Universidade Federal de Santa Catarina
Centro Tecnológico de Joinville
Programa de Pós-Graduação em Engenharia de Sistemas Eletrônicos

Tópicos – Aprendizado de Máquina
Prof. Lucas Weihmann / Prof. Pablo Andretta Jaskowiak

Trabalho Prático

O presente trabalho consiste na utilização e avaliação de modelos de classificação para a tarefa de predição de spam. Para tanto, deve ser utilizada a base de dados *spambase*, disponível no UCI Machine Learning Repository: https://archive-beta.ics.uci.edu/dataset/94/spambase. Note que, embora a tarefa a ser abordada esteja relacionada com corpos textuais (emails), já ocorreu um pré-processamento que converteu cada email em um conjunto de atributos numéricos descritivos. Maiores informações acerca deste pré-processamento podem ser obtidas no artigo original (fonte disponível no link acima). Note ainda, que no site da UCI há valores de precisão e revocação tipicamente obtidos para esta base de dados com modelos populares da literatura. Entretanto, nada é dito acerca de quais hiperparâmetros foram utilizados para tanto.

Tendo em vista o exposto, seu trabalho consiste em:

1. Realizar uma análise exploratória na base de dados e reportar seus achados *mais relevantes*. Existem atributos irrelevantes? redundantes? Podem ser removidos atributos? Quais? Por que? Se você optar por realizar sua análise considerando diferentes subconjuntos de atributos, justifique.
2. Definir um conjunto de modelos de classificação para serem avaliados no contexto do problema (pelo menos três modelos diferentes). Você não precisa se limitar a modelos vistos em sala de aula, podendo explorar outros existentes. Não é necessário implementar os modelos do zero, isto é, você pode utilizar bibliotecas e/ou códigos prontos que forneçam estes modelos para utilização.
3. Definir quais hiperparâmetros e respectivos valores serão avaliados para cada modelo. A partir desta definição, realizar uma busca em grade (*grid search*) para encontrar a *melhor* configuração de hiperparâmetros de cada modelo. Note que a forma com a qual a qualidade do modelo será estimada e as métricas que serão utilizadas para tal estimativa são de livre escolha. Sua escolha, entretanto, deve ser minimamente justificada e embasada. Esta etapa deve ser programada por você.
4. Fornecer, avaliar e interpretar os resultados obtidos na etapa anterior por meio de tabelas e principalmente gráficos. Importante: você deve fazer aqui uma análise crítica, discutindo os resultados. A atribuição de nota considerará não só a qualidade final dos resultados obtidos, mas também a forma como eles foram apresentados e discutidos. Portanto, pense bem em como organizar, sumarizar e discutir os resultados obtidos.

A utilização de códigos e bibliotecas para modelos de classificação, estimação de qualidade e métricas é *livre*. A utilização de códigos de terceiros para a realização dos experimentos computacionais, isto é, avaliação e obtenção dos resultados aqui solicitados *não é permitida*, visto que o foco do trabalho é justamente este: a aplicação, avaliação e interpretação dos resultados de diferentes modelos.

Você deve entregar: o código fonte utilizado para realizar as análises; um relatório, no formato de artigo, seguindo o modelo da IEEE, disponível em: https://www.ieee.org/conferences/publishing/templates.html. O relatório possui limite rígido de 2 páginas, incluindo referências. O trabalho é individual.