Métodos de Regressão
Aprendizagem de Máquina
2023

André Luiz Brun[1]

[1]Colegiado de Ciência da Computação
Campus de Cascavel - UNIOESTE

Resumo. *Este documento consiste na especificação formal do terceiro trabalho da disciplina de Aprendizagem de Máquina (Csc3040) para o ano letivo de 2023. Aqui são descritas as atividades a serem desenvolvidas e como cada processo deverá ser realizado. Além disso, o documento contém as informações sobre a formação das equipes, o objeto de trabalho de cada uma e as datas de entrega e apresentação dos relatórios.*

1. Introdução

O objetivo do terceiro trabalho da disciplina consiste em comparar o comportamento, em termos de precisão, de métodos de regressão baseados em diferentes conceitos sobre uma mesma base de dados. Como critério de avaliação serão computadas as medidas do erro médio absoluto (MAE), erro médio quadrático (MSE) e raiz quadrada do erro médio (RMSE).

Espera-se, através da execução dos experimentos, que cada equipe possa identificar a abordagem que foi mais adequada, segundo os critérios definidos, ao seu conjunto de dados.

2. Implementação

Nesta seção é descrito como cada etapa do desenvolvimento deve ser realizada segundo os conceitos vistos durante a disciplina. Deverão ser implementadas uma estratégia de regressão linear múltipla (que não usa Aprendizagem de Máquina), três abordagens monolíticas de regressão (KNR, SMR e MLP) e duas estratégias baseadas em sistemas de múltiplos regressores (Random Forest e Gradient Boosting).

2.1. Análise descritiva dos dados

Nesta etapa deve-se fazer uma análise descritiva dos dados, apresentando características da base como tamanho, dimensão, origem, número de classes, tipos dos atributos, valores médios, máximos e mínimos dos atributos etc. **Além disso, é necessário fazer uma explanação sobre o problema em questão, falando um pouco da aplicação e do significado dos dados coletados**.

Ademais, **deve feita uma análise de correlação entre os atributos**. Este processo pode ser feito utilizando-se o coeficiente de correlação de Pearson ou mesmo através de uma representação gráfica bidimensional (Heatmap) em que cada eixo representa os valores de um dos atributos.

As bases estão serão distribuídas aleatoriamente para cada equipe e serão disponibilizadas, em formato .csv através dos links disponíveis na Seção 3. Cada equipe terá seu link específico, de acordo com o arquivo que lhe é destinado.

2.2. Divisão do conjunto de dados

O primeiro passo consistirá na divisão da base original em três subconjuntos mutuamente exclusivos: treino, teste e validação (conforme apresentado na Figura 1). A instância que for designada para um conjunto não deve aparecer nos outros.

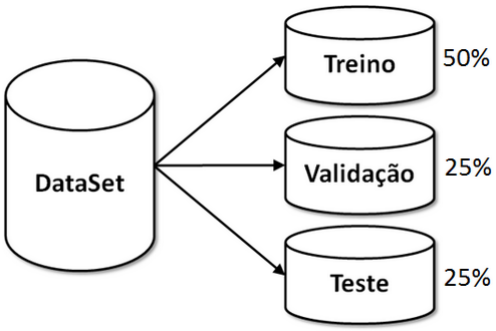

Figura 1. Divisão do conjunto de dados original

O conjunto de treino deverá possuir 50% do tamanho do arquivo original. Já as bases de validação e teste, terão 25% do número de instâncias da base de entrada. Como não há classes específicas para os registros, a divisão pode ser feita randomicamente.

Importante 1: a escolha das instâncias que formarão cada um dos conjuntos deve ser totalmente aleatória.

Importante 2: lembrem-se de sempre "bagunçar" os conjuntos de dados **antes** de fazer a divisões e de realizar o treinamento. A adoção de aleatoriedade adiciona robustez ao processo.

2.3. Treinamento e Calibração dos Modelos

Depois de formados os conjuntos, o passo seguinte será o treinamento dos modelos de regressão. Nesta tarefa deverão ser implementadas as estratégias dos K Vizinhos mais próximos (KNR), Máquina de Vetores de Suporte (SVR), Multilayer Perceptron (MLP), Random Forest (RF) e Gradient Boosting (GB). Além disso, deverá ser implementada a abordagem de Regressão Linear Múltipla clássica (RLM), que não usa Aprendizado de Máquina.

