Universidade Federal de Alagoas

Instituto de Computação

Ciência de Dados

Professor: Bruno Pimentel

Aluno: Maxwell Esdra Acioli Silva

Lista de Exercícios 1

1. Qual a diferença entre Big Data e Ciência de Dados? (0,5 ponto)

Big Data pode ser definido como um grande volume de dados que são gerados com uma alta velocidade e de formas variadas, também precisam de técnicas inovadoras para serem processados e armazenados, de tal forma que possibilitem uma melhor compreensão no processo de tomada de decisão ou automação de processos. Já Ciência de dados é responsável pela criação de soluções para modelagem destes dados, a fim de extrair insights de dados, ou seja, descobrir coisas novas nos dados, e como consequência o conhecimento adquirido possa auxiliar na tomada de decisão.

2. De que forma Estatística, Mineração de Dados e Aprendizagem de Máquina interagem com Ciência de Dados? (1 ponto)

Estatística se relaciona com ciência de dados através de fornecimento de métodos estatísticos que possibilitam entender melhor os dados, a validação dos mesmos, bem como avaliar os modelos de aprendizagem de máquina. Mineração de dados por sua vez, se alia a ciência de dados através do pré-processamento dos dados a partir da aquisição destes, ou seja, fazer a seleção de atributos e remoção de outliers ou ruídos, para que após este tratamento os dados fiquem prontos para serem processados pelos algoritmos de aprendizagem de máquina. Aprendizagem de máquina se alia a ciência de dados através de possibilitar automação do processo de análise dos dados que estão sendo processados, este processamento são realizados através algoritmos definidos pela inteligência artificial.

3. Mostre a importância do conhecimento de domínio para o cientista de dados. (0,5 ponto)

Um cientista de dados é responsável por fazer a análise de um grande volume de dados, e através de insights deve extrair algum conhecimento que possa auxiliar no processo

de tomada de decisão. Porém, para fazer esta análise, este deve possuir o domínio de algumas ferramentas que possam auxiliar em cada etapa da análise de dados. Por exemplo, precisa ter um bom domínio de estatística, porque esta fornece métodos que ajudam a fazer a análise descritiva dos dados ou realizar testes de hipóteses para validar alguma inferência. Outro domínio importante que o cientista deve ter é sobre a mineração de dados, através desta ele pode fazer o pré-processamento dos dados obtidos, e preparar os dados para sem processados por um modelo de aprendizagem de máquina. A programação é outro domínio que o cientista deve ter, porque sem ela não é possível utilizar o poder que grandes linguagens de programação tem de facilitar as etapas do processo de análise de dados, como Python ou R. Ele também deve ter um espírito investigativo, por através dessa mentalidade será capaz de explorar ao máximo todas as informações que podem ser obtidas em um volume de dados.

4. Crie um conjunto de dados com duas variáveis V1 e V2, tal que:

Para gerar os dados foi criada uma função que gera aleatoriamente um conjunto de dados contendo 100 elementos que variam entre 0 e 1. Segue abaixo um exemplo de conjunto gerado:

a. Mediana de V1 < Média de V1 (0,5 ponto)

