A aprendizagem mecânica é um método de análise de dados que automatiza a construção de modelos analíticos. Utilizando algoritmos que aprendem iterativamente com os dados, a aprendizagem de máquinas permite que os computadores encontrem conhecimentos ocultos sem serem explicitamente programados onde procurar.

A aprendizagem mecânica é um método de análise de dados que automatiza a construção de modelos analíticos. Utilizando algoritmos que aprendem iterativamente com os dados, a aprendizagem de máquinas permite que os computadores encontrem conhecimentos ocultos sem serem explicitamente programados onde procurar.

Para que é utilizada?

Detecção de fraudes.

Resultados de pesquisa na Web.

Anúncios em tempo real em páginas web

Pontuação do crédito.

Predição de falhas de equipamento.

Novos modelos de preços.

Detecção de intrusão na rede.

Motores de Recomendação

Segmentação de clientes

Análise de Sentimento de Texto

Rotatividade do cliente

Padrão e reconhecimento de imagem.

O que são redes neurais?

As Redes Neurais são uma forma de modelar matematicamente os sistemas de neurónios biológicos.

Estas redes podem então ser utilizadas para resolver tarefas que muitos outros tipos de algoritmos não conseguem (por exemplo, classificação de imagens).

Deep Learning refere-se simplesmente a redes neurais com mais do que uma camada oculta.

Aprendizagem mecânica

Modelos analíticos automatizados.

Redes Neuronais

Um tipo de arquitectura de aprendizagem mecânica modelada após neurónios biológicos.

Aprendizagem profunda

Uma rede neurológica com mais do que uma camada oculta.

Aprendizagem supervisionada

Diagram

Description automatically generated

Os algoritmos de aprendizagem supervisionados são treinados utilizando exemplos rotulados, tais como uma entrada onde a saída desejada é conhecida.

Por exemplo, um segmento de texto poderia ter uma etiqueta de categoria, como por exemplo:

Spam vs. Email Legítimo

Revisão de Filme Positivo vs. Negativo

A rede recebe um conjunto de entradas juntamente com as saídas correctas correspondentes, e o algoritmo aprende comparando a sua saída real com as saídas correctas para encontrar erros.

Modifica então o modelo em conformidade.

A aprendizagem supervisionada é normalmente utilizada em aplicações em que os dados históricos prevêem prováveis eventos futuros.

Sobreajustamento e subajustamento

Agora que compreendemos todo o processo de aprendizagem supervisionada, vamos abordar os importantes tópicos de sobre e subapetrechamento.

Sobreajustamento

O modelo adapta-se demasiado ao ruído dos dados.

Isto resulta frequentemente em erros baixos nos conjuntos de treino mas erros elevados nos conjuntos de teste/validação.

Subajustamento

O modelo não capta a tendência subjacente dos dados e não se adapta suficientemente bem aos dados.

Baixa variância, mas com elevado enviesamento.

O subajuste é frequentemente o resultado de um modelo excessivamente simples.

Quando pensamos em sobreajustamento e subajustamento, queremos ter em mente a relação entre o desempenho do modelo no conjunto de treino versus o conjunto de teste/validação.

Avaliar o desempenho

CLASSIFICAÇÃO

As principais métricas de classificação que precisamos de compreender são:

Accuracy

Recall

Precision

Recordar

Precisão

F1-Score

Precisão

A precisão nos problemas de classificação é o número de previsões correctas feitas pelo modelo dividido pelo número total de previsões.

A precisão é útil quando as classes-alvo estão bem equilibradas

No nosso exemplo, teríamos aproximadamente a mesma quantidade de imagens de gato que temos imagens de cão.

A precisão não é uma boa escolha com classes desequilibradas!

Imagine que tínhamos 99 imagens de cães e 1 imagem de um gato.

Se o nosso modelo fosse simplesmente uma linha que sempre previu cão, teríamos 99% de exactidão!

