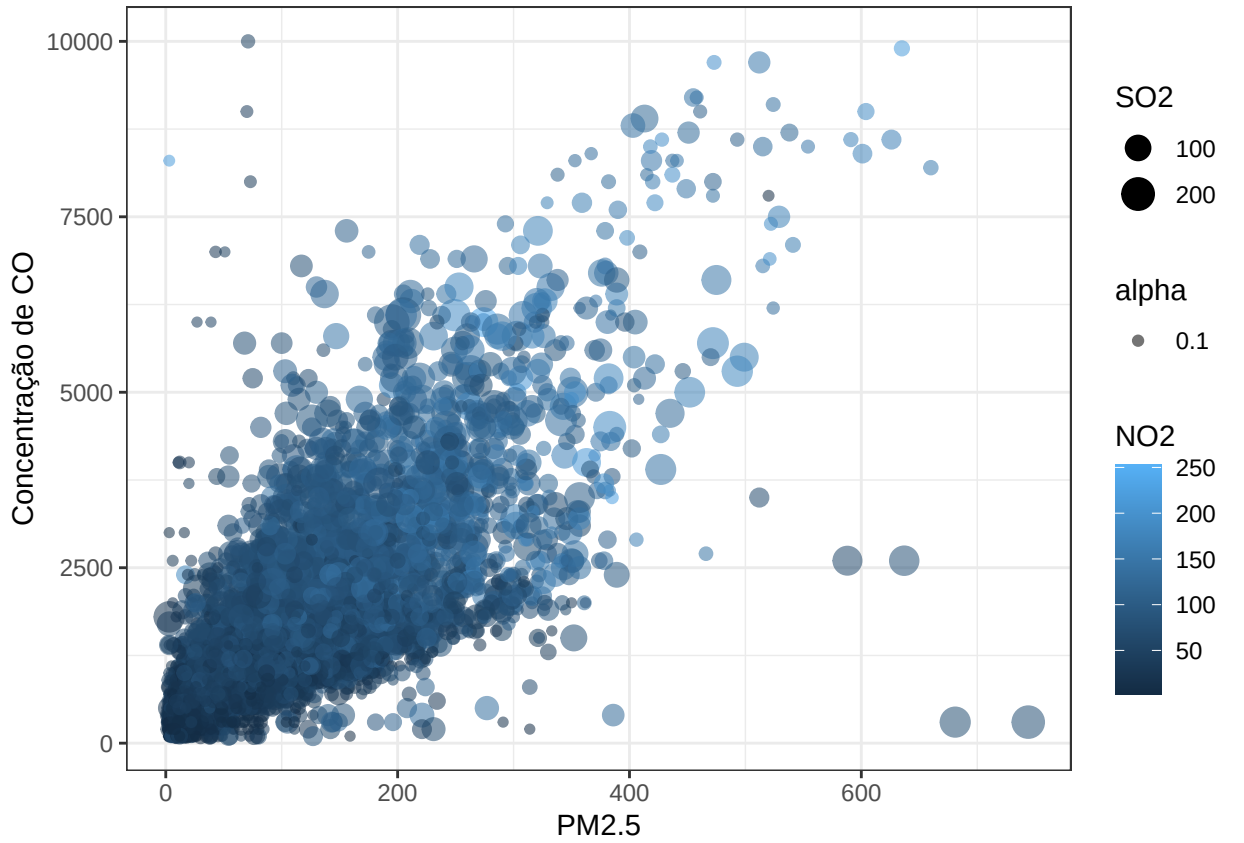

Figura 2: Distribuição da concentração de CO e de PM2.5 variando o NO2 e SO2.

Além disso, como é possível notar por meio da Tabela 1, as colunas estão em escalas diferentes, por isso, aplicamos a normalização Z-norma nos dados de treino e os mesmos parâmetros para a validação. Dessa forma, cada feature está padronizada com média zero e desvio padrão correspondente.

Metodologia

Para predizer o valor da concentração do monóxido de carbono, primeiro é importante avaliar um modelo de regressão linear mais simples com todas as variáveis, esse será o baseline. A Tabela 2 apresenta as medidas do erro médio (MAE), erro quadrático médio (MSE) e o R2 do baseline para os dados de treino, validação e teste. O baseline apresentou um MAE de 372.4 no conjunto de teste.

Tabela 2: Resultado do baseline nos conjuntos de treino, validação e teste

	MAE	MSE	R2
Treino	370.09	354861.52	0.73
Validação	371.07	366334.64	0.73
Teste	372.37	362478.31	0.73

Para melhorar a predição do CO, vamos criar um modelo com combinação de features. Para esse modelo, vamos utlizar todas as variáveis e combinar 2 a 2 as variáveis PM2.5, PM10, SO2, NO2, O3, TEMP, PRES e DEWP. Os resultados do MAE, MSE e R2 para o conjunto de validação podem ser observados na Tabela 3. A primeira linha corresponde aos resultados do modelo com combinação 2 a 2 enquanto a segunda linha é

Tabela 3: Resultado das combinações nos conjuntos de treino, validação e teste

Modelo	MAE	MSE	R2
2 a 2	307.03	271113.02	0.8
3 a 3	291.87	253683.35	0.81

o modelo utilizando a combinação 3 a 3. É possível observar que o o melhor modelo desta categoria foi o modelo com combinação 3 a 3. Aplicando esse modelo no conjunto de teste obtivemos um MAE de 292.9.