v1 median is: 0.5026226956692676

v1 mean is: 0.5090271749126098

[0.5315860518335002, 0.5468301823251933, 0.4361104043793795, 0.4351534602339402, 0.4092553775197598, 0.6055087421447591, 0.5828134843038235, 0.47663050031048004, 0.48954191868579655, 0.545408990952919, 0.48064815139393885, 0.48806389032665465, 0.6730131490846069, 0.5694868349670904, 0.5783971331658174, 0.48168971039072733, 0.6557508368837383, 0.5067434700204975, 0.518599612400269, 0.5199419993819814, 0.3522884942835687, 0.4822223712625863, 0.49740899549841106, 0.45051026267168204, 0.47538833722349594, 0.5491209521345366, 0.4187400844913074, 0.5750232152226209, 0.5196321756317261, 0.5131276821442755, 0.5951778742895261, 0.388335594726394, 0.5771615033831761, 0.5005968833535535, 0.43762979942100527, 0.468620049306495, 0.5985385667946831, 0.705340112502919, 0.435256631548799, 0.4756192623061669, 0.6355680460322788, 0.4427730681216472, 0.46826772185404153, 0.5221753577188687, 0.638776147127259, 0.28210709893619046, 0.25539251851502914, 0.6528524753905321, 0.5308408355888377, 0.37253090617834894, 0.6632678673776417, 0.6233015028630582, 0.42117610649712733, 0.4876982429614586, 0.5986365530500004, 0.5828814797488437, 0.5511884798187517, 0.5055600092881023, 0.6398222504540663, 0.5198264178928542, 0.41857711506572864, 0.442104217644054, 0.538998456737936, 0.6086652964889893, 0.38706588001042763, 0.7501861239774834, 0.39044599022348575, 0.49053314188969116, 0.47473793458389846, 0.39145640418703637, 0.5961979849223477, 0.4474645810946833, 0.5019483752016238, 0.46422067592880367, 0.59633533508743, 0.5800043917924403, 0.4476681020477126, 0.36669004258139837, 0.3248608054312757, 0.4816480728855277, 0.43952146445911094, 0.5777930848598974, 0.44175570713106016, 0.4126942825688532, 0.6192664514696606, 0.33778915299255774, 0.5176618883517484, 0.5526409318376756, 0.5163482277002546, 0.6855573878776782, 0.7455877063133654, 0.49782003913550094, 0.5425267904380824, 0.4532262989497247, 0.4440377678069622, 0.6289120694396363, 0.4865942053041368, 0.3508910663698762, 0.5074311900195843, 0.5032970161369114]

b. Mediana de V2 > Média de V2 (0,5 ponto)

