UFPA PPGCC: Aprendizado de Máquina
Lista de exercício #1

1. (1.0 pt) Os dados abaixo se referem a taxas de colesterol total (mg/100ml) de 30 indivíduos. Utilize duas casas decimais para o cálculo.

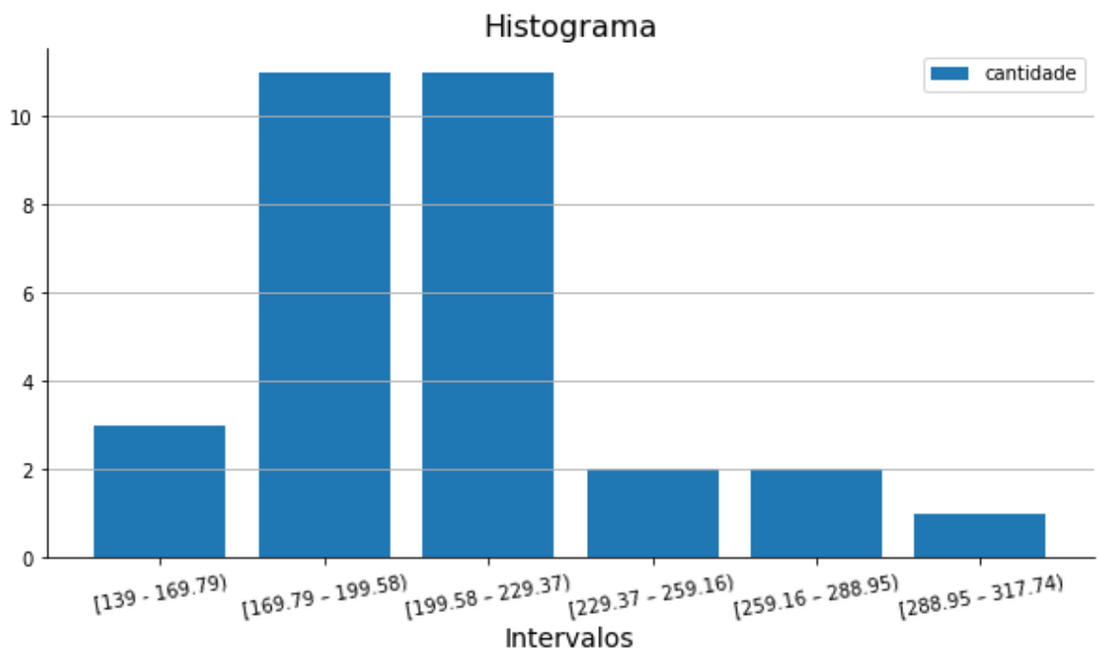
140	160	168	180	180	180	180	184	185	190
190	192	192	196	200	200	200	205	205	208
214	214	220	220	225	230	240	260	280	315

a) Montar uma tabela de distribuição de frequência por intervalo para as taxas (utilize a regra de Sturges para calcular o número de classes – intervalos).

Número de clases k = 1 + 3.3*log(30) = 6
Amplitud L = 315 – 140 = 175
Largura da classe h = L/k = 29.79

Classe	Intervalo
1	**[**139 – 169,9**)**
2	**[**169,79 – 199,58**)**
3	**[**199,58 – 229,37**)**
4	**[**229,37 – 259,16**)**
5	**[**259,16 – 288,95**)**
6	**[**288,95 – 317,74**)**

b) Calcule o histograma



c) Calcule as frequências relativas, as frequências acumuladas absolutas e relativas e os pontos médios para todas as classes.

Classe	Frequência absoluta	Frequência relativa	Frequência acumulada absoluta	Frequência acumulada relativa	Pontos médios
1	3	0,1	3	0,1	154,395
2	11	0,366	14	0,4666	184,685
3	11	0,366	25	0,4666	214,475
4	2	0,066	27	0,9	244,265
5	2	0,066	29	0,9666	274,055
6	1	0,0333	30	1	303,345

