DCE 88 ___/___/___

TEMA 02 - APRENDIZADO DE MÁQUINA NÃO-SUPERVISIONADO

01 - INTRODUÇÃO

O APRENDIZADO DE MÁQUINA NÃO-SUPERVISIONADO, TAMBÉM CONHECIDO COMO MODELOS DESCRITIVOS, SÃO UMA TÉCNICA DE INTELIGÊNCIA ARTIFICIAL (IA) ONDE SUA PRINCIPAL CARACTERÍSTICA É TRABALHAR COM DADOS NÃO ROTULADOS. DIFERENTEMENTE DO APRENDIZADO SUPERVISIONADO, ONDE TODA A BASE DE DADOS ESTA DEVIDAMENTE ROTULADA, NO APRENDIZADO NÃO-SUPERVISIONADO ESSE PROCESSO DE MARCAÇÃO DE BASES PODE SER CARO, TANTO EM RECURSO HUMANO QUANTO EM TEMPO, NO CASO FOR EXEMPLO DE MARCAÇÃO DE BASES DE DADOS DE IMAGENS, OU MUITAS VEZES IMPRATICÁVEL, COMO NO CASO DE RECOMENDAÇÃO DE NOVAS MÚSICAS BASEADO NO QUE UM USUÁRIO OUVE EM UM APLICATIVO DE STREAMING DE MÚSICAS.

O APRENDIZADO DE MÁQUINA NÃO-SUPERVISIONADO BUSCA, ATRAVÉS DE ~~SUAS ALGORITMOS~~ TÉCNICAS: 1- ENCONTRAR SIMILARIDADES ENTRE OS DADOS ATRAVÉS DE AGRUPAMENTO EM CLUSTERS, OU SEJA, AGRUPAR DADOS QUE POSSUAM CARACTERÍSTICAS SEMELHANTES; 2- ASSOCIAR CONJUNTOS DE DADOS QUE POSSUEM UMA FREQUÊNCIA DE USO EM CONJUNTO, POR EXEMPLO, QUEM COMPRA WHEY PROTEIN, FREQUENTEMENTE COMPRA CREATINA; 3- REDUZIR A DIMENSIONALIDADE DE UM GRANDE CONJUNTO DE DADOS, SIMPLIFICANDO-O PARA QUE O MESMO POSSA SER ANALISADO.

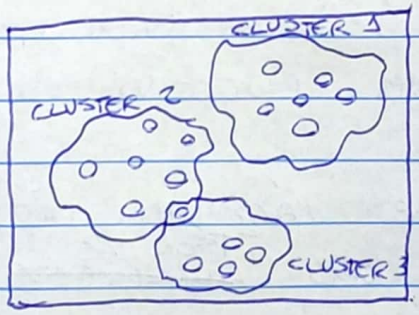
__/__/____ DCE88

02 - ~~Algoritmos~~ Técnicas de Aprendizado Não-Supervisionado

As ~~Algoritmos~~ Técnicas de aprendizado não-supervisionado são divididos em três classes: agrupamento, associação e redução de dimensionalidade. Nas próximas seções serão apresentados os detalhes de cada classe.

02.1 - ~~Algoritmos~~ Técnicas de Agrupamento

As técnicas de agrupamento de dados visam agrupar em clusters conjuntos de pontos que são similares entre si e que diferem de outro conjunto de pontos. Na imagem abaixo é apresentado um exemplo de agrupamento em cluster:



As técnicas de agrupamento são divididas em agrupamentos exclusivos, sobrepostos, hierárquicos e probabilísticos. Nos agrupamentos exclusivos, um ponto (dado) pode pertencer só e somente só a um único cluster. O algoritmo frequentemente utilizado é o K-means, onde inicialmente são criados K clusters aleatórios, o centroide médio de cada cluster é calculado e os clusters então são atualizados através de uma função de distância, geral

MENTE A DISTÂNCIA EUCLIDIANA OU A DISTÂNCIA MANHATTAN, ENTRE A DISTÂNCIA DE CADA PONTO PARA O ~~MEDIA~~ CENTRÓIDE MÉDIO DE CADA CLUSTER. OS AGRUPAMENTOS SOBREPOSTOS SE DIFEREM DOS EXCLUSIVOS POIS PERMITE QUE UM MESMO PONTO PERTENÇA A MAIS DE UM CLUSTER COM NÍVEIS DE PERTENCIMENTO DISTINTOS.

OS AGRUPAMENTOS HIERÁRQUICOS ATUALIZAM OS PONTOS ENTRE OS CLUSTERS ATRAVÉS DE REGRAS DE LIGAÇÃO ENTRE ~~OS PONTOS~~ CADA DOIS PONTOS DO CONJUNTO DE DADOS. AS LIGAÇÕES PRINCIPAIS SÃO:

- LIGAÇÃO MÉDIA: É A MÉDIA DA DISTÂNCIA ENTRE DOIS PONTOS
- LIGAÇÃO MÁXIMA: É A DISTÂNCIA MÁXIMA ENTRE DOIS PONTOS
- LIGAÇÃO MÍNIMA: É A DISTÂNCIA MÍNIMA ENTRE DOIS PONTOS

POR FIM, OS AGRUPAMENTOS PROBABILÍSTICOS ATUALIZAM OS PONTOS ENTRE OS CLUSTERS ATRAVÉS DA PROBABILIDADE DE UM DETERMINADO DADO PERTENCER A UM CLUSTE. NORMALMENTE A DISTRIBUIÇÃO GAUSSIANA É O MÉTODO UTILIZADO PARA NORMALIZAR OS DADOS.