Por fim, vamos testar o modelo de regressão aumentando o grau das features. O gráfico da Figura 3 apresenta o MAE para cada polinômio no conjunto de treino e validação. Podemos reparar que até o grau 3 estamos com underfitting pois o erro no treino e na validação são altos e muito similares. Porém, a partir do grau 6 temos overfitting já que o erro médio no conjunto de validação começa crescer enquanto no conjunto de treino cai. Dessa forma, o ponto ótimo do modelo dessa categoria é o modelo de grau 5. Ajustando o nosso melhor modelo polinomial de grau 5 no conjunto de teste obtivemos um MAE de 344.16.

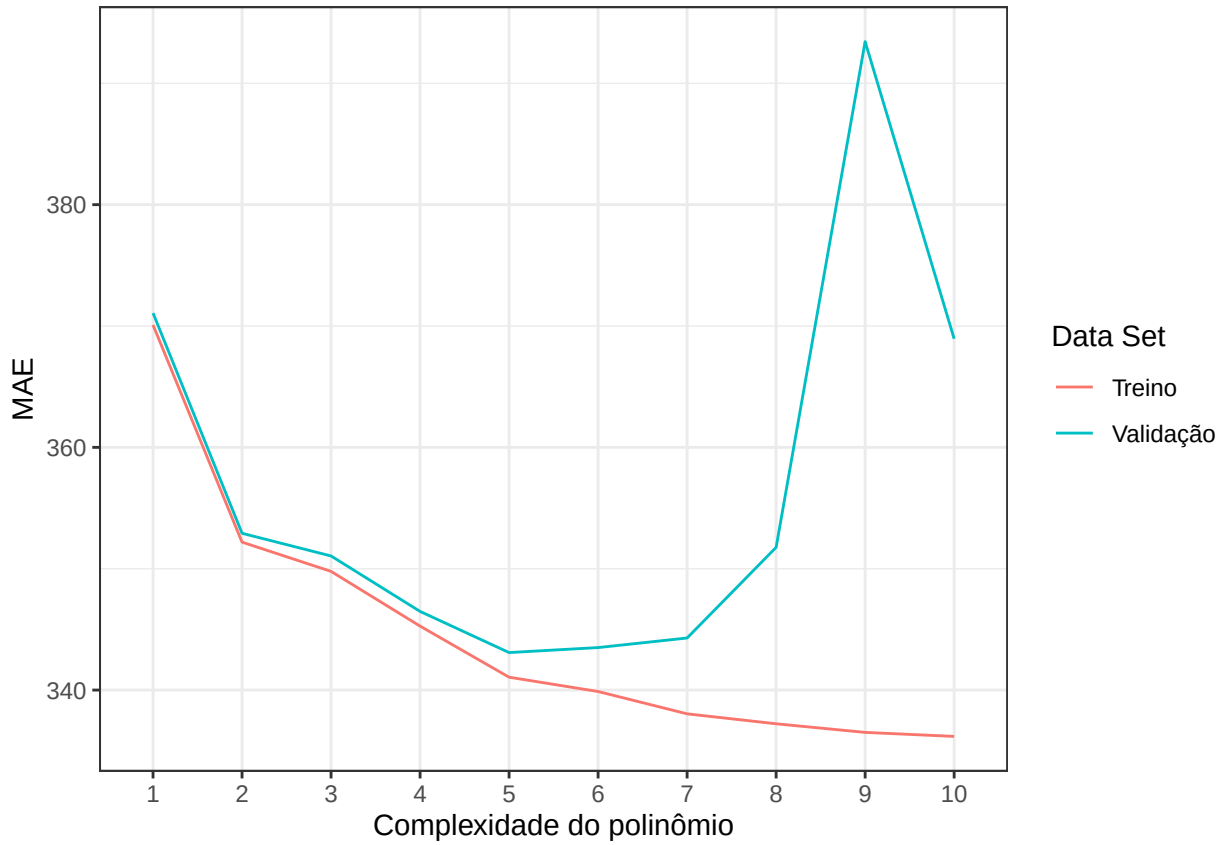

Figura 3: MAE do treino e validação para cada grau.

Conclusão

O modelo que apresentou menor erro médio foi o modelo utilizando todas as variáveis com grau 1 e a combinação 3 a 3 das variáveis que possuem maior correlação com a concentração de monóxido de carbono. O baseline, modelo mais simples, apresentou o pior desempenho quando comparado com os outros. O modelo polinomial apresentou resultados razoáveis porém possui uma complexidade alta e exige muito processamento

para treinar. Assim, o modelo com combinação 3 a 3 tem vantagem além de possuir o menor MAE. É interessante observar que, selecionar melhor quais variáveis são mais relevantes para predizer a concentração de CO trouxe ganhos significativos no erro médio.