**Exercícios: aprendizagem baseada em instâncias**

1. Considere o conjunto de 6 exemplos de pontos no plano real e as suas correspondentes classificações. Dada uma nova instância a classificar, (4, 4), qual seria a classificação atribuída:
   1. Utilizando o algoritmo kNN, com k = 1 e k = 4?
   2. Que resultado seria obtido utilizando a ponderação das distâncias?



1. A tábua adjunta contem 8 instâncias que constituem o conjunto de treino para um classificador kNN. As duas últimas instâncias, numeradas 9 e 10, constituem o conjunto de teste. Classifique as instâncias de teste, utilizando a distância euclidiana:
   1. Utilizando o algoritmo kNN básico (k = 1, 3)
   2. Utilizando kNN com ponderação de distâncias (k = 3)



1. Suponha que são dados 11 exemplos de mensagens de correio electrónico as quais são classificadas como sendo e-Lixo ou não. Considere os 6 atributos como sendo booleanos e indicando o resultado do teste de uma determinada propriedade: o endereço do autor da mensagem tem como domínio edu? O campo “De” está vazio? O assunto está vazio? Existe a palavra “dinheiro” no corpo ou no assunto da mensagem? Existe o caracter “$” no corpo ou assunto da mensagem? Existe a palavra “pagamento” no corpo ou assunto da mensagem? Os dados são apresentados a seguir:



* 1. Utilizando o algoritmo do vizinho mais próximo (k = 1), como seria classificada a mensagem de teste convertida no vector de atributos-valores seguinte?



* 1. E no caso de utilizar o algoritmo dos k vizinhos mais próximos, com k = 5, qual seria o resultado para o mesmo exemplo de teste.
  2. Vamos fazer uma ligeira alteração aos 3 últimos atributos: estes deixam de ser booleanos para passarem a ser inteiros correspondentes ao número de ocorrências de uma palavra na mensagem de e-mail, no corpo ou no assunto. Os dados de treino são os seguintes:



Utilizando o algoritmo dos k vizinhos mais próximos, com k=3, qual é a classificação do exemplo de treino seguinte, utilizando como métrica a distância euclidiana?



1. Um fabricante de máquinas cortadoras de relva deseja encontrar uma maneira de classificar famílias em duas classes: os compradores de cortadores de relva e os não compradores. Esse fabricante teve acesso a informação sobre um conjunto de 24 famílias, habitantes em vivendas com jardim, em que 12 delas possuem cortadores de relva. Os atributos relevantes são: o rendimento da família e a área do jardim. O fabricante deseja aplicar o algoritmo de aprendizagem em instâncias dos k vizinhos mais próximos, mas não sabe que valor escolher para k. Para escolher k resolveu seguir a estratégia seguinte. Dividiu o conjunto de dados num conjunto de treino, com 18 famílias, e num conjunto de teste, com 6 famílias. O conjunto de teste consiste nas famílias 6, 7, 12, 14, 19 e 20 da tabela abaixo indicada. As restantes 18 observações constituem o conjunto de treino. O fabricante resolveu fazer experiências com vários valores de k, atribuindo a k valores no conjunto {1,3,5}, e recolher a taxa de classificações erradas. Resolveu escolher para k o valor com menor taxa de erro.



* 1. Que valor de k seria escolhido? Para justificar a sua resposta, apresente, para cada um dos valores de k acima indicados, as classificações obtidas para os exemplos de teste, e a taxa de erro correspondente.