Para se determinar quais os melhores parâmetros dos métodos deve-se adotar o conjunto de validação (conforme ilustrado na Figura 2). Por exemplo, digamos que estamos treinando um KNR e queremos decidir qual o melhor K a ser empregado. Deve-se treinar o regressor com o conjunto de treino e então variar o valor de K e analisar quão próxima é a estimativa de precisão do modelo. O valor de K que obter o menor RMSE é usado no momento de estimar o conjunto de teste. Os parâmetros que deverão ser definidos para cada modelo de regressão são apresentados na Tabela 1.

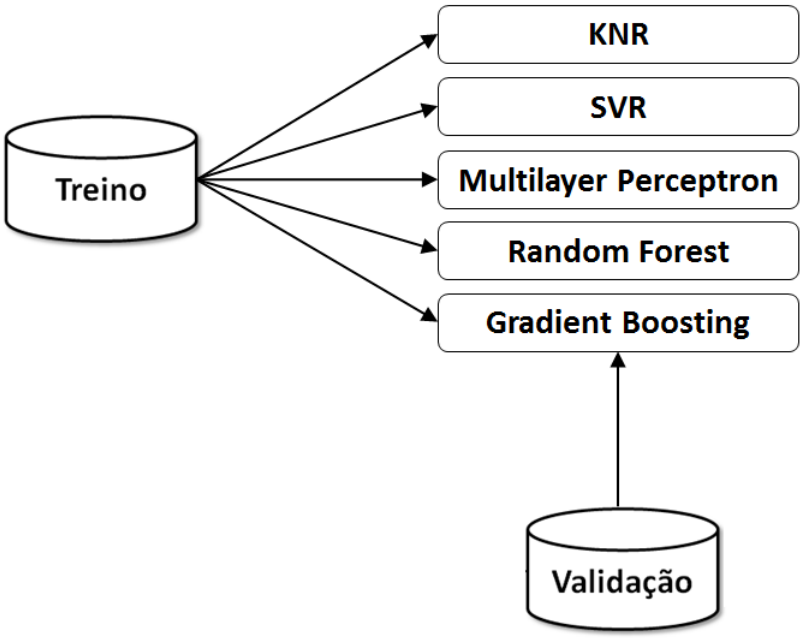

Figura 2. Adoção do conjunto de validação na estimação dos parâmetros

Tabela 1. Conjunto de parâmetros a serem calibrados através do Grid-search

Classificador	Parâmetros
KNR	K distance
SVR	kernel C
MLP	hidden_layer_sizes activation max_iter learning_rate
RF	n_estimators criterion max_depth min_samples_split min_samples_leaf
GB	n_estimators loss max_depth learning_rate min_samples_split min_samples_leaf
RLM	—

2.4. Avaliação dos Modelos

Definidos os melhores parâmetros para cada método de regressão, o passo seguinte consiste em estimar as medidas de avaliação especificadas sobre o conjunto de teste, conforme ilustrado na Figura 3. Para cada uma das abordagens testadas deverão ser obtidas as métricas de Erro Médio Absoluto (MAE), Erro Médio Quadrático (MSE) e Raiz do Erro Médio Quadrático (RMSE).

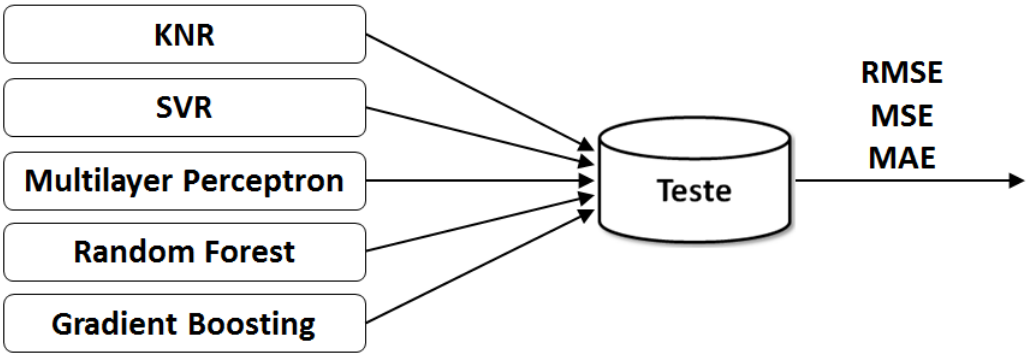

Figura 3. Ilustração do processo de avaliação dos métodos de regressão.

Para que o processo tenha base para análise estatística, deverão ser executadas 20 repetições. Os valores a serem comparados deverão ser os valores médios das 20 execuções. Um exemplo de representação dos resultados é ilustrado na Tabela 2 onde cada coluna corresponde ao desempenho de um método de regressão ao longo das 20 execuções do experimento. Na última linha são apresentados a RMSE média e o desvio padrão do longo das execuções.

