练习

2023 年 9 月 18 日

练习 1：加州房屋价格预测

数据集来源：
https://www.kaggle.com/datasets/fedesoriano/california-housing-prices-data-extra-features

数据集位置：
‘数据/California_Houses.csv’

数据集描述：

This is the dataset is a modified version of the California Housing Data used in the paper Pace, R. Kelley, and Ronald Barry. “Sparse spatial autoregressions.” Statistics & Probability Letters 33.3 (1997): 291-297.. It serves as an excellent introduction to implementing machine learning algorithms because it requires rudimentary data cleaning, has an easily understandable list of variables and sits at an optimal size between being too toyish and too cumbersome.

The data contains information from the 1990 California census. So although it may not help you with predicting current housing prices like the Zillow Zestimate dataset, it does provide an accessible introductory dataset for teaching people about the basics of machine learning.

Modifications with respect to the original data

This dataset includes 5 extra features defined by me: “Distance to coast”, “Distance to Los Angeles”, “Distance to San Diego”, “Distance to San Jose”, and “Distance to San Francisco”. These extra features try to account for the distance to the nearest coast and the distance to the centre of the largest cities in California.

The distances were calculated using the Haversine formula with the Longitude and Latitude:

where:

phi_1 and phi_2 are the Latitudes of point 1 and point 2, respectively lambda_1 and lambda_2 are the Longitudes of point 1 and point 2, respectively r is the radius of the Earth (6371km) Content

The data pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data. The columns are as follows, their names are pretty self-explanatory:

1) Median House Value: Median house value for households within a block (measured in US Dollars) [$]
2) Median Income: Median income for households within a block of houses (measured in tens of thousands of US Dollars) [10k$]
3) Median Age: Median age of a house within a block; a lower number is a newer building [years]
4) Total Rooms: Total number of rooms within a block
5) Total Bedrooms: Total number of bedrooms within a block
6) Population: Total number of people residing within a block
7) Households: Total number of households, a group of people residing within a home unit, for a block
8) Latitude: A measure of how far north a house is; a higher value is farther north [°]
9) Longitude: A measure of how far west a house is; a higher value is farther west [°]
10) Distance to coast: Distance to the nearest coast point [m]
11) Distance to Los Angeles: Distance to the centre of Los Angeles [m]
12) Distance to San Diego: Distance to the centre of San Diego [m]
13) Distance to San Jose: Distance to the centre of San Jose [m]
14) Distance to San Francisco: Distance to the centre of San Francisco [m]

Source

This data was entirely modified and cleaned by me. The original data (without the distance features) was initially featured in the following paper: Pace, R. Kelley, and Ronald Barry. "Sparse spatial autoregressions." Statistics & Probability Letters 33.3 (1997): 291-297.

The original dataset can be found under the following link: https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html

[ ]: