Machine Learning II - Trabalho 1

14 de Abril de 2018

Este trabalho objetiva avaliar os conhecimentos na resolução de problemas de classificação, regressão e clusterização 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 Problema 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

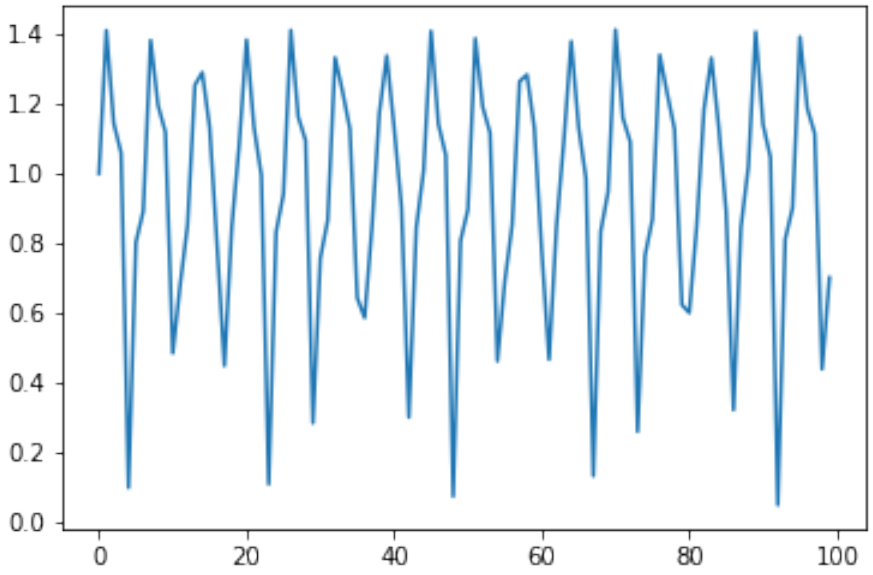
o aparelho. Para mais informações e download do dataset, basta acessar este link.

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. **Um** deles pelo menos deverá ser um método baseado em ensembles.
4. Comparar os resultados dos algoritmos considerando **acurácia, precisão e sensibilidade (*recall*)**.

2 Problema de Regressão

Utilize a rede neural perceptron de múltiplas camadas para fazer a predição de alguns passos da série temporal $\sqrt{1 + \sin(n + \sin^2(n))}$, onde n é um passo na série.



Avalie o desempenho mostrando o erro de predição e os resíduos da regressão:

1. Considerando 1 passo a frente;
2. Considerando 10 passos a frente;
3. Considerando 100 passos a frente

A utilização de gráficos para mostrar o resultado é encorajado.

3 Problema de Clusterização

Considere um problema de apenas 2 dimensões gerado artificialmente, faça a visualização dos dados para ter uma ideia inicial sobre a quantidade de grupos. Os dados estarão contidos no arquivo texto **unlabeled_data.txt**.

1. Selecione alguns métodos de identificação de clusters como K-means, DBSCAN, entre outros.

2. Avalie a performance desses algoritmos com métricas como: Silhouette Coefficient, Calinski-Harabaz (C-H), etc.

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.