d) Calcule a taxa de colesterol média

$$média = \sum_{i=1}^{30} x_i = 205.1$$

e) Calcule a taxa de colesterol mediana

$$mediana = \frac{200+200}{2} = 200$$

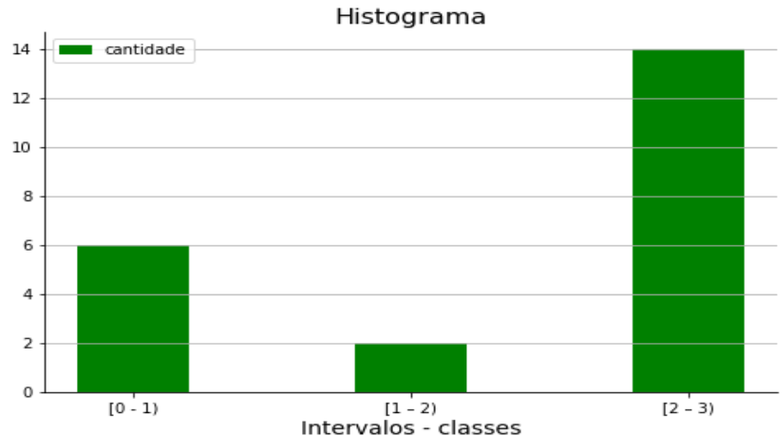
f) Calcule a variância e o desvio padrão amostral

$$variância = \sigma^2 = \sum_{i=1}^{N} \left(\frac{(x_i - \mu)^2}{N} \right) = 1184{,}1566$$

$$desvio\ padrão\ amostral = \sqrt{\sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})^2}{n-1} \right)} = 34{,}9998$$

2. (1.5 pt) Considere que os valores assumidos por um dado atributo numérico são listados no vetor x = {1, 3, 2, 3, 2, 2, 0, 1, 0, 0, 3, 0, 2, 3, 2, 2, 3, 3, 0, 3, 2, 0} .

a) Calcule o histograma de x (utilize o bom senso para definir o número de classes).



b) Supondo que tais valores correspondem aos assumidos em um experimento por uma variável aleatória X, estime sua média E[X] = μ, E[X²], variância σ_x^2, o desvio padrão σ_x e o desvio médio absoluto.

$$E[x]=\sum_{i=1}^{n} p_i x_i$$

$$E[x]=(0{,}2727)*0+(0{,}0909)*1+(0{,}3181)*2+(0{,}3181)*3=1{,}6818$$

$$E[x^2]=(0{,}2727)*0^2+(0{,}0909)*1^2+(0{,}3181)*2^2+(0{,}3181)*3^2=4{,}2272$$

$$\sigma^2=E[x^2]-E[x]^2=4{,}2272-1{,}6818^2=1{,}3987$$

$$\sigma=\sqrt{\sigma^2}=\sqrt{1{,}3987}=1{,}1826$$

$$Desvio\,médio\,absoluto=D_m=\sum_{i=1}^{N}\left(\frac{|x_i-\bar{x}|}{N}\right)$$

$$D_m=1{,}0413$$

c) X é uma variável aleatórioa ou contínua?

Eu acho que X é contínuo, já que as possibilidades de cada número são diferentes, eu acho que elas deveriam ser mais semelhantes.

3. (2.0 pt) Use um editor de texto ASCII para verificar o conteúdo do arquivo iris.arff (o qual vem com Weka). Estude-o também usando a GUI chamada Explorer do pacote Weka. Copie a iris.arff para um novo arquivo chamado iris.csv, elimine o header (primeirais linhas, antes de @data), e leia o arquivo iris.cvs no Excel. Escreva código em Java ou outra linguagem de sua preferência para calcular a variância do terceiro parâmetro (terceiro elemento de x) a partir da leitura do arquivo iris.csv. Compare o resultado com as variâncias estimadas pelos programas Weka e Excel. Inclua a listagem de seu código.

Os cálculos foram arredondados para quatro casas decimais.

Variância	Weka	Excel	Var Python	Var amostral Python
	3,1117	3,1132	3,0924	3,1132

```
1 df = pd.read_csv("./iris.csv", names=["sepallength","sepalwidth","petallength","petalwidth", "CLASSE"])
2 print("varianza por Python: ", round(np.var(df["petallength"]), 4))
3 print("varianza Amostral por Python: ", round(np.var(df["petallength"], ddof=1), 4))
4 df.head()
```

