UFV - Universidade Federal de Viçosa
DPI - Departamento de Informática
Trabalho 4 - Detecção de SPAM - 2019

Algoritmo de Bayes Ingênuo
Entrega: 28/06/19 (23:59)

Inteligência Artificial (INF 420)

Informações Gerais

Neste trabalho você irá implementar o algoritmo de classificação Bayes Ingênuo para solucionar o problema de detecção de SPAMs. O trabalho deve ser desenvolvido individualmente. A colaboração entre colegas de classe é permitida, mas sempre através da troca de ideias, nunca através da troca de código fonte.

Política de Atraso Cada aluno poderá atrasar um total de 7 dias a entrega de trabalhos durante o semestre. Uma vez esgotados os dias de atraso, a nota do trabalho atrasado será zero. Por exemplo, ao entregar o primeiro trabalho com 2 dias de atraso, a nota do aluno não sofrerá nenhuma penalização. No entanto, se no segundo trabalho o aluno atrasar 6 ou mais dias, a nota do segundo trabalho será zero.

Implementação e Entrega Implemente um sistema de detecção de SPAMs utilizando o algoritmo de classificação de Bayes Ingênuo. Utilize a estratégia de dicionário de palavras com suavização aditiva para a implementação do algoritmo. Uma vez implementado e treinado, o sistema receberá como entrada um email (no formato que lhe for mais conveniente) e retornará um rótulo dizendo se o email é SPAM ou não. O trabalho deverá ser desenvolvido em Python 3 através de um Jupyter Notebook. A entrega será o html do seu notebook via PVANet, com a data limite de 28/06 (23:59).

Base de Dados A base de dados para treinamento e teste está disponível no PVANet. Cada arquivo .eml contém um email. Os rótulos se encontram no arquivo *SPAMTrain.label*, onde 0 indica SPAM e 1 indica HAM. O arquivo README indica um conjunto de teste, mas esse conjunto não está disponível, vocês terão que separar um conjunto de teste a partir do que é fornecido no trabalho (vejam instruções abaixo).

O Que Reportar

Além do código fonte do algoritmo, o html do seu notebook deve conter os seguintes resultados:

1. Os valores médios de acurácia de um procedimento de validação cruzada para escolher o parâmetro de suavização aditiva. Lembrem-se que a validação cruzada deverá ser feita sem considerar o conjunto de teste, que deve ser deixado de lado.

2. Matriz de confusão da versão final do algoritmo no conjunto de teste, que deve ser selecionado por você a partir dos dados rotulados do trabalho. O conjunto de teste deve ser pelo menos 15% dos dados rotulados fornecidos.