Lista de exercícios #1

Lista de Exercícios 1

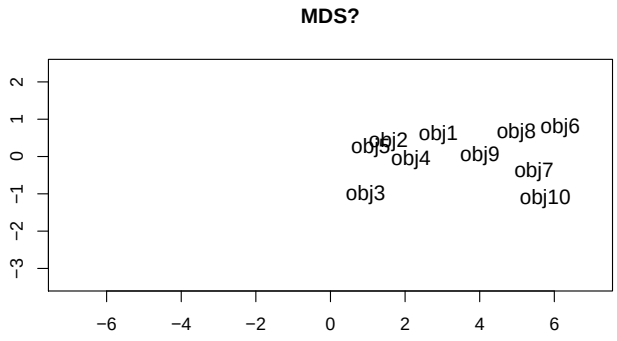
1. Estime a densidade de cada um dos conjuntos de amostras abaixo utilizando: *Parzen Windows* com *kernel* hiper-cubo unitário ($h = 1$ e $h = 4$) e Estimador de Máxima Verossimilhança assumindo que os dados são distribuídos de acordo com uma distribuição Normal. Para cada conjunto, indique qual abordagem obteve um resultado mais apropriado.

a) 11, 14, 14, 16, 15, 15, 16, 15, 13, 14, 14, 16, 15, 20, 22
b) 12, 12, 11, 12, 12, 11, 12, 15, 15, 15, 15, 15, 15, 15, 16
c) 15, 12, 11, 17, 18, 13, 16, 10, 13, 12, 18, 14, 18, 18, 17

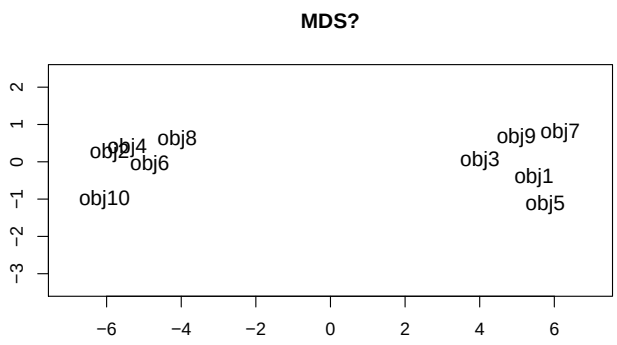
2. Foi utilizado escalonamento multidimensional para aproximar a base de dados que gerou a matriz de distâncias abaixo. Indique qual(is) gráfico(s) representa(m) a base de dados aproximada obtida (**não** é necessário realizar o cálculo para responder essa questão).

obj1	0.0	1.4	2.5	1.0	1.8	10.3	9.6	9.1	8.1	10.0
obj2	1.4	0.0	1.5	0.8	0.5	11.6	10.9	10.4	9.5	11.3
obj3	2.5	1.5	0.0	1.5	1.3	12.3	11.5	11.2	10.1	11.8
obj4	1.0	0.8	1.5	0.0	1.1	11.0	10.3	9.9	8.8	10.7
obj5	1.8	0.5	1.3	1.1	0.0	12.1	11.4	10.9	9.9	11.8
obj6	10.3	11.6	12.3	11.0	12.1	0.0	1.4	1.2	2.3	2.0
obj7	9.6	10.9	11.5	10.3	11.4	1.4	0.0	1.2	1.5	0.8
obj8	9.1	10.4	11.2	9.9	10.9	1.2	1.2	0.0	1.2	1.9
obj9	8.1	9.5	10.1	8.8	9.9	2.3	1.5	1.2	0.0	2.1
obj10	10.0	11.3	11.8	10.7	11.8	2.0	0.8	1.9	2.1	0.0

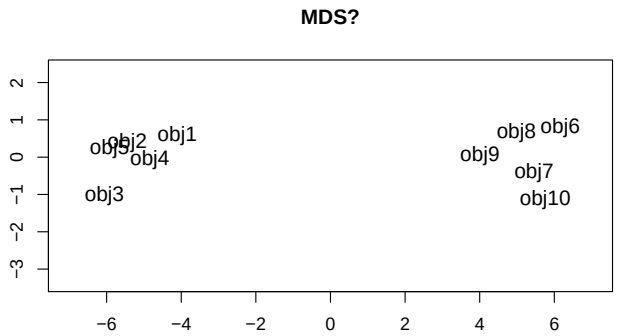
a)



b)



c)



4. Classifique os seguintes atributos como discretos ou contínuos. Também os classifique como qualitativos ou quantitativos. Note que alguns casos podem ter mais de uma interpretação, justifique a sua escolha em caso de ambiguidade.

