Machine Learning

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.

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.

Filtragem de spam por e-mail.

Há diferentes tipos de aprendizagem de máquinas em que nos vamos concentrar durante as próximas secções do curso:

Aprendizagem supervisionada

Aprendizagem sem supervisão

Aprendizagem mecânica

Modelos analíticos automatizados.

Redes Neuronais

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

Aprendizagem profunda

Uma rede neural com mais do que uma camada oculta.

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 corretas 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.

Diagram

Description automatically generated

Obtenha os seus dados! Clientes, Sensores, etc...

Limpe e formate os seus dados (usando Pandas)

O que acabámos de mostrar é uma abordagem simplificada à aprendizagem supervisionada, que contém um problema!

Será justo usar a nossa divisão única dos dados para avaliar o desempenho dos nossos modelos?

Afinal de contas, foi-nos dada a oportunidade de atualizar os parâmetros do modelo uma e outra vez.

Para resolver este problema, os dados são frequentemente divididos em 3 conjuntos

Dados de Formação: Utilizado para treinar parâmetros do modelo

Dados de Validação: Utilizado para determinar que hiperparâmetros de modelo a ajustar

Dados de teste: Usado para obter alguma métrica de desempenho final

Isto significa que depois de vermos os resultados no conjunto final do teste não podemos voltar atrás e ajustar quaisquer parâmetros do modelo!

Esta medida final é o que rotulamos o verdadeiro desempenho do modelo a ser.

Neste curso, em geral, simplificamos os nossos dados utilizando uma simples divisão treino/teste.

Iremos simplesmente treinar e depois avaliar num conjunto de testes (deixando a opção aos alunos de voltar atrás e ajustar parâmetros).

Depois de passar pelo curso, poderá facilmente realizar outra divisão para obter 3 conjuntos de dados, se desejar.

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.

Estes dados eram fáceis de visualizar, mas como podemos ver sub e sobreajustamento quando lidamos com conjuntos de dados multidimensionais?

Primeiro vamos imaginar que treinámos um modelo e depois medimos o seu erro ao longo do tempo de treino.

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

Acabámos de saber que, após a conclusão do nosso processo de aprendizagem da máquina, utilizaremos métricas de desempenho para avaliar a forma como o nosso modelo se saiu.

Vamos discutir a métrica de classificação com mais detalhe!

As principais métricas de classificação que precisamos de compreender são:

Precisão

Recordar

Precisão

F1-Score

Mas primeiro, devemos compreender o raciocínio por detrás destas métricas e como elas funcionarão realmente no mundo real!

Tipicamente, em qualquer tarefa de classificação o seu modelo só pode alcançar dois resultados:

Ou o seu modelo estava correto na sua previsão.

Ou o seu modelo estava incorrecto na sua previsão.

Felizmente, incorrecto vs correcto expande-se para situações em que tem múltiplas classes.

Para efeitos de explicação das métricas, imaginemos uma situação de classificação binária, em que só temos duas classes disponíveis.

No nosso exemplo, vamos tentar prever se uma imagem é um cão ou um gato.

Uma vez que se trata de aprendizagem supervisionada, vamos primeiro ajustar/formar um modelo sobre dados de treino, depois testar o modelo sobre dados de teste.

Assim que tivermos as previsões do modelo a partir dos dados do X\_test, comparamo-lo com os verdadeiros valores y (as etiquetas corretas).

Repetimos este processo para todas as imagens nos nossos X dados de teste.

No final teremos uma contagem dos resultados correctos e uma contagem dos resultados incorrectos.

A realização chave que precisamos de fazer, é que no mundo real, nem todos os resultados incorrectos ou correctos têm o mesmo valor!

Também no mundo real, uma única métrica não contará a história completa!

Para compreender tudo isto, vamos trazer de volta as 4 métricas que mencionámos e ver como elas são calculadas.

Poderíamos organizar os nossos valores previstos em comparação com os valores reais numa matriz de confusão.

Precisão

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

Por exemplo, se o conjunto X\_test foi 100 imagens e o nosso modelo previu correctamente 80 imagens, então temos 80/100.

0,8 ou 80% de precisão.

A exactidão 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!

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

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

Nesta situação, vamos querer compreender a recordação e a precisão

Recordar

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.

Precisão

Capacidade 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.

Recolha 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

Nos casos em que pretendemos encontrar uma mistura óptima de precisão e recordação, podemos combinar as duas métricas usando o que se chama a pontuação F1.

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 with medium confidence

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

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

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

Avaliação do modelo

O principal ponto a lembrar com a matriz de confusão e as várias métricas calculadas é que todas elas são fundamentalmente formas de comparar os valores previstos versus os verdadeiros valores.

O que constitui uma "boa" métrica, dependerá realmente da situação específica!

Ainda confuso com a matriz de confusão?

Não há problema! Consulte a página da Wikipédia, tem um diagrama realmente bom com todas as fórmulas para todas as métricas.

Ao longo de toda a formação, normalmente só imprimimos métricas (por exemplo, exatidão).

Pensemos nesta ideia de:

O que é uma exatidão suficientemente boa?

Tudo isto depende do contexto da situação!

Foi criado um modelo para prever a presença de uma doença?

A presença da doença é bem equilibrada na população em geral? (Provavelmente não!)

Muitas vezes os modelos são utilizados como testes de diagnóstico rápido antes de se fazer um teste mais invasivo (por exemplo, fazer um teste de urina antes de se fazer uma biopsia)

Temos também de considerar o que está em jogo!

Muitas vezes temos uma troca de precisão/rechamada. Temos de decidir se o modelo deverá concentrar-se na fixação de Falsos Positivos vs. Falsos Negativos.

No diagnóstico de doenças, é provavelmente melhor ir na direção dos Falsos Positivos, por isso certificamo-nos de classificar corretamente o maior número possível de casos de doença!

Tudo isto é dizer que a aprendizagem mecânica não é realizada num "vácuo", mas sim num processo de colaboração em que devemos consultar especialistas no domínio (por exemplo, médicos).

Avaliação da Regressão

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!

Por 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

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.

Erro Quadrado Médio de Raiz (RMSE)

Esta é a raiz da média dos erros ao quadrado.

Mais popular (tem as mesmas unidades que y)

**Aprendizagem sem supervisão**

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.

**Aprendizagem sem supervisão**

É importante notar que estas são situações em que não temos a resposta correcta para os dados históricos!

O que significa que a avaliação é muito mais difícil e mais matizada!

Diagram

Description automatically generated