Machine Learning I - Trabalho 1

11 de Setembro de 2018

Este trabalho objetiva avaliar os conhecimentos na resolução de problemas de classificação e regressão usando Aprendizado de Máquina.

Todo o desenvolvimento deverá ser documentado para poder ser avaliado. Se preferir, poderá colocar a documentação em Markdown e código direto nos notebooks do Jupyter.

1 Problemas de Classificação

1.1 *Human Activity Recognition Using Smartphones Dataset*

O dataset de reconhecimento de atividade humana (HAR) [1] consiste de dados coletados pelos sensores de um telefone (acelerômetro e giroscópio) em um experimento envolvendo 30 voluntários com idade entre 19 e 48 anos. Cada um dos voluntários executou seis tipos de atividade (**walking, walking upstairs, walking downstairs, sitting, standing, laying**) enquanto carregava

o aparelho. Para mais informações e download do dataset, basta acessar este link.

A resolução deste problema deverá ser toda documentada e terá que seguir os seguintes passos:

1. Analisar os dados e tomar algumas ações, por exemplo:
    - Limpeza dos dados, caso necessário;
    - Seleção de features;
    - Normalização dos dados;
    - Balanceamento dos dados, caso necessário;
    - etc.
2. O dataset já disponibiliza um split de treino e outro de teste, logo não há necessidade de utilizar K-folds. A separação de uma parte do treino para fazer validação é recomendável.
3. Escolher **ao menos** 3 algoritmos diferentes para resolverem este problema.
4. Comparar os resultados dos algoritmos considerando **acurácia, precisão e sensibilidade (*recall*)**.

1.2 *Credit Card Fraud Detection*

Ser cobrado por algo que não foi adquirido é péssimo, não é? Por isso as companias de cartão de crédito se esforçam em garantir que isso não aconteça, reconhecendo transações fraudulentas. Este dataset contém informações (a maioria

anonimazada) disponibilizadas por companias européias sobre transações feitas por cartões de crédito em Setembro de 2013. Mais especificamente, ele contém dados sobre transações que ocorreram em 2 dias, onde temos 492 fraudes contra 284,807 transações normais, ou seja, um dataset altamente desbalanceado.

As colunas $V1, V2, ..., V28$ são as componentes principais obtidas com o uso de PCA, as únicas colunas que não foram transformadas são "*Time*"e "*Amount*". Este dataset está disponível em `https://www.kaggle.com/mlg-ulb/creditcardfraud`.

Para resolver este problema, use um modelo baseado em **Máquinas de Vetor de Suporte** (SVM) e faça comparação entre os resultados obtidos por uma **SVM linear, SVM com kernel polinomial e SVM com kernel RBF (*Radial Basis Function*)**.

Referências

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013.