UFU CAMPUS MONTE CARMELO

Mineração de Dados

Trabalho 1

Marcus Antonio Rufino Ribeiro
31511BSI023

Monte Carmelo

1 Introdução.

A minha base de dados foi a Adult, que consite em dizer se a pessoa vai receber <=50k por ano .Os tratamentos dos dados foi pegar a media para atributos ordinais e pegar a moda para atributos nominas, o algoritmo foi coeficiente de jaccard.

Numero total de instâncias: 32.561, sendo 7.841 casos negativos e 24.720

positivos. Serão 3.256 duplas de teste e 29.305 de treino.

Número de variáveis: 14 mais a classe.

As variáveis consistem em:

- age
- workclass
- fnlwgt
- education:
- education-num
- marital-status
- occupation
- relationship
- race
- sex
- capital-gain
- capital-loss
- hours-per-week
- native-country

O algoritmo de classificação KNN é o algoritmo usado, sendo K =

3 e K = 57, isso significa que pegaremos os 3 ou os 57 vizinhos mais

próximos, e olhando a classe deles, pegando a classe que mais se

repetiu entre eles, ou seja, a moda.

2 Desenvolvimento

Após abrir a planilha o primeiro passo foi normalizar a tabela com os dados que estava faltando.

Após normalizada foi salva a nova planilha

Depois foi considerado as primeiras 90%linhas como minha base de teste e os 10% linha final da tabela considerada como base de treino.

Execução:

Para cada linha de treino eu percorro a minha base de teste por completa e vejo quais as 3 primeiras linhas ou as 57 linhas tem a maior similaridade com minha linha da base de treino e pego a moda da coluna 15 que e aonde indica o salario se vai ser <=50k ou >50K e jogo o valor da moda na minha linha da base de treino que foi trabalhada.

Resultados obtidos em termos de acurácia KNN=3:

		Resultado obtido	Resultado obtido
		<=50K	>50K
Resultado esperado	<=50K	1634	251
Resultado esperado	>50K	673	698

Total de acertos:2.332

Total de erros:924

Acurácia :73,1%

Erro:26,8%

precisão :70,8%

sensitividade:70%

Resultados obtidos em termos de acurácia KNN=57:

		Resultado obtido	Resultado obtido
		<=50K	>50K
Resultado esperado	<=50K	1720	165
Resultado esperado	>50K	708	663

Total de acertos:2.383

Total de erros:873

Acurácia :71%

Erro:28%

precisão :70,8%

sensitividade:72,1%

A precisão se manteve a mesma nos 2 casos, e com o valor do knn alto a tendência e que a pessoa receba o salario “<=50K” pois ele é a moda.

3 conclusão:

Como não e um algoritmo para a área da saúde, não vai arriscar a vida de ninguém a acurácia maior que 70% eu considero ainda como bom pois a cada 10 pessoas 7 ele esta correto.

O que me chamou atenção foi a linguagem R que eu não tinha muito pratica e nem conhecimento com a linguagem ,e outro fato foi que o como tinha muitos dados o algoritmo demorou mais de 12 horas para se executar por completo.