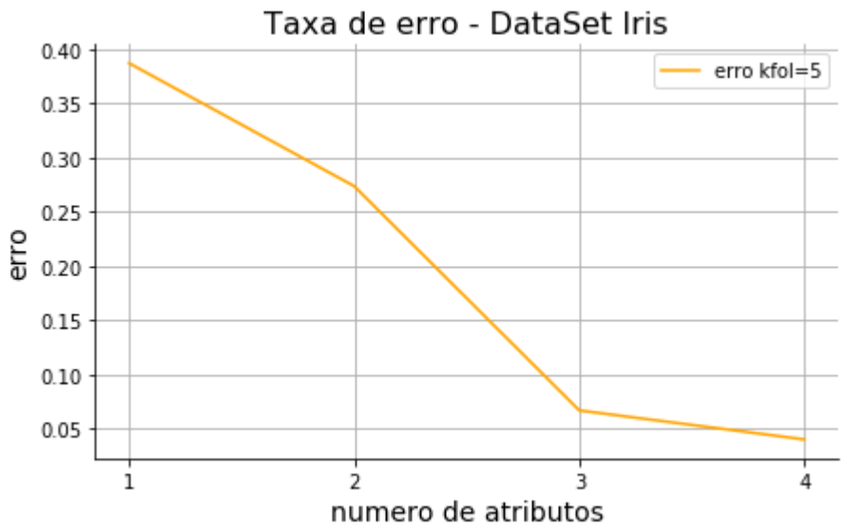
```
varianza por Python:  3.0924
varianza Amostral por Python:  3.1132
```

	sepallength	sepalwidth	petallength	petalwidth	CLASSE
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

4. (2.5 pt) O Coeficiente de variação (CV) é uma medida relativa de variabilidade que independe da unidade de medida utilizada CV = (Desvio p adrao/M edia). É possível utilizar o CV para selecionar os ”melhores” atributos, ou seja, aqueles que contenham os menores valores de CV. Selecione duas bases de dados do UCI e construa um gráfico (Taxa de erro versus conjunto de atributos) para cada base. Utilize o classificador 1-NN para estimar a taxa de erro. Os conjuntos de atributos serão formados da seguinte maneira: inicialmente o conjunto irá conter o atributo com o menor CV: no passo seguinte o conjunto irá conter os dois atributos com os menores CVs; e assim por diante até que o conjunto final seja formado por todas os atributos.

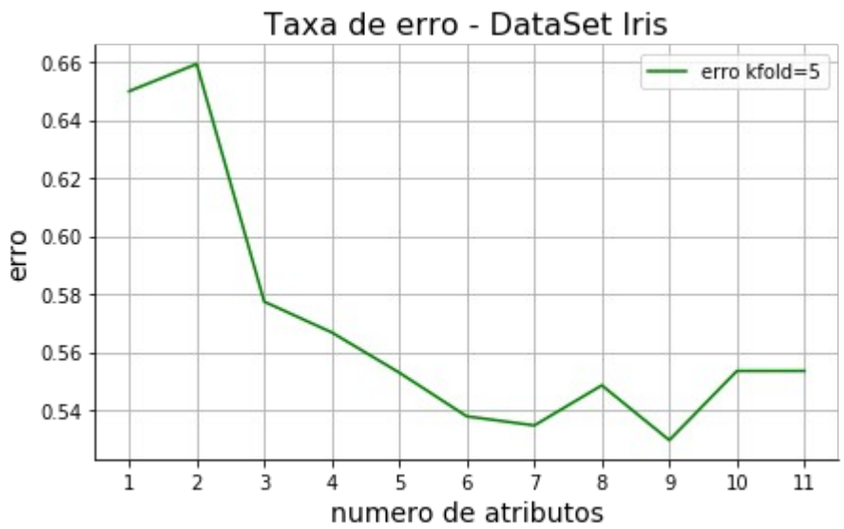
DataSet Iris:

- CV sepallength: 0,1412
- CV sepalwidth: 0,1415
- CV petallength: 0,4678
- CV petalwidth: 0,6345



DataSet winequality-red

- CV density: 0,0018
- CV pH: 0,0466
- CV Calcohol: 0,1022
- CV fixed acidity: 0,2092
- CV sulphates: 0,2574
- CV volatile acidity: 0,3391
- CV chlorides: 0,5379
- CV residual sugar: 0,5551
- CV free sulfur dioxide: 0,6587
- CV total sulfur dioxide: 0,7076
- CV citric acid: 0,7186



5. (2.0 pt) Classifique o dataset iris usando o classificador DecisionStump. Descreva a saída em texto que o Weka fornece, tentando explicar cada um dos itens (e.x., confusion matrix, etc.). Usando o Weka Explorer, verifique se é possível encontrar um outro classificador que alcance uma taxa de erro menor que o Decision Stump. Caso positivo, diga qual o classificador usado (e.x., uma árvore decisão).