(a) Tempo em termos de AM ou PM;
(b) Claridade mensurada por um medidor de luz;
(c) Claridade mensurada por pessoas;
(d) Ângulos mensurados em graus entre 0° e 360°;
(e) Medalhas olímpicas (bronze, prata e ouro);
(f) Altura acima do nível do mar;
(g) Número de pacientes em um hospital;
(h) ISBN de um livro;

5. Um diretor de marketing de uma empresa quer sua opinião sobre um critério para mensurar satisfação do consumidor que ele desenvolveu. Segundo ele: "O critério é tão simples que eu não acredito que ninguém pensou nisso antes. Eu conto o número de reclamações de cada produto. Li, em um livro de Mineração de Dados, que contagens são atributos proporcionais, portanto, meu critério é um atributo proporcional. Fiz a análise com nossos dados reais, nosso produto mais vendido é o que tem maior número de reclamações. Levei ao meu chefe o critério, porém, ele me dispensou dizendo que ignorei algo muito básico, logo, meu critério não serve para nada. Você pode me ajudar a convencê-lo?". O chefe tem razão? Se sim, qual a razão e como consertar o critério?

6. Alguns meses depois, o mesmo diretor de marketing te procura novamente. Dessa vez ele pensou em uma nova forma de mensurar o quanto um cliente prefere um produto sobre outros produtos similares. Segundo ele: "Quando desenvolvemos novos produtos, geramos variações e avaliamos quais os consumidores preferem. Nosso procedimento padrão é oferecer para cada pessoa no grupo de teste todas as variações, de uma só vez, e pedir para eles ordenarem de acordo com sua preferência. No entanto, eles são muito indecisos, especialmente quando tem mais de duas variações, e o processo acaba demorando muito. Sugeri que fossem feitas comparações em pares e que dessas comparações fossem geradas as ordenações. Portanto, com três variações apenas três comparações são feitas (1 vs 2, 2 vs 3 e 1 vs 3). Nosso procedimento agora leva um terço do tempo de antes, mas os avaliadores estão reclamando que não conseguem gerar uma ordenação consistente com os resultados. Você pode me ajudar?". Qual a origem do problema do novo procedimento de teste? Proponha uma correção para o procedimento.

7. Responda as seguintes questões sobre ruídos e *outliers*:

(a) Ruídos podem ser interessantes ou desejáveis? E *outliers*?
(b) Objetos com ruído podem ser *outliers*?
(c) Objetos com ruídos são sempre *outliers*?
(d) *Outliers* são sempre objetos com ruído?
(e) É possível que ruído cause um valor normal se transformar em um incomum e vice-versa?

8. Considere um conjunto de pontos $\mathcal{S}$ em um espaço Euclideano e uma medida de distância de cada

ponto em $\mathcal{S}$ e um ponto $\mathbf{x}$ (que pode estar ou não em $\mathcal{S}$). Considere o problema de encontrar todos os pontos em $\mathcal{S}$ que possuem distância de $\mathbf{y}$ menor que ϵ, sendo $\mathbf{y} \neq \mathbf{x}$. Explique como você pode usar a desigualdade triangular e distâncias já computadas em relação a $\mathbf{x}$ para reduzir o número de cálculo de distâncias. Lembrando que a desigualdade triangular é $d(\mathbf{a}, \mathbf{c}) \leq d(\mathbf{a}, \mathbf{b}) + d(\mathbf{b}, \mathbf{c})$.

9. Você acabou de receber uma nova base de dados para um problema de classificação. Seu chefe te pediu um valor razoável (*educated guess*) de taxa de acerto nessa base de dados para **ontem**. Um colega seu te disse que normalmente isso significa rodar os classificadores: discriminante linear (ou quadrático) e/ou Naïve Bayes em 70% dos dados e avaliar a taxa de acerto nos 30% restantes. Logo, você deve escolher entre essas três opções. Justifique quais informações são importantes para esta escolha, informando qual seria a escolha em cada caso. Adicionalmente, existe alguma equivalência entre essas escolhas? Se sim, explique qual.

10. Considerando a base de dados Iris[1], com o auxílio do computador, gere os modelos de análise de discriminante linear e Naïve Bayes considerando todos os 150 objetos e os seguintes atributos: área da pétala e área da sépala (devem ser gerados a partir dos originais). Gere o gráfico com a fronteira de decisão obtida no espaço formado por esses dois atributos. Dica: você pode gerar as equações das fronteiras ou avaliar cada posição do espaço de acordo com uma certa resolução (*e.g.*, valores com incrementos de 0,1). Essa última opção é mais simples e de fácil implementação.

[1] https://archive.ics.uci.edu/ml/datasets/iris