Tabela 2. Exemplo de estrutura para análise dos resultados dos sistemas monolíticos

Repetição	KNR	SVR	MLP	RF	GB	RLM
1	$RMSE_1$	$RMSE_1$	$RMSE_1$	$RMSE_1$	$RMSE_1$	$RMSE_1$
2	$RMSE_2$	$RMSE_2$	$RMSE_2$	$RMSE_2$	$RMSE_2$	$RMSE_1$
...	... ...	...	...	...	...	
20	$RMSE_{20}$	$RMSE_{20}$	$RMSE_{20}$	$RMSE_{20}$	$RMSE_{20}$	$RMSE_{20}$
	Média (DP)	Média (DP)	Média (DP)	Média (DP)	Média (DP)	Média (DP)

2.5. Análise Comparativa

A última etapa consiste na comparação dos erros dos métodos para descobrir qual deles obteve o melhor desempenho. Para tanto, deve-se executar dois testes estatísticos. O primeiro servirá para detectar se há diferença entre o desempenho dos algoritmos (independente de qual foi melhor ou pior). O segundo teste estatístico serve para comparar, dois a dois, os regressores com o objetivo de avaliar se eles têm desempenhos significativamente diferentes e quem é o melhor.

Para avaliar se há pelo menos um modelo de regressão com desempenho diferente dos demais utilizem o teste de Kruskal-Wallis com 5% de significância. Caso haja

pelo menos um regressor com comportamento diferente deve-se aplicar o teste de Mann-Whitney (bicaudal), também com 5% de significância, para identificar quais modelos apresentaram comportamento discrepante.

Os testes podem ser realizados via código em python, usando a biblioteca scipy (conforme exemplo visto em sala) ou pelos endereços Kruskal-Wallis e Mann-Whitney.

2.6. Como fazer?

A linguagem adotada é de escolha da dupla. Entretanto, é fortemente indicado o uso de Python ou Java.

Não é necessário implementar os métodos de regressão. Neste caso, pode-se e é indicado, que sejam utilizadas implementações prontas dos métodos, ficando a carga da dupla apenas a implementação do framework e análise dos parâmetros e resultados.

3. Equipes

Na Tabela 3 são apresentadas as composições de cada equipe bem como o problema sobre qual cada uma trabalhará. Além disso, são apresentados os endereços eletrônicos onde as bases de dados podem ser obtidas.

Tabela 3. Formação das equipes e conjunto de dados para o trabalho

ID	Equipe	Fonte
1	Felipi Lima Matozinho João Luiz Reolon	link
2	Gabriel Norato Claro Maria Eduarda Crema Carlos	link
3	Jaqueline Cavaller Faino Davi Marchetti Giacomel	link
4	Bruno Stafuzza Maion Rodrigo da Rosa	link
5	Heloisa Aparecida Alves Vinicius Muller de Freitas	link
6	Gustavo Pauli da Luz Guilherme de Oliveira Correia	link
7	Rodrigo Brickmann Rocha Gabriel Alves Mazzuco	link

4. O que deve ser entregue

4.1. Relatório

Deve ser elaborado um relatório técnico em formato pdf contendo:

- Detalhamento de quais foram os parâmetros empregados em cada método de regressão e em qual faixa de valores cada parâmetro foi variado. Por exemplo, no KNR, seria possível variar o valor de k entre 1 e 50.
- Análise detalhada das métricas de desempenho (MAE, MSE e RMSE) obtidas para cada modelo, inclusive a tabela com os dados de cada iteração.

- Análise pertinente indicando quais métricas melhor representam o desempenho dos algoritmos perante o conjunto de entrada.
- Comparação adequada e embasada das seis estratégias de regressão testadas.

O formato do relatório deve ser a formatação presente neste texto. As regras para tal podem ser obtidas no link download. No arquivo disponível pode-se utilizar a formatação em arquivo .doc ou em latex.

4.2. Código-fonte

Além do relatório citado, cada equipe deverá enviar os códigos-fonte construídos para a execução dos experimentos. Ambos arquivos podem ser compactados e enviados como arquivo único.

5. Para quando?

O trabalho deverá ser submetido no link disponibilizado na turma de disciplina dentro do ambiente Microsoft Teams até as **13:30 do dia 04/12/2023**.

As apresentações serão realizadas na aula do dia 05/12/2023.

Cada grupo terá 15 minutos para apresentar o trabalho realizado, focando na descrição do problema, nos desempenhos obtidos e na comparação dos desempenhos alcançados.