CH 2 End-to-End ML Project

mrevanisworking

August 14, 2020

1. Look at big picture
2. Get the data
3. Discover and visualize the data to gain insights
4. Prepare the data for ML algorithms
5. Select a model and train it
6. Fine-tune your model
7. Present your solution
8. Launch, monitor, and maintain your system

0.1 Working with