INF-0615– Aprendizado de Máquina Supervisionado I

Trabalho 01

Predição da Qualidade do Ar pela Concentração de Monóxido de Carbono

Data de Entrega: 05/09/2021

1 Descrição do Dataset

O Monóxido de Carbono (CO) é um gás extramente tóxico para os seres humanos e animais e pode causa sérios problemas respiratórios e circulatórios se inalado em grandes quantidades. Ele é emitido por automóveis, fábricas, fornalhas, equipamentos de aquecimento por meio da queima incompleta de combustíveis fósseis. Nesse contexto, a concentração de CO pode ser um excelente indicativo para mensurar a qualidade do ar, a qual pode ser utilizada para averiguar e diagnosticar os riscos que uma população está propensa em uma determinada região.

O objetivo desse trabalho é desenvolver modelos de regressão linear para predizer a concentração de Monóxido de Carbono no ar. Para isso, um cojunto de atributos em relação às características do ar também são coletados e descritos abaixo:

- **No**: Índice do exemplo na base de dados. É um atributo não preditivo e **não** deve ser utilizado durante o treinamento dos modelos.
- **Year**: Ano no qual a mensuração foi realizada.
- **Month**: Mês no qual a mensuração foi realizada.
- **Day**: Dia no qual a mensuração foi realizada.
- **Hour**: Hora em que a mensuração foi realizada.
- **PM2.5**: Concentração de PM2.5 no ar (em ug/m^3).
- **PM10**: Concentração de PM10 no ar (em ug/m^3).
- **SO2**: Concentração de SO_2 no ar (em ug/m^3).
- **NO2**: Concentração de NO_2 no ar (em ug/m^3).
- **O3**: Concentração de O_3 no ar (em ug/m^3).
- **TEMP**: Temperatura em graus Celsius no momento da coleta.
- **PRES**: Pressão Atmosférica (em hPa) no momento da coleta.
- **DEWP**: Ponto de Condensação da Água na região da coleta.
- **RAIN**: Precipitação da água na região da coleta (em mm).
- **WD**: Direção do vento no momento da coleta.
- **WSPM**: Velocidade do Vento.
- **Target**: Concentração de Monóxido de Carbono (CO) no ar em ug/m^3. **Esse é o valor alvo que vocês devem predizer**.

2 Tarefas

Pedimos que vocês:

1. Inspecionem os dados. Quantos exemplos vocês tem? Como vocês irão lidar com as features (atributos) discretas, se houverem? Há exemplos com features sem anotações? Como vocês lidariam com isso?
2. Apliquem alguma técnica de normalização de forma a deixa os dados mais bem preparados para o treinamento (Min-Max, Z-Norma, etc).
3. Como *baseline*, treinem uma regressão linear utilizando todas as features para predizer a concentração de Monóxido de Carbono no ar. Reportem o erro nos conjuntos de treinamento, validação e teste.
4. Implementem soluções alternativas baseadas em regressão linear através da combinação das features existentes para melhorar o resultado do *baseline*. Comparem suas soluções reportando os erros no conjunto validação. Tomem **apenas a melhor solução baseada no conjunto de validação** e reportem o erro no conjunto de teste.
5. Implementem soluções alternativas baseadas em regressão linear aumentando os graus das features (regressão com polinômios) para melhorar o resultado obtido no *baseline*. Plotem o erro no conjunto de treinamento e validação pelo grau do polinômio. Identifiquem as regiões de *underfitting*, ponto ótimo e *overfitting*. Tomem **apenas o melhor modelo polinomial baseado no conjunto de validação** e reportem seu erro no conjunto de teste.
6. Escrevam um relatório de no máximo 5 páginas:
   (a) Descrevam o que foi feito, bem como as diferenças entre o seu melhor modelo e o seu baseline;
   (b) Reportem o erro do melhor modelo de todos no conjunto de teste. Lembrem-se que o melhor modelo de todos deve ser escolhido baseado no erro no conjunto de validação.
   (c) Uma Seção de conclusão do relatório explicando a diferença entre os modelos e o porquê que estas diferenças levaram a resultados piores ou melhores.

3 Arquivos

Os arquivos disponíveis no Moodle são:

- *trabalho01_nome_dos_membros.R*: Código de apoio ao Trabalho 01. Desenvolvam o trabalho a partir dele.
- *training_set_air_quality*: conjunto de dados para treinamento;
- *validation_set_air_quality*: conjunto de dados para validação;
- *test_set_price_variation*: conjunto de dados de teste retido pelo professor (**será disponibilizado no sábado anterior ao prazo final da submissão**).

4 Avaliação

O dataset foi previamente dividido aleatoriamente em três conjuntos — treino, validação e teste — e apenas os dois primeiros serão disponibilizados para que vocês implementem as suas soluções.

No sábado anterior ao prazo final de submissão, iremos disponibilizar no Moodle o conjunto de teste e iremos avisá-los pelo canal da disciplina no Slack. No relatório, vocês devem reportar tudo que foi pedido na seção Tarefas.

A avaliação consistirá da análise do relatório e do código submetidos no Moodle. Iremos avaliar se as tarefas pedidas foram realizadas, como o treinamento e validação foram feitos, os resultados reportados e as conclusões.

Observações sobre a avaliação:

- O trabalho deverá ser feito em duplas ou trios, podendo haver repetição dos membros a cada trabalho;

- O código (arquivo .R) e o relatório (formato .pdf) deverão ser submetidos no Moodle por **apenas um integrante do grupo**;
- Não se esqueçam de listar os nomes dos integrantes do grupo no início do relatório e no código;
- As notas do trabalho serão divulgadas em até uma semana após o prazo da submissão;