**Descrição dos conjuntos de dados**

Durante esta atividade prática será utilizado 3 conjuntos de dados que são:

1. DataSet\_2D\_LinearlySeparable
2. DataSet\_2D\_NoLinearlySeparable
3. Diabetes

O conjunto de dados DataSet\_2D\_LinearlySeparable , é um conjunto de dados que está composto por 16 instâncias, 3 atributos, 16 folhas, e é um conjunto de dados do tipo numérico.

O conjunto de dado DataSet\_2D\_NoLinearlySeparable ,que é um conjunto de dados que está composto por 9 instâncias, 3 atributos, 9folhas, e é um conjunto de dado do tipo numérico.

O conjunto de dado diabetes,que é um conjunto de dados que está composto por 768 instâncias, 9 atributos, 768 folhas, e é um conjunto de dados do tipo numérico.

**Tarefa 1: classificação com dados linearmente separáveis**

Os dados linearmente separáveis o seu objectivo é encontrar um hiperplano h, de dimensão n – 1, que separe os exemplos da classe -1 dos exemplos da classe +1.

O seu Processo de aprendizagem consiste em determinar o vector de pesos (w) e o umbral (b).

Nesta atividade utilizaremos o Weka para treinar um classificador SVM para o caso de um problema simples, cujo objectivo é o reforço do entendimento básico dos conceitos relacionados às máquinas de vectores de suporte.

Treinemos primeiro uma SVM para um problema linearmente separável. Para tal utilizaremos o método SMO. SMO, abreviatura de *Sequential Minimal Optimization*, é um algoritmo que permite resolver o problema de optimização quadrática subjacente ao treino das SVM.

1- No painel *Preprocess*, carregue o conjunto de dados *DataSet\_2D\_LinearlySeparable*. Observe quais são os atributos e os valores correspondentes e as classes. Inspeccione os dados utilizando a secção *Visualize*.

R: os atributos do conjunto de dados são: x, y.

E os valores correspondentes são:

Para o atributo x são: Minimum 0; Maximum 15; Mean 6,438; StdDev 4,83;

Para o atributo y são: Minimum 0; Maximum 16; Mean 6,563; StdDev 4,98;

As classes são: A, B.

Os dados utilizados na secção visualize são:

Para o atributo x: 10, 6;

Para o atributo y: 11, 5;

Para a classe A: 8;

Para a classe B: 8; assim como ilustra a nossa figura abaixo.

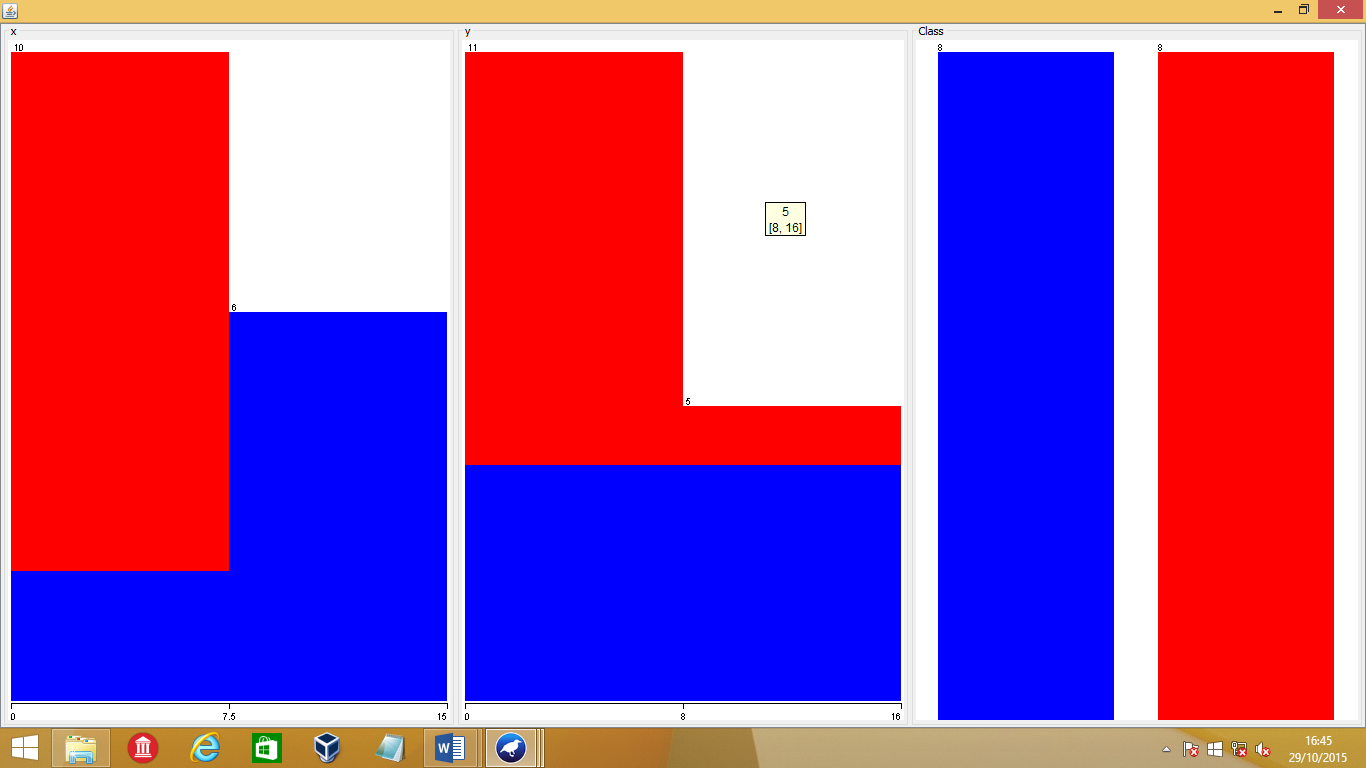


Fig.nº1

B-A outra forma consiste em utilizar a aplicação *ClassificationBoundaryVisualizer* do WEKA. Esta não está incluída no Explorador, porisso, é necessário abrir a aplicação numa nova janela.Construa a superfície de separação para o mesmo conjunto de dados e classificador utilizado no passo anterior.