02.2 - TÉCNICAS DE ASSOCIAÇÃO

AS TÉCNICAS DE ASSOCIAÇÃO LIGAM CONJUNTOS DE DADOS QUE ~~SÃO~~ POSSUEM UMA FREQUÊNCIA DE USO EM CONJUNTO. ESSAS TÉCNICAS SÃO PARTICULARMENTE VALIOSAS NO RAMO DE SUGESTÕES PARA USUÁRIOS TANTO EM SITES DE E-COMMERCE QUANTO STREAMING, OU SEJA, O PRÓPRIO USUÁRIO, ATRAVÉS DE SUAS COMPRAS OU REPRODUÇÕES AJUDAM A CONSTRUIR UM MODELO

DE IA QUE DARÁ RECOMENDAÇÕES PARA TODOS OS USUÁRIOS DA PLATAFORMA. SÃO EXEMPLOS DE RECOMENDAÇÕES O "QUEM COMPROU X TAMBÉM COMPRA Y" DA AMAZON E TAMBÉM AS PLAYLISTS "RECOMENDED FOR YOU" DO SPOTIFY.

O ALGORITMO MAIS UTILIZADO NAS TÉCNICAS DE ASSOCIAÇÃO É O APriori. ESTE PARTE DO PRINCÍPIO QUE SE UM ITEM É FREQUENTE, TODOS OS SEUS SUBITENS TAMBÉM SÃO FREQUENTES E ASSIM O MODELO É CONSTRUÍDO E É ATUALIZADO.

02.3 - TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE

AS TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE SÃO PARTICULARMENTE ÚTEIS EM CASOS ONDE A BASE DADOS É MUITO GRANDE OU REDUNDANTE. ESSAS TÉCNICAS SÃO FREQUENTEMENTE UTILIZADAS EM PRÉ-PROCESSAMENTO DE DADOS PARA UTILIZAÇÃO EM OUTROS MODELOS DE APRENDIZAGEM OU ATÉ MESMO OUTRAS APLICAÇÕES. UM EXEMPLO COTIDIANO ONDE SÃO APLICADAS TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE SÃO OS "TRENDING TOPICS" DO SITE X (ANTIGO TWITER). CADA TWEET É UM CONJUNTO DE DADOS ~~E JUNTAMENTE COM TE~~ DE UMA VASTA BASE DE USUÁRIOS QUE, SEM A APLICAÇÃO DA REDUÇÃO DE DIMENSIONALIDADE, É IMPRATICÁVEL TRATAR OS DADOS. JUNTAMENTE COM TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL, AS PALAVRAS MAIS FALADAS DO MOMENTO CONSTITUEM O "TRENDING TOPICS".

O ALGORITMO MAIS CONHECIDO É O DE ANÁLISE DOS COMPONENTES PRINCIPAIS (PCA). NELE, CADA COMPONEN-

TE EXTRAÍDA DO CONJUNTO DE DADOS REPRESENTAM A VARIAÇÃO MÁXIMA DOS DADOS. CADA COMPONENTE É COMPLETAMENTE DIFERENTE UMA DAS OUTRAS.

03 - Considerações Finais

NESTE TEXTO ~~FORAM~~ FOI APRESENTADO O APRENDIZADO DE MÁQUINA NÃO-SUPERVISIONADO E SUAS PRINCIPAIS TÉCNICAS. QUANDO A MARCAÇÃO DE UMA BASE DE DADOS É CUSTOSA, AS TÉCNICAS NÃO-SUPERVISIONADAS PODEM SER DE GRANDE UTILIDADE, INCLUSIVE SERVINDO DE PRÉ-PROCESSAMENTO DE DADOS PARA OUTRAS APLICAÇÕES DE INTELIGÊNCIA ARTIFICIAL, POR EXEMPLO, AS REDES NEURAIS.

AS TÉCNICAS NÃO-SUPERVISIONADAS SÃO DE GRANDE VALOR PARA EMPRESAS DE E-COMMERCE E STREAMING POIS AS PREFERÊNCIAS DE COMPRAS E GOSTOS DOS USUÁRIOS MUDAM CONSTANTEMENTE, DIFERENTE DE UMA BASE DE DADOS FIXAS DE IMAGENS DE MAÇÃS E BANANAS.

É FUNDAMENTAL ~~TEM~~ SER CRITERIOSO COM AS CONDIÇÕES DE PARADA DOS ALGORITMOS DE APRENDIZADO NÃO-SUPERVISIONADO, UMA VEZ QUE NESSE MODELO NÃO É POSSÍVEL APLICAR AS MÉTRICAS DE AVALIAÇÃO DE DADOS CONVENCIONAIS, COMO PRECISION, RECALL OU F1-SCORE.

O APRENDIZADO DE MÁQUINA NÃO-SUPERVISIONADO POSSUI UMA VASTA GAMA DE APLICAÇÕES. QUANDO BEM UTILIZADO É DE GRANDE VALIA NO ESPECTRO DA INTELIGÊNCIA ARTIFICIAL.