__/__/____

DCE33 Páginas 1 até 9

Sumário

DCE 33 ___/___/___

Aprendizado de Máquina não supervisionado

1. Introdução

Aprendizado de máquina é o ramo da inteligência artificial que permite que os sistemas aprendam sem que sejam explicitamente programados. Ele envolve a construção e treinamento de modelos (algoritmos), capazes de resolver tarefas complexas, sem que as regras de resolução das tarefas, sejam explicitamente colocadas nos códigos destes modelos.

A classificação geral do aprendizado de máquina distingue duas categorias: aprendizado de máquina supervisionado e o aprendizado de máquina não-supervisionado. Outras classificações incluem o aprendizado semi-supervisionado, aprendizado por reforço. O aprendizado de máquina supervisionado distingue-se do não-supervisionado pelos rótulos (ou respostas) dos dados de treinamento, enquanto o primeiro (supervisionado) possuí rótulos, o segundo (não supervisionado) não tem.

Para exemplificar, supomos que os dados de treinamento são referente a imóveis, tais como: área (m^2), nº de cômodos,

localização. No caso do aprendizado supervisionado poderíamos ter como rotulos o tipo de imóvel: popular ou luxo no caso do aprendizado não supervisionado não existem rótulos.

O aprendizado não supervisionado resolve muitas tarefas da vida real, já que a realidade da maioria das situações é um grande conjunto de dados sem rotulos.

2. Aprendizado de máquina não supervisionado

O aprendizado não supervisionado busca encontrar padrões ocultos nos dados. Possui três componentes básicos: entradas, métricas de distância ou semelhança e algoritmos. As entradas também conhecidas pelo termo inglês "features", referem-se aos dados, que trazem atributos e caracteristicas, que sendo usados para treinar os modelos. As métricas de distância/semelhança sendo utilizadas para avaliar a qualidade dos padrões encontrados. Algoritmos são usados para treinar os modelos e dependem do tipo de problema ou tarefa a ser executa.

2.1 Tarefas

Há três principais tarefas no

DCE33 ___/___/___

aprendizado não supervisionado: agrupamento (clusterização), redução de dimensionalidade e associação.

2.1.1 Agrupamento

Agrupamento é uma das tarefas mais populares do aprendizado não supervisionado e consiste em reunir (agrupar) as entradas de acordo com critérios de distância e/ou semelhança entre os pontos destes dados. É importante lembrar que as máquinas precisam que os dados sejam numéricos para conseguir processá-los. Assim, após esta conversão os dados de entrada podem ser representados por pontos num espaço de dimensão "n".

O algoritmo mais conhecido para a tarefa de agrupamento é o K-means. O algoritmo acha os centróides, que são os pontos centrais dos pontos próximos e forma um número "K" de grupos ou clusters. No caso do K-means, o centróide é a média aritmética dos pontos contidos no grupo. Outros algoritmos populares são o DBSCAN e o agrupamento hierárquico. Duas métricas importantes para o agrupamento são coesão e distância entre clusters. Coesão diz respeito a quão próximos uns dos outros os pontos contidos num cluster estão. A dis-

PCE33

tância entre clusters mede o quão distantes os grupos estão entre si. O ideal é um sistema com seus pontos próximos dentro de cada cluster (coesão) e com seus diferentes clusters distantes entre si, indicando um bom agrupamento. Portanto, o ideal seria combinar ambos coesão e distância entre clusters para avaliar a qualidade do modelo. Assim existem métricas que se preocupam com coesão e distância entre grupos como: coeficiente de silhouette, coeficiente Dunn e coeficiente Davis Bouldin.

O agrupamento é aplicado em segmentação de clientes, estratégias de marketing para encontrar perfis de usuários, entre outras aplicações.

