**Prática de Laboratório 1: pré-processamento dos dados e aprendizagem de árvores de decisão**

Nesta prática serão examinadas algumas operações de preprocessamento dos dados e a aprendizagem de árvores de decisão com o Weka.

Na mesma serão utilizados vários conjuntos de dados, *weather.nominal*, *iris*, *contact-lens*, *german-credit*, *glass* e *ionosphere*.

Com esta prática se pretende (a) aplicar algumas operações de preprocessamento sobre os dados, (b) utilizar dois métodos distintos, ID3 e C4.5, para a aprendizagem de árvores de decisão.

Como pré-requisitos para a mesma é importante o estudo de diversos aspectos:

* Operações de pré-processamento
  + normalização, discretização
  + consultar:
    - Theodoridis, pg. 164 – 166
    - Witten, pg. 314 – 322
* Avaliação dos modelos e métodos de validação
  + medidas de desempenho, holdout, validação cruzada, matriz de confusão
  + consultar: Witten, pgs. 147 – 150; 152 – 157; 180 - 182
* Árvores de decisão
  + Consultar: Mitchell, pg. 52 – 77

Abra o WEKA e entre na janela do Explorador.

**Tarefa 1**

Através do painel *Preprocess* carregue o conjunto de dados denominado *weather.nominal.arff*. Este contém uma versão do conjunto de dados *weather* com todos os atributos nominais.

Observe que o conjunto possui 14 instâncias e 5 atributos (*outlook*, *temperature*, *humidity*, *windy* e *play*), constituindo este último o atributo classe (*yes*, *no*). Click no nome de um atributo no subpainel da esquerda para ver a informação acerca do mesmo no subpainel da direita (os correspondentes valores e quantas instâncias no conjunto possuem um valor em particular). Quais são os possíveis valores do atributo *temperature*?

Carregue outro conjunto de dados, *iris.arff*. Quantas instâncias existem neste conjunto de dados? Quantos atributos? Qual é o rango de valores possíveis do atributo *petallength*?

É possível ver e editar dados a partir do Weka. Para tal carregue novamente o ficheiro *weather.nominal.arff*. Click o botão *Edit* na linha de botões do painel de preprocessamento. Com isto abre-se a janela do visualizador/editor (*viewer*), mostrando todas as instâncias do conjunto de dados. Qual é a função da primeira coluna no visualizador? Qual é a classe da instância 8 do conjunto de dados?

Carregue o conjunto de dados *iris* e abra-o no editor. Quantos atributos numéricos e quantos nominais tem o conjunto de dados?

**Tarefa 2**

Uma das vantagens das árvores de decisão é a sua "interpretabilidade", ou seja, as regras aprendidas para classificação são fáceis de ser seguidas por uma pessoa, a diferença de muitos métodos que funcionam como "caixas pretas". Trataremos de ver isso utilizando dois conjuntos de dados, *contact-lens* e *credit*.

No explorador do Weka, abra o ficheiro *contact-lens.arff*.

Vamos primeiro treinar o algoritmo ID3 com o referido conjunto de dados. Para tal, no separador *classify*, seleccione *classifiers > trees > ID3*. Seleccione *Use training set* na secção *Test options* e click no botão *start*.

Observe os resultados apresentados na janela da direita e trate de seguir a árvore de decisão aprendida. É compreensível a referida árvore?

Verifique que classificação seria atribuída pela árvore à instância *<age = young, spectacle-prescrip = myope, astigmatism = no, tear-prod-rate = normal>*.

Pode consultar informação ou as opções associadas ao algoritmo fazendo click no seu nome a direita do botão *Choose*. Click a continuação nos botões *More* e *Capabilities* e leia a informação associada.

Treine novamente a árvore de decisão utilizando como método de teste a validação cruzada em 5 folhas. Para tal, na secção *Test Options*, seleccione *crossvalidation* e *folds* = 5. Observe a quantidade (ou percentagem) de instâncias correctamente classificadas e compare com o caso anterior. O que é que isto sugere?

