TP555 - AI/ML

Nome: Mayomona Lando Filipe

Matricula: 836

Lista de Exercícios #3

Regressão Polinomial

Suponha que você esteja usando regressão polinomial. Você plota as curvas de aprendizado e percebe que há uma grande diferença entre o erro de treinamento e o erro de validação. **O que está acontecendo?**

R: deve esta acontecendo um Overfitting! isso quer dizer ,que nosso algoritmo funciona bem no conjunto de treinamento, mas não consegue ter um desempenho melhor nos conjuntos de teste; por outro o mesmo erro de treinamento tende a diminuir conforme aumentamos o grau do polinômio. Devido ao superajuste e o erro de validação tende a diminuir conforme diminuirmos o grau do polinômio, mas apenas até um certo ponto. Após esse ponto, esse erro irá aumentar, formando uma curva de erro convexa. **Quais são as três maneiras de resolver isso?**

OBS: Curvas de aprendizado: são gráficos mostrando o desempenho do modelo no conjunto de treinamento e no conjunto de validação em função do tamanho do conjunto de treinamento (ou da iteração do treinamento).

R: as 3 maneiras de resolução do problemas é usando algumas técnicas tais como:

Técnicas de Regularização.

Curvas de aprendizado são gráficos que mostram o desempenho do modelo no conjunto de treinamento.

Técnica validação cruzada é uma técnica para avaliar modelos por meio de treinamento de vários modelos de ML em subconjuntos de dados de entrada disponíveis e avaliação deles no subconjunto complementar dos dados. Usamos a validação cruzada para detectar sobreajuste, ou seja, a não generalização de um padrão.

Ao criarmos e treinarmos um modelo devemos selecionar o que faz as melhores previsões, o que significa escolher o modelo com as melhores configurações, que pode definir as etapas, regularização, tamanho do modelo e tipo de ordem aleatória.

No entanto, se nos selecionar-mos configurações de parâmetro de modelo que produzam o melhor desempenho de previsões com dados de avaliação, pode sobreajustar o modelo. O sobreajuste ocorre quando um modelo memoriza padrões que aparecem nas fontes de dados de avaliação e, mas não consegue generalizar os padrões nos dados. Isso geralmente acontece quando os dados de treinamento incluem todos os dados usados na avaliação. Um modelo sobreajustado tem bom desempenho durante

Para evitar a seleção de um modelo sobreajustado como o melhor modelo, devemos reservar dados adicionais para validar o desempenho do modelo. Por exemplo: dividir os dados em 60 por cento para treinamento, 20 por cento para avaliação e outros 20 por cento para validação. Após selecionar os parâmetros do modelo que funcionam bem com os dados de avaliação, execute uma segunda avaliação com os dados de validação para ver o desempenho do modelo com os dados de validação. Se o modelo atende às expectativas com os dados de validação, o modelo não está sobreajustando os dados.

O uso de terceiro conjunto de dados para validação ajuda a selecionar os parâmetros de modelo adequados para evitar o sobreajuste. No entanto, a retenção de dados do processo de treinamento tanto para avaliação como para validação disponibiliza menos dados para treinamento. Isso é um problema principalmente com conjuntos pequenos de dados porque é sempre melhor usar o máximo de dados possível para treinamento. Para resolver esse problema, devemos executar a validação cruzada.