Relembrar

Capacidade de um modelo para encontrar todos os casos relevantes dentro de um conjunto de dados.

A definição precisa de recolha é o número de verdadeiros positivos dividido pelo número de verdadeiros positivos mais o número de falsos negativos.

Habilidade de um modelo de classificação para identificar apenas os pontos de dados relevantes.

A precisão é definida como o número de verdadeiros positivos dividido pelo número de verdadeiros positivos mais o número de falsos positivos.

Recall e Precisão

Muitas vezes há uma troca entre a Recall e a Precision.

Enquanto a Rechamada expressa a capacidade de encontrar todas as instâncias relevantes num conjunto de dados, a precisão expressa a proporção dos pontos de dados que o nosso modelo diz ter sido relevante eram realmente relevantes.

F1-Score

A pontuação de F1 é a média harmónica de precisão e recordação tendo em conta ambas as métricas na seguinte equação:

Table

Description automatically generated

Utilizamos a média harmónica em vez de uma simples média, porque castiga valores extremos.

Um classificador com uma precisão de 1,0 e uma recordação de 0,0 tem uma média simples de 0,5 mas uma pontuação F1 de 0.

Também podemos ver todas as imagens classificadas correctamente versus imagens classificadas incorrectamente sob a forma de uma matriz de confusão.

Table

Description automatically generated

Utilizamos a média harmónica em vez de uma simples média, porque castiga valores extremos.

Um classificador com uma precisão de 1,0 e uma recordação de 0,0 tem uma média simples de 0,5 mas uma pontuação F1 de 0.

Também podemos ver todas as imagens classificadas correctamente versus imagens classificadas incorrectamente sob a forma de uma matriz de confusão.

Avaliando o desempenho

Vamos agora discutir a avaliação dos Modelos de Regressão

A regressão é uma tarefa quando um modelo tenta prever valores contínuos (ao contrário dos valores categóricos, que é a classificação)

Já deve ter ouvido falar de algumas métricas de avaliação como precisão ou recordação.

Este tipo de métricas não são úteis para problemas de regressão, precisamos de métricas concebidas para valores contínuos!

ou exemplo, tentar prever o preço de uma casa dadas as suas características é uma tarefa de regressão.

A tentativa de prever o país em que uma casa está, dadas as suas características, seria uma tarefa de classificação.

Vamos discutir algumas das métricas de avaliação mais comuns para a regressão:

Erro Médio Absoluto

Erro médio quadrático

Erro Quadrado Médio de Raiz

Erro Médio Absoluto (MAE)

Esta é a média do valor absoluto dos erros.

Fácil de compreender

Text

Description automatically generated with medium confidence

Erro médio quadrático (MSE)

Este é o meio dos erros ao quadrado.

Erros maiores são mais notados do que com o MAE, tornando o MSE mais popular.

Text

Description automatically generated

Erro Quadrado Médio de Raiz (RMSE)

Esta é a raiz da média dos erros ao quadrado.

Mais popular (tem as mesmas unidades que y)

Text

Description automatically generated with medium confidence

**Unsupervised Learning**

Cobrimos a aprendizagem supervisionada, onde o rótulo era conhecido devido a dados históricos etiquetados.

Mas o que acontece quando não temos rótulos históricos?

Há certas tarefas que se enquadram na aprendizagem não supervisionada:

Clustering

Detecção de anomalias

Redução da dimensionalidade

Clustering

Agrupamento de pontos de dados não etiquetados em categorias/classificadores

Os pontos de dados são atribuídos a um cluster com base na semelhança

Detecção de anomalias

Tentativas de detectar aberrantes num conjunto de dados

Por exemplo, transacções fraudulentas com um cartão de crédito.

Redução da dimensão

Técnicas de processamento de dados que reduzem o número de características de um conjunto de dados, quer para compressão, quer para melhor compreender as tendências subjacentes dentro de um conjunto de dados.

Diagram

Description automatically generated