UFPA PPGCC: Aprendizado de Máquina

Lista de exercício #1 - Data de entrega:24/04/2019

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).

   b. Calcule o histograma

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

   d. Calcule a taxa de colesterol média

   e. Calcule a taxa de colesterol mediana

   f. Calcule a variância e o desvio padrão amostral

2. (1.5 pt) Considere que os valores assumidos por um dado atributo numérico são listados no vetor $\mathbf{x} = \{1, 3, 2, 3, 2, 2, 0, 1, 0, 0, 3, 0, 2, 3, 2, 2, 3, 3, 0, 3, 2, 0\}$. a) *Calcule* o histograma de $\mathbf{x}$ (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 $\mathbf{X}$, *estime* sua média $\mathbf{E}[\mathbf{X}] = \mu$, $\mathbf{E}[\mathbf{X}^2]$, variância σ_x^2, o desvio padrão σ_x e o desvio médio absoluto. c) $\mathbf{X}$ é uma variável aleatórioa ou contínua?

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.

4. (2.5 pt) O Coeficiente de variação (CV) é uma medida relativa de variabilidade que independe da unidade de medida utilizada $CV = (Desvio_padrao/Media)$. 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.

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).