Introdução

A correlação de Pearson é uma estatística que mede a correlação linear entre duas variáveis aleatórias X e Y . Por exemplo, sabemos que as notas de física e matemática tem correlação: um aluno com boas notas em matemática geralmente tem também boas notas em física e vice-versa. Outro exemplo é a relação entre a idade e altura (até a puberdade). O coeficiente da correlação de Pearson pode assumir valores entre -1 e 1. O valor de +1 representa uma correlação linear positiva (quando X ou Y cresce, a outra variável também cresce linearmente), 0 representa ausência de correlação linear e -1 representa uma correlação linear negativa (quando X ou Y cresce, a outra variável decresce linearmente).

Este método foi desenvolvido por Karl Pearson, baseado nas ideias de Francis Galton de 1880.

Antes de descrevermos o cálculo da correlação de Pearson, é necessário introduzir alguns conceitos básicos de estatística: a covariância e o desvio padrão.

A covariância é uma medida de variabilidade conjunta de duas variáveis aleatórias X e Y . Sua interpretação é semelhante à da correlação de Pearson. No entanto, sua magnitude não é de fácil interpretação, pois ela não é normalizada pelas magnitudes das variáveis. Sejam (X ,Y) o par de variáveis aleatórias com valores (xi , yi) para i=1,2,…, n , ´x= ∑ i=1 n xi n e ´y= ∑ i=1 n yi n as médias dos valores observados de X e Y , respectivamente, então a covariância entre X e Y é dada por: cov (X ,Y)= 1 n−1∑ i=1 n (xi−´x)(yi−´y) .

O desvio padrão é uma medida que quantifica a variação de dispersão de um conjunto de valores. Um baixo desvio padrão indica que os valores estão próximos da média, enquanto valores altos indicam valores mais espalhados. Sejam x1 , x2 ,…, xn os valores observados e ´x= ∑ i=1 n xi n a média dos valores observados, o desvio padrão de X é dado por: s X=√ 1 n−1∑ i=1 n (xi−´x) 2 .

Finalmente, a correlação de Pearson é a covariância normalizada pelo desvio padrão, i.e., Pearson(X ,Y)= cov (X ,Y) sX sY .

Uma segunda medida de associação entre variáveis é a correlação de Spearman. Ela nada mais é que o cálculo da correlação de Pearson nos postos dos dados. Em outras palavras, Sp ear man (X ,Y)=Pearson (posto(X), posto(Y)) . A medida de correlação de Spearman é capaz de identificar associações não-lineares monotônicas como relações sigmoides e logaritmicas.

Para o cálculo do posto, existem diversas formas. Uma das mais usadas é a do fracionamento (fractional ranking). Neste método, números iguais recebem a média do valor no caso do posto usual. Vejamos um exemplo. Para o conjunto de dados: (1, 1, 2, 3, 3, 4, 5, 5, 5), o posto usual seria (1, 2, 3, 4, 5, 6, 7, 8, 9). Agora no caso do fractional ranking, os postos são (1.5, 1.5, 3, 4.5 4.5, 6, 8, 8, 8). Para o cálculo do coeficiente de Spearman, usaremos o fractional ranking.

Tarefa: Você deve implementar as seguintes funções abaixo:

recebe uma lista x e devolve a média dos valores de x. def media(x):

recebe uma lista x e devolve o desvio padrão de x. Você deve usar a função media aqui. def desvpad(x):

recebe duas listas x e y e devolve a covariância entre x e y. Você deve usar a função media aqui. def cov(x,y):

recebe duas listas x e y e devolve o coeficiente de correlação de Pearson entre x e y. Você deve usar as funções cov e desvpad aqui. def pearson(x,y):

recebe uma lista x e devolve o fractional ranking de x. Para o cálculo do posto, utilize o algoritmo de ordenação por inserção visto em aula. def posto(x):

recebe duas listas x e y de devolve o coeficiente de correlação de Spearman entre x e y. Você deve usar a função pearson aqui. def spearman(x,y):

recebe o nome de um arquivo texto (ex: in01.txt) contendo os dados das variáveis aleatórias e devolve a matriz de covariância dos dados. O valor de cada posição (i , j) da matriz é dado pela covariância dois a dois entre as variáveis das colunas i e j do arquivo de entrada. Você deve usar a função cov aqui. def matriz_covariancia(nome_arquivo):

O arquivo texto de entrada corresponde a uma planilha no formato CSV, tal que os dados são separados por ponto-e-vírgula da seguinte forma:

1. A primeira linha deve ter os nomes das variáveis

2. A partir da segunda linha estão os valores correspondentes a variável

Exemplo de arquivo de entrada (in01.txt):

altura;peso;salario;QI

175;74;16500;151

160;48;4000;140

178;82;18800;147

160;52;2300;137

168;72;5750;124

As funções main(), imprime_lista(L)e imprime_matriz(M) já estão sendo fornecidas prontas e não devem ser alteradas. A função principal possui 7 diferentes modos de operação, visando testar cada uma das sete funções solicitadas. Exemplos de entrada e saída para cada modo de operação do programa são apresentados abaixo. As entradas fornecidas estão em azul e as saídas esperadas em vermelho.

Exemplo 1:

```
Digite modo do programa: 1
Digite n: 5
Digite x1: 175
Digite x2: 160
Digite x3: 178
Digite x4: 160
Digite x5: 168
media: 168.2000
```

Exemplo 2:

```
Digite modo do programa: 2
Digite n: 5
Digite x1: 74
Digite x2: 48
Digite x3: 82
Digite x4: 52
Digite x5: 72
desvpad: 14.7919
```

Exemplo 3:

```
Digite modo do programa: 3
Digite n: 5
Digite x1: 16500
Digite x2: 4000
Digite x3: 18800
Digite x4: 2300
Digite x5: 5750
Digite y1: 151
Digite y2: 140
Digite y3: 147
Digite y4: 137
Digite y5: 124
cov: 55917.5000
```

Exemplo 4:

```
Digite modo do programa: 4
Digite n: 5
Digite x1: 16500
Digite x2: 4000
Digite x3: 18800
Digite x4: 2300
Digite x5: 5750
Digite y1: 151
Digite y2: 140
Digite y3: 147
Digite y4: 137
Digite y5: 124
pearson: 0.7048
```

Exemplo 5:

```
Digite modo do programa: 5
Digite n: 8
Digite x1: 159
Digite x2: 160
Digite x3: 159
Digite x4: 178
Digite x5: 160
Digite x6: 168
Digite x7: 160
Digite x8: 160
posto: 1.5000 4.5000 1.5000 8.0000 4.5000 7.0000 4.5000 4.5000
```

Exemplo 6:

```
Digite modo do programa: 6
Digite n: 5
Digite x1: 16500
Digite x2: 4000
Digite x3: 18800
Digite x4: 2300
Digite x5: 5750
Digite y1: 151
Digite y2: 140
Digite y3: 147
Digite y4: 137
Digite y5: 124
spearman: 0.6000
```

Exemplo 7:

```
Digite modo do programa: 7
Digite o nome do arquivo: in01.txt
       69.2000       118.1000     60907.5000        42.8000
      118.1000       218.8000     95510.0000        36.4000
    60907.5000     95510.0000  57909500.0000     55917.5000
       42.8000        36.4000     55917.5000       108.7000
```