* + 1. No WEKA abra a interface de comandos (*SimpleCLI*) e introduza o comando *java weka.gui.boundaryvisualizer.BoundaryVisualizer*.
    2. Seleccione o conjunto de dados e o classificador.
    3. Active a opção *Plot training data* e pressione o botão *Start*. Assim como ilustra a nossa figura.

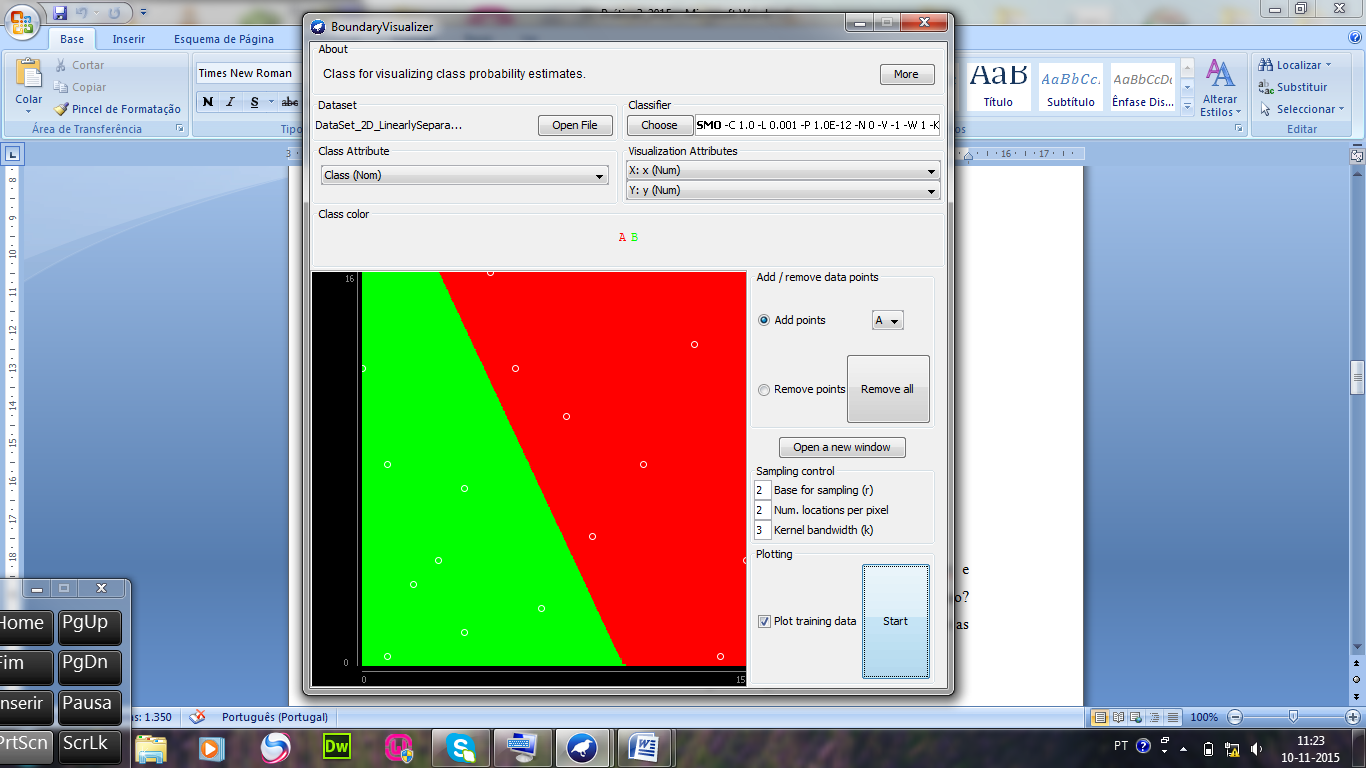


Fig.nº02

7. Analise as superfícies de decisão obtidas e descreva os erros de classificação.

7.R: Para um conjunto de dados linearmente separável existem várias superfícies de decisão

Possíveis.

8. R: Descrever os erros cometidos e acrescentar mais pontos para ver se altera alguma coisa

Visto que para a classe A existem 8 erros, e para a classe B também existem 8 erros.

Acrescentar pontos na classe A, B

*Tarefa*

Analise do conjunto de **dados não separáveis linearmente(***DataSet\_2D\_NoLinearlySeparable***).**

**O conjunto de dados tem 3 atributos,9 instacias ,9 sum of weigth ,composto por 2 atributos do tipo numérico e um do tipo nominal.**

1. Visualize os resultados da classificação. O que é que se observa?
2. **Após a execução do conjunto de dados** *DataSet\_2D\_NoLinearlySeparable com* As opções por defeito correspondem a um *kernel* linear (*kernel* polinómico com expoente 1). É possível observar o tempo que o weka levou para testar o conjunto de dados(0,02 segundos), a percentagem de instacias corectamente classificadas(55,556 %), a percentagem de instacias mal classificadas(44,4444 %) e com o total de 9 instacias...
3. Na caixa de diálogo de opções do classificador experimente alterar o valor do parâmetro de penalização (C) para 10, 100 e depois para 1000. Se observa algum efeito nos resultados da classificação?
4. Sim quando se troca o parâmetro do c para 10 nos temos 6 instancias corectamente classificadas que conrresponde a 66,6667%.
5. Quando c é 100