Carregue o ficheiro *credit.arff*. Este conjunto descreve, a partir de um conjunto de atributos, as pessoas como "boas" ou "de risco" para a concessão de crédito. Pode ler uma descrição deste conjunto de dados no site da UCI (<https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29>).

Ao utilizar um conjunto de dados, geralmente é boa ideia visualizá-lo primeiro. Vá para a secção *Visualize*. Click em qualquer dos gráficos para abrir uma nova janela que mostra um gráfico para dois atributos dados. Tente visualizar o gráfico para os atributos *age* e *duration*. Se observa algo inusual? Faça click sobre qualquer ponto para ver todos os valores dos seus atributos.

No ponto anterior, deve ter sido encontrado uma instância que parece estar corrompida, visto que alguns dos valores dos seus atributos não fazem sentido. Até um único ponto do gênero pode afectar significativamente o rendimento de um classificador. Uma boa forma de testar esta ideia é verificar o rendimento do classificador antes e depois de remover o referido ponto.

Para remover o referido ponto do conjunto de dados utilizaremos uma operação de pré-processamento (um filtro). Queremos remover todas as instâncias cujos valores de idade (atributo *age*) são inferiores a 0 anos. Na secção *Preprocess*, click em *Choose* no painel de filtros. Seleccione *filters > unsupervised > instance > RemoveWithValues*. Click no texto do filtro para alterar os seus parâmetros. Introduza 13 para o valor do índice do atributo (*attributeIndex*) e 0 para o ponto de corte (*splitPoint*). Click *Ok* para definir os parâmetros e *Aplly* para aplicar o filtro aos dados. Visualize novamente os dados para verificar que o ponto de dados inválido foi removido.

No painel de classificação, seleccione *Percentage split*, na secção *Test options*, e altere o seu valor para 90%. Com isto treinamos o classificador utilizando 90% dos dados e testamos o seu rendimento utilizando os restantes 10%.

Treinemos uma árvore de decisão. Para tal, seleccione *classifiers > trees > J48* e click *start*. J48 é a implementação no Weka do algoritmo C4.5 para a aprendizagem de árvores de decisão.

Uma vez treinado o classificador, a árvore aprendida é apresentada na janela da direita. A mesma também pode ser visualizada em forma gráfica fazendo click direito sobre o classificador, na lista de resultados à esquerda, e seleccionando *visualize tree*. Observe a saída do classificador e trate de responder às seguintes questões:

* Qual é o rendimento do classificador? A percentagem de instâncias correctamente classificadas é uma medida suficiente neste caso? Porquê? Sugestão: analise o número casos bem e mal classificados utilizando a matriz de confusão.
* Analise a árvore de decisão. As regras estabelecidas são razoáveis? Existem ramos que parecem absurdos? Sugestão: verifique as regras correspondentes às seguintes trajectórias (a) *CheckingAccount = <0, Foreign = Yes, Duration >11, job = Skilled, OtherDebtors = None, Duration <=30* e (b) *CheckingAccount = <0, Foreign = yes, Duration >11, job = unskilled*.
* Como maneja a árvore a classificação nos casos em que não existe nenhuma instância no conjunto de treino correspondente à determinada trajectória (os nodos folha assinalados com (0.0))?

**Tarefa 3**

Nesta tarefa aplicaremos outra operação de pré-processamento que é a discretização de dados numéricos. Esta é uma operação utilizada nos casos em que temos conjuntos de dados com atributos numéricos e pretendemos converter os referidos valores à uma escala nominal.

Existem dois tipos de discretização: *não supervisionada*, não tem em conta as classes das instâncias, e *supervisionada*, tem em conta as classes das instâncias ao criar os intervalos. O principal método de *discretização não supervisionada* existente no Weka é *Weka.filters.unsupervised.attribute.Discretize*. O mesmo implementa duas variantes: *discretização com intervalos igual largura* (seleccionado por defeito) e com *intervalos de igual frequência*.