v2 median is: 0.5317107768605931

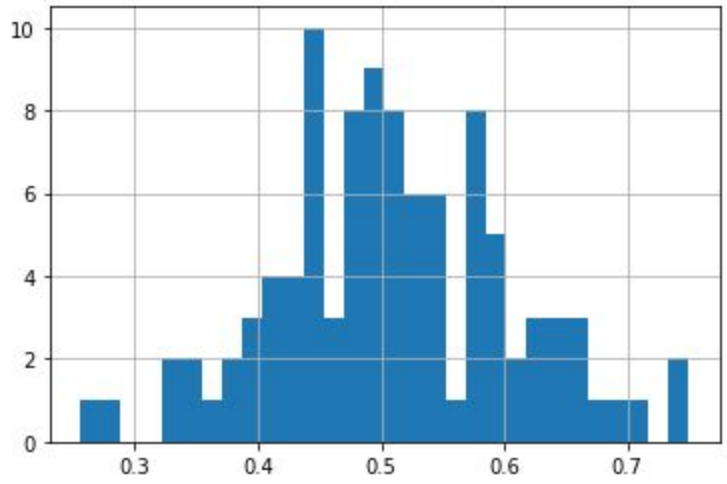
v2 mean is: 0.5089958779577891

[0.8177215185002511, 0.4331977231891503, 0.568138390239395, 0.7655021247050549, 0.887963798413877, 0.4264000415289405, 0.2905903648711837, 0.17830782339731688, 0.7082045898550351, 0.9255542272875457, 0.6125226694139385, 0.6730227018441015, 0.2913046161921785, 0.09024073680911471, 0.17311755272001406, 0.6815131906476147, 0.026561320678148648, 0.07314579903723772, 0.8464232153170624, 0.8198078165328189, 0.2884548634584546, 0.9053001090384263, 0.1609692595715858, 0.8346198976932536, 0.7760749349202869, 0.6414978290136161, 0.874750185512281, 0.3819045430172918, 0.060900674764950646, 0.34954507332167317, 0.7484931162298714, 0.4743700184996631, 0.41709111143237, 0.8570632436680583, 0.7329655049419109, 0.7718156195899986, 0.40183174179049785, 0.815768731023719, 0.2599213511574142, 0.5470530365444487, 0.6516473379345962, 0.6771640075068184, 0.4785014847166811, 0.1676619848038402, 0.9371504850641957, 0.7721953624453662, 0.11297880072745137, 0.7597404826952121, 0.3150523213512738, 0.5539586881073535, 0.6225956404912337, 0.5672019226951867, 0.7141569410013209, 0.4473116323144729, 0.24726105071162607, 0.29293417931859655, 0.841071048673258, 0.014279679198782147, 0.4193672323017338, 0.5163685171767376, 0.33578180015138914, 0.7761047389451391, 0.10061607912786363, 0.6129075255772941, 0.10261048138732098, 0.0637027130470379, 0.7798021804249154, 0.4279337636573627, 0.9855670508386448, 0.4203289053135886, 0.02942732644453705, 0.28529904250162075, 0.7598220212674887, 0.7528632849919058, 0.5808663267937175, 0.4709376815867089, 0.1075558115854488, 0.9485012570110032, 0.076060283399625, 0.2721986665406281, 0.9757190977729137, 0.7851249574280188, 0.6690140242020935, 0.6118934240824425, 0.38763582023504906, 0.884722333082451, 0.2592193382716518, 0.8860199182242137, 0.31184756014588744, 0.8037187972504206, 0.5969264501718318, 0.08448845549982997, 0.2649775015757234, 0.6278523917209511, 0.37606178966773707, 0.19505294359282443, 0.20663119027447885, 0.1480900434247645, 0.8506288617533146, 0.3868460912035827]

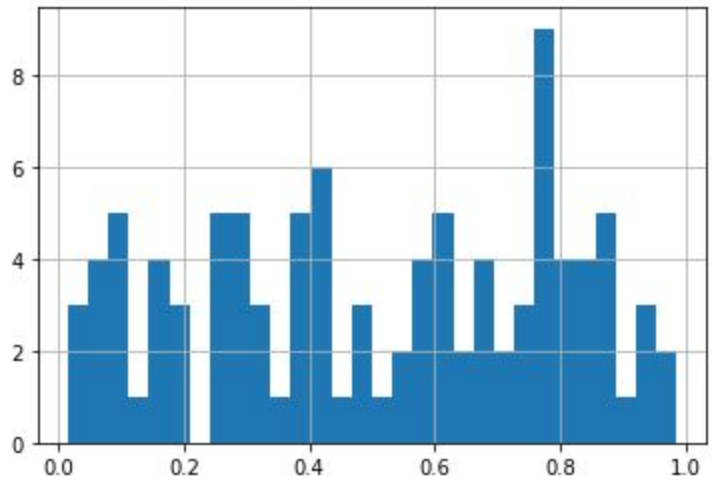
5. Baseando-se no conjunto de dados criado na questão 4, crie uma função em Python que:

a. Mostra o histograma de cada variável; (1 ponto)

Histograma v1:



Histograma v2:



b. Verifica se as variáveis seguem uma distribuição Normal (use teste de hipótese) (1 ponto)

O código utilizado para realizar o teste de hipótese está anexado no código Lista_1.py

a. Verificação da normalidade do conjunto v1:

Pvalue = 0.9454444646835327

SignificantLevel = 0.05

The distribution is normal

b. Verificação da normalidade do conjunto v2:

Pvalue = 0.0005203643231652677

SignificantLevel = 0.05

The distribution is not normal

6. Cite 2 técnicas para remoção de ruídos e, para cada uma, mostre uma vantagem e uma desvantagem. (1 ponto)

Regression é uma técnica que consiste de encontrar uma função linear que representa a distribuição dos dados, ou seja, definir uma função onde podemos fazer o mapeamento de uma variável para outra. Se tratando de detecção de ruídos, quando definimos uma linha de regressão que representa a distribuição dos dados, os pontos que ficam muito distantes da linha são considerados dados ruidosos. Uma vantagem da utilização desta técnica é que ela manipula bem dados cuja distribuição é considerada inclinada. Por outro lado, podemos citar como uma desvantagem desta o fato de ter um custo computacional muito alto quando os dados contêm um número elevado de dimensões.

