Fundamentos de IA
Projeto - Aprendizado de Máquina **09/08/2016**
Nome: _____
Prof. Dr. Sylvio Barbon Jr

Projeto: K-NN e Árvores de Decisão

As Redes Sociais Digitais (RSD), também conhecidas como Redes Sociais *Online* (RSO), são consideradas ambientes digitais nos quais as pessoas discutem ideias e expressam opiniões sobre qualquer assunto. Atualmente, RSD representam uma fonte relevante de informações a serem exploradas em áreas como avaliação de opiniões e pesquisas de marketing.
Entretanto, se por um lado as pessoas conseguem se beneficiar do conteúdo disponível em RSD, é possível também causar danos. A grande popularidade e facilidade de acesso às RSD também resultou na presença de usuários não desejados. Juntamente com problemas relacionados a privacidade de seus próprios usuários, RSD apresentam problemas para identificar casos que envolvam usuários desempenhando atividades maliciosas como o *spam*, envio repetido e excessivo de conteúdo não solicitado, sendo o mais comum.
Contas falsas, são em grande parte, representadas por contas com comportamento automático denominadas robôs e popularmente conhecidas como *bots* (do inglês *robots*). O objetivo dessas contas varia entre a postagem de *spam* de produtos, links maliciosos para prática de *phishing*, atividade uma tentativa de fraude em que se tenta obter dados sensíveis como senhas da vítima, ou simplesmente fazem volume para aparentar que uma dada entidade é mais popular do que realmente é. No Twitter, *bots* tem como objetivo se passar por humanos para assim ganharem seguidores e serem capazes de disseminar suas atividades em grande escala.

1. (2 points) Considerando o contexto descrito, use a base de dados disponível aqui e realize tarefas a seguir. O conjunto de dados apresenta três aquivos isolados. Cada arquivo apresenta exemplos de uma classe do problema. São três classes ao todo: Humano, Bot Legítimo e Bot Fraudulento. Para os experimentos, recomenda-se o uso de mais de 30 repetições (holdout 30/70) e apresentação dos resultados com gráficos que possibilitem a interpretação dos resultados. As tarefas devem ser apresentadas em forma de um relatório com a apresentação e discussão dos valores alcançados.

    1. Utilizando o algoritmo k-NN, explore qual o melhor valor de k para realizar a classificação da base de dados considerando todos os atributos apresentados.
    2. Utilizando uma árvore de decisão, realize experimentos de classificação e identifique quais os atributos selecionados e desempenho de classificação.
    3. Agora, com base nos atributos selecionados pela AD, execute experimentos com o k-NN com o k que obteve o melhor desempenho. Discuta os resultados comparando com a Questão 1.
    4. Qual a classe mais complexa de ser modelada? Algum algoritmo é superior para uma dada classe?
    5. Conclua sobre os resultados obtidos e discuta possíveis estratégias para melhoria dos modelos obtido.