Trabalho Prático – SISTEMAS INTELIGENTES APLICADOS

Trabalho 1

DOCENTE: Thiago Naves

Integrantes:
O trabalho deve ser feito de forma individual ou em dupla.

Objetivo:
Consolidar uma base de dados (dataset) e teste para uso no algoritmo de classificação probabilística Naive Bayes disponível no Weka.

Escopo:
Vamos criar uma solução que seja capaz de identificar se um comentário sobre alguma coisa passa uma mensagem positiva ou negativa. Isso pode ser útil para identificar a satisfação de seus clientes sobre um respectivo produto, por exemplo.

O conjunto de dados que vamos utilizar são três fontes públicas que contém comentários dos sites: IMDb, Amazon e Yelp. Juntando todas as fontes de dados teremos três mil registros. Vamos começar coletando os dados que usaremos para treinar o nosso algoritmo classificador. Existe um dataset pronto para consumo que pode ser encontrando em https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences.

Percebam que existe três fontes de dados comentários e suas avaliações, é preciso consolidar essas três em uma única base que conterá 3003 registros. Em seguida é preciso tratar os dados, ou seja, colocar a base no formato preferencial de trabalho do Weka que é o .arff, mesmo formato que estão as bases de exemplo do próprio software.

Cada registro, ou seja, cada linha da base contém um comentário, seguido pela sua classificação. Onde o resultado é colocado como ‘0’ para comentários negativos e ‘1’ para comentários negativos, veja este exemplo: *Worst movie ever! 0*. Assim, verifique a base com objetivo de remover possíveis inconsistências. Por inconsistência você pode entender: os registros que não estão no formato correto, os registros que não estão classificados e os registros que não possuem o comentário.

A ferramenta Weka possui o algoritmo Naive Bayes implementado e com funções para visualização de performance e do resultado da classificação. O objetivo é testar a base de dados e coletar os dados de performance e resultado da classificação para elaborar um relatório sobre esses.

O relatório deve conter os seguintes itens:

- Resultados de performance e da classificação do algoritmo em relação a base de dados, cole a imagem do resultado com foco nos resultados de acurácia, verdadeiro positivo e falso positivo e precisão obtidos. Gere gráficos com os valores das métricas e discorra sobre elas. Pesquisem o significado e uso de cada uma das métricas para utilizar seus valores na descrição dos resultados obtidos.

- Utilizando três diferentes valores na estratégia de testes cross-validation, discorra sobre os resultados obtidos de acordo com:
    - Instancias classificadas de forma correta e incorreta
    - Taxa de Verdadeiro Positivo e Falso Positivo
    - Precisão
    - Matriz de Confusão

- Clicando com o botão esquerdo na caixa de texto ao lado do botão 'Choose' no Weka, abra a janela de opções. Modifique a opção 'UseSupervisedDiscretization' para 'true'. Nessa opção, ao invés de tratar cada atributo como uma variável real gaussiana, o algoritmo primeiro discretiza os atributos e depois utiliza contagens para estimar probabilidades. Execute o algoritmo e reporte a taxa de acerto obtida com essa opção e responda com sua anális:
    - A mudança foi significativa?
    - Por que isso aconteceu?
    - Resumo dos resultados e qual algoritmo possui melhor performance e qual possui melhor resultado de classificação

Requisitos do Trabalho:

Relatório do trabalho:

- Introdução sobre o algoritmo Naive Bayes.
- Exemplo da base de treino depois de configurada.
- Descrever os resultados de acordo com os itens descritos acima.

Forma de Avaliação:

Será avaliado se o trabalho atendeu a todos os requisitos especificados anteriormente. Quaisquer elementos adicionais como novas funções e uso de métricas para avaliar a qualidade da classificação serão adicionados pontos extras.

Forma de Entrega:

Entrega será feita pelo moodle, a data e tarefa para entrega dos trabalhos já está disponível no site.