Aula2 – Modelos de Previsão –

Tratamento de numéricos, tratamento de nulos,

1. Comparar Naives Bayes com e sem discretização usando o dataset segment-challange e avaliando usando o ficheiro (segment-test).

Utilizando o algoritmo Naive Bayes, sem discretização, a percentagens de instancias corretamente classificadas foi de 77.037% e instancias incorretamente classificadas foi de 22.963%. O erro RMSE obtido foi de 0.2464

Para o mesmo algoritmo, mas utilizando um filtro de discretização, o número de instancias corretamente classificadas diminuiu para 91.9753% e instancias incorretamente classificadas foi de 8.0247%. O erro RMSE obtido foi de 0.1339

A aplicação dos filtros de discretização foi conseguida alterando as definições do algoritmo. Não é assim uma filtragem realizada no pré processamento.

Estes resultados são os esperados pois, ao realizar a discretização do conjunto de dados, é esperado que o RMSE baixe e que o número de casos corretamente identificados aumente.

1. A mesma avaliação usando o algoritmo J48. Considerar a árvore gerada e analisar os testes. Comparar estes valores com os valores obtidos na alínea anterior.

Ao utilizar o algoritmo de árvores J48, e com um intervalo de confiança alterado para 0.1, a percentagens de instancias corretamente classificadas foi de 96.4198% e instancias incorretamente classificadas foi de 3.5802%.

Neste algoritmo não faz sentido a aplicação do filtro de discretização, pois sendo dados numéricos, o algoritmo realiza automaticamente este processo.

O erro RMSE obtido foi de 0.0968, sendo assim muito mais baixo e, portanto, melhor que qualquer valor de RMSE obtido com o método Naive Bayes.

1. Avalie o dataset “vote”. Considere a árvore em relação ao treino com nulos. Analise os nós da árvore e as suas estatísticas.

Para avaliar o dataset “vote” foi utilizado o algoritmo de árvores J48 selecionando como opção de teste o método “Cross Validation Folds 10”.

1. Crie um conjunto de teste para o dataset vote (vote-test) com cinco casos onde há dois ou três atributos com valores nulos. Analise as previsões obtidas. Tente identificar o “caminho” escolhido para a previsão de cada caso.
2. Pruning: Volte a usar o dataset de treino e teste de “segment”. Avalie as consequências do pruning no J48 (opções “unpruning” e “subtree raising”).
3. Comparar modelos J48 usando os datasets anteriores. Compare a avaliação usando conjunto de treino versus conjunto de teste em modelos podados e não podados.