2.1.2. Redução de Dimensionalidade

Em muitos casos, quando os dados são convertidos em pontos num espaço de dimensão "n", este "n" é muito alto, o que além de dificultar a compreensão, torna-se computacionalmente custoso processá-los. A redução de dimensionalidade visa reduzir este espaço dimensional mantendo o máximo de informações e variedade dos dados de entrada. Os algoritmos mais conhecidos PCA (Análise do Componentes Principais) e o t-SNE. Um desafio é saber qual o número ideal de

DCE33 ___/___/___

componentes os dados terão, de modo a manter suas características e diversidade. Este número ideal depende do algoritmo utilizado, mas há métricas e ferramentas que nos ajudam a encontrá-lo, por exemplo a variância explicada e o erro. Com a variância explicada podemos saber o percentual de variância dos dados que foi mantida com um número "x" de componentes. Traçando o gráfico da variância explicada acumulada podemos verificar quantos componentes são necessários para atingir o percentual de variância desejado. Já o erro, calcula o quanto o componente corresponde ao valor original (diferença entre o ponto gerado e o ponto original). De modo similar, podemos encontrar o nº de componentes que produza um erro até um certo limite, por exemplo, $erro_{MAX}=0.01$.

A redução de dimensionalidade é bastante usada no pré-processamento de texto e imagens, que possuem alta dimensão para seus dados

2.1.3. Associação

Um pouco menos popular a associação é uma tarefa que busca encontrar padrões que juntos indiquem uma tendência. Por exemplo, baseado nos itens de

DCE33

compras de um consumidor, o sistema pode sugerir novos itens. Um algoritmo utilizado é o Apriari, que busca encontrar padrões baseado na variância dos dados.

3. Ciclo de vida do processo do aprendizado não supervisionado (ANS)

O ciclo de vida do ANS inclue como principais etapas: a coleta de dados, o préprocessamento, o treinamento, a avaliação do desempenho e a implantação.

A coleta de dado refere-se a etapa onde os dados são reunidos para serem utilizados. Usa-se conjuntos de dados publicos, API, técnicas de raspagem de dados da internet ou utiliza-se dados gerados internamente.

O pré-processamento dos dados é uma etapa crucial para garantir dados de qualidade, pois não há como gerar bons modelos se a mátéria-prima (dados de entrada) não é de boa qualidade. O pré-processamento inclue a limpeza dos dados (duplicidades, erros, etc), o tratamento de dados faltantes (exclusão de amostras, imputação de dados), normalização/padronização, redução de dimensionalidade, entre outras.

DE33 ___/___/___

O treinamento do modelo inclui utilizar o algoritmo escolhido, ajustando os parâmetros para reduzir o erro. Técnicas como GridSearch e validação cruzada são usadas em alguns treinamentos para evitar o super ajuste dos dados (overfitting) ou um sub ajuste dos dados (underfitting).

A avaliação de desempenho é a etapa onde usamos as métricas para avaliar o quão bom nosso modelo está. É importante usar métricas apropriadas para a tarefa e seu objetivo e impacto dos resultados

Implantação do modelo refere-se a etapa em que o modelo é colocado em um sistema num ambiente de produção para ser usado pelos usuários finais. Inclui monitorar, manter e escalar o modelo.

4. Desafios e limitações

Os modelos de aprendizado de máquina não supervisionado podem ser difíceis de explicar e compreender os padrões, sendo muitas vezes necessária a atuação de especialistas do domínio que está sendo trabalhado. Outro problema é quando tratamos dados de altas dimensões ou com um volume grande

DCC33

de dados, tornando-se bastante custoso computacionalmente. A presença de dados anômalos (outliers) pode distorcer os resultados dos modelos.

5. Tendências

A integração do aprendizado de máquina não supervisionado com as novas tecnologias como "Internet das Coisas" e Big Data, podem trazer uma grande evolução. Assim como a combinação deste tipo de aprendizado com outros tipos. Fala-se em Auto-aprendizagem criando modelos que criam e treinam outros modelos e no aprendizado com redes neurais profundas, como forte tendência associada ao aprendizado não-supervisionado

6. Conclusão

O aprendizado de máquina não supervisionado torna-se um campo de bastante interesse, tendo em vista a produção exponencial de dados não rotulados. Importantes algoritmos com Kmeans, DBSCAN e PCA, ainda são bastante usados e tem respondido com eficiência em muitos problemas reais. Porem, é evidente que pesquisas neste campo de es

DCE33 ___/___/____

tudo são mecanismos para adaptar ou desenvolver novas técnicas para lidar com tecnologias como Internet das Coisas e Big Data.