Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Márcio Alexandre Costa
27 de abril de 2019

Proposta

Predizer o valor de casas

Construir um modelo que possa predizer o valor de casas com base na localização e característica da casa ideal.

Descrição do problema

O objetivo é predizer o valor de venda de cada casa em Ames / Iowa nos Estados Unidos.

Com base na descrição da casa ideal que o comprador descrever poder predizer o valor de venda destas casas incluindo uma variedade de características além de metragem do terreno, números de quarto e banheiros.

O modelo poderá servir em balizar o sonho da casa ideal verso a valor a ser pago por esta "casa dos sonhos".

Conjuntos de dados e entradas

Estarei utilizando a conjunto de dados do Kaggle no link https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data.

O conjunto de dados contém 79 características e 1.460 entradas, sendo 43 características com valores categóricas e 36 características com valores numéricas. O conjunto de dados contém 19 características com valores ausentes.

O conjunto de dados está com uma distribuição normal, gaussiana, porém apresenta alguns outliers.

Descrição da solução

Utilizar técnica avançadas de Regressão para criação do modelo.

Será utilizado algoritmo de Regressão da aprendizagem de máquina supervisionada. Os algoritmos de regressão que iremos trabalhar nesta proposta são: Lasso, Simple e/ou Multiple Linear Regression, Support Vector Regression (SVR), Decision Tree Regression e Random Forest regression.

Modelo de referência (benchmark)

Iremos criar um modelo com o algoritmo 'Naive' de Regressão para utilizar como modelo de benchmark comparando com o modelo de regressão mais avançados.

Métricas de avaliação

Será avaliado pelo RMSE (Root Mean Squared Error) entre o logaritmo do valor previsto e o logaritmo do valor de venda observado.

Design do projeto

O Fluxo de trabalho será:

- Entendimento dos dados;
- Limpeza dos dados;
- Correção das características (features);
- Verificar se é possível reduzir as características (features) para a criação do modelo;
- Treinar vários algoritmos de regressão;
- Avaliar vários algoritmos de regressão comparando com o benchmark;
- Definir o melhor modelo.

Referências

De Cook, Dean. “Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project.” Journal of Statistics Education, vol. 19, no. 3, 2011 - https://amstat.tandfonline.com/doi/abs/10.1080/10691898.2011.11889627#aHR0cHM6Ly9hbXN0YXQudGFuZGZvbmxpbmUuY29tL2RvaS9wZGYvMTAuMTA4MC8xMDY5MTg5OC4yMDExLjExODg5NjI3P25lZWRBY2Nlc3M9dHJ1ZUBAQDA= .

De Hujia Yu, Jiafu Wu. “Real Estate Price Prediction with Regression and Classification.” CS 229 Autumn 2016 Project Final Report - http://cs229.stanford.edu/proj2016/report/WuYu_HousingPrice_report.pdf