Carregue o conjunto de dados *glass.arff* no explorador.

Aplique o método de *discretização não supervisionada* nos dois modos mencionados. Para tal no separador *Preprocess*, faça click no botão *Choose* e seleccione *filters > unsupervised > Discretize*.

* Para mudar de modo, faça click com o botão direito do rato na barra de edição diante do botão *Choose* e seleccione *Show properties* no menú de contexto. Na caixa de diálogo altere o valor da propriedade *useEqualFrequency*.

O que é que se observa ao comparar os histogramas obtidos? Os histogramas correspondentes à *discretização com intervalos de igual frequência* são bastante enviesados para alguns atributos. Porquê?

A principal técnica de *discretização supervisionada* de atributos numéricos no Weka é *weka.filters.supervised.attribute.Discretize*.

Carregue o conjunto de dados *iris*, aplique a *discretização supervisionada* e observe os histogramas obtidos. A *discretização supervisionada* tenta criar intervalos nos quais a distribuição das classes é consistente, ainda que a distribuição das instâncias varie de um intervalo para outro.

Com base nos histogramas obtidos, qual dos atributos discretizados acha que possui maior poder preditivo?

Carregue novamente o conjunto *glass* e aplique a *discretização supervisionada*.

Para alguns atributos existe uma única barra no histograma. O que é que isto significa?

Os atributos discretizados são normalmente codificados como atributos nominais, com um único valor para cada rango. Ambos os filtros permitem a criação de atributos com valores binários em vez de multivalorados, activando a opção *makeBinary*, na caixa de diálogo de propriedades.

Escolha um dos métodos e use-o para criar atributos binários. Compare os resultados com as saídas produzidas quando a opção *makeBinary* é falsa. O que é que representam os atributos binários?

Examinemos agora o efeito da discretização ao construir uma árvore de decisão com o algoritmo *C4.5* (designado *J48* no Weka) com os dados do conjunto *ionosfera*. Este conjunto de dados contém informação sobre sinais de radar retornados da ionosfera. Consideram-se “Boas” instâncias aquuelas que mostram evidências da existência de algum tipo de estrutura na ionosfera, enquanto que os sinais “Maus” são aqueles que atravessam directamente a ionosfera. Para mais informações, veja os comentários no ficheiro ARFF.

Iniciemos com a *discretização não supervisionada*.

Para o algoritmo *C4.5* (*J48*), *weka.classifiers.trees.J48*, compare taxas de erro da validação cruzada e o tamanho das árvores geradas pelos (1) dados em bruto, (2) dados discretizados usando o método não supervisionado de discretização no modo por defeito e (3) dados discretizados pelo mesmo método com atributos binários.

Vejamos agora a *discretização supervisionada*. Neste caso, se utilizamos a totalidade dos dados para a discretização os resultados da validação cruzada são demasiadamente optimistas. O procedimento correcto é costruir o filtro utilizando apenas os dados de treino e avaliar o classificador sobre os dados de teste, utilizando os intervalos de discretização determinados com os dados de treino. Para tal se utiliza um *metaclassificador* incluído no Weka, *FilteredClassifier* (*weka.classifiers.meta. FilteredClassifier*).

Utilizando *FilteredClassifier* e *J48*, compare a taxa de erro da validação cruzada e o tamanho das árvores geradas pela (4) *discretização supervisionada*, no modo definido por defeito, e (5) *discretização supervisionada* com atributos binários.

Compare os resultados do último exercício com os obtidos ao usar os dados em bruto determinados anteriormente. Como poderiam possivelmente ser as árvores de decisão geradas a partir de dados discretizados melhores predictores que as árvores de decisão geradas a partir de dados numéricos em bruto?