Uma segunda técnica é a de análise de outliers baseada em agrupamentos, esta pode ser definida como uma técnica que tem o objetivo de organizar os dados em grupos, estes grupos são definidos a partir de alguma métrica que define uma representatividade para o mesmo. Quando alguns dados não pertencem a nenhum dos grupos pré definidos estes podem ser considerados outliers ou dados ruidosos. Podemos citar como vantagem deste tipo de técnica é a velocidade na detecção de outliers, pois uma vez definidos os grupos, há apenas a necessidade de comprar o objeto em questão com os grupos existentes para definir se ele é ou não um dado ruidoso, isto acontece de forma muito rápida porque o número de grupos é muito menor comparado ao número total de dados. Porém, podemos definir como uma desvantagem desta técnica o alto custo computacional para definição dos clusters, bem como definir qual métrica será utilizada para estabelecer os grupos.

7. Qual é a importância de utilizar as seguintes abordagens de redução de dados no contexto de Ciência dos Dados?

a. Redução de dimensionalidade (0,5 ponto)

A redução de dimensionalidade é importante, pois através dela é possível eliminar atributos que não são relevantes numa modelagem, e que se não forem retiradas podem levar o modelo a fazer uma análise errada.

b. Redução de numerosidade (0,5 ponto)

Já a redução de numerosidade é importante para detectar um grupo com maior representatividade dos dados, ou seja, este pode representar um grupo original com um volume muito grande de instâncias com uma menor quantidade, isso pode resultar num ganho em performance na análise dos dados pois podemos reduzir drasticamente o número de instâncias a serem analisadas.

8. De que forma pode-se detectar overfitting em um classificador? (0,5 ponto)

O overfitting pode ser identificado verificando métricas de validação, que geralmente aumentam até um ponto em que elas estagnam ou começam a declinar quando o modelo é afetado pelo ajuste excessivo.

Outra forma de detectar um overfitting é verificar se o modelo tem uma performance muito discrepante nos dados de treinamento do que nos dados de teste.

9. Em quais tipos de problemas é preferível utilizar leave-one-out a utilizar K-fold cross-validation? (0,5 ponto)

A utilização do K-fold cross-validation não é recomendável em um conjunto de testes pequeno, ele deve ser utilizado quando queremos testar todo o nosso conjunto de dados. Já o leave-one-out serve para ser testado em um conjunto de dados pequeno, além disso por não ter o fator de aleatoriedade da seleção das folds ele é considerado mais confiável.

10. Crie um script em Python que avalie a diferença de desempenho do classificador K-NN para o conjunto de dados Iris (https://archive.ics.uci.edu/ml/datasets/iris). Use F-measure e K-fold cross-validation. (2 pontos)

O script criado pode ser encontrado no arquivo Lista_1.py anexado. Para fazer a avaliação do classificador foram feitos testes para ver qual o número de vizinhos mais próximos que tornavam o resultado do mesmo mais adequado, nos experimentos realizados o valor mais adequado foi o de n igual a 3. Com relação a utilização da técnica K-fold cross-validation, para definir o número de folds foi utilizada como referência o valor estabelecido na maioria das literaturas, ou seja, o valor 10. Sobre o uso da métrica F-measure, onde foi utilizada a implementação definida na biblioteca sklearn, foram feitos alguns testes sobre o parâmetro “*average*” para ver qual valor retornava o melhor desempenho, após vários

testes o valor escolhido para parâmetro foi “*weighted*”, este parâmetro calcula a métrica para cada label utilizando uma média ponderada. Para obter o desempenho do classificador utilizando as parametrizações citadas anteriormente, foi feita uma média da métrica f-measure das iterações sobre as 10 folds pré definidas, e o resultado final foi um desempenho de 97% para a métrica utilizada.