Saída da Weka

=== Run information ===
Scheme: weka.classifiers.trees.DecisionStump
Relation: iris
Instances: 150
Attributes: 5
sepallength
sepalwidth
petallength
petalwidth
class
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Decision Stump

Classifications: O Decision Stump só tem em conta dois classes de acordo com o atributo "*petallength"*, então ele não tem em conta a classe "*Iris-setosa".*

petallength <= 2.45 : Iris-setosa
petallength > 2.45 : Iris-versicolor
petallength is missing : Iris-setosa

Class distributions: As classes tem a mesma quantidade de dados, 0.33 para cada uma.

petallength <= 2.45
Iris-setosa Iris-versicolor Iris-virginica
1.0 0.0 0.0
petallength > 2.45
Iris-setosa Iris-versicolor Iris-virginica
0.0 0.5 0.5
petallength is missing
Iris-setosa Iris-versicolor Iris-virginica
0.3333333333333333 0.3333333333333333 0.3333333333333333

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

- O classificador tem um 66.66% de precisão, ele classificou 100 dados bem:
Correctly Classified Instances 100 66.6667 %

- O classificador tem um 33.33% de erro, ele classificou 50 dados ruim:
Incorrectly Classified Instances 50 33.3333 %

- O classificador tem um 50% de precisão de acordo com a metrica Kappa, ela tem em conta a escolha aleatória.
Kappa statistic 0.5

- O erro de acordo com a metrica da diferença de distância na previsão e na saída esperada.
Mean absolute error 0.2222

- O classificador tem um 33.33% de erro dada pela raiz da diferençã ao cuadrado
Root mean squared error 0.3333

- O classificador tem um 50% de erro de acordo com a metrica *Relatice absolute error* a qual é dada pela soma da diferença absoluta entre a saída do classificador e o valor esperado, que é dividido pela soma da diferença entre o valor esperado e a média esperada.
Relative absolute error 50 %

- A metrica é dada pela raiz da diferença ao cuadrado, tendo em conta a média esperada.
Root relative squared error 70.7107 %

Total Number of Instances 150

=== Detailed Accuracy By Class ≡

TP Rate: True Positive Rate é uma metrica que diz se o classificador acertou todos os dados de uma classe dada (neste caso a classe positiva).
FP Rate: False Positive Rate é uma metrica que diz quantos dados o classificador achava que eram da classe positiva, mas não era assim, isso em uma escala de 0 a 1.
Precision: A precisão do classificador.
Recall: É uma metrica que tem como visa saber com precisão quantos acertó propiamente de uma classe específica, do total dessa classe. **TP/(TP + FN)**
F-Measure: É uma metrica de accuracy, que envolve a *Precision* e o *Recall.*
MCC: Envolve o *TP Rate e o FP Rate, é uma metrica que está entre -1 e 1, sendo 1 o melhor caso, y -1 o ruim.*
ROC Area: *é o area baixo a curva, uma métrica para medir o desempenho do classificador.*
PRC Area: *Precision Recall Curve,* é uma metrica de desempenho que mide a precisão vs. Recall.

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1	0	1	1	1	1	1	1	**Iris-setosa**
1	0,5	0,5	1	0,667	0,5	0,75	0,5	**Iris-versicolor**
0	0	?	0	?	?	0,75	0,5	**Iris-virginica**
0,667	0,167	?	0,667	?	?	0,833	0,667	**Weight Avg.**

=== Confusion Matrix ===

A matriz de confusão diz a relação das classes, sendo o melhor score a matriz com só a diagonal diferente de cero, aquí se pode ver que o *Decision stump* classificou a classe "*c*" como se fosse "*b*".

```
 a  b  c   <-- classified as
50  0  0 |  a = Iris-setosa
 0 50  0 |  b = Iris-versicolor
 0 50  0 |  c = Iris-virginica
```

Melhor Classificador (Random Forest)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         143               95.3333 %
Incorrectly Classified Instances         7                4.6667 %
Kappa statistic                          0.93
Mean absolute error                      0.0408
Root mean squared error                  0.1621
Relative absolute error                  9.19   %
Root relative squared error             34.3846 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     Iris-setosa
                 0,940    0,040    0,922      0,940    0,931      0,896    0,991     0,984     Iris-versicolor
                 0,920    0,030    0,939      0,920    0,929      0,895    0,991     0,982     Iris-virginica
Weighted Avg.    0,953    0,023    0,953      0,953    0,953      0,930    0,994     0,989

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  4 46 |  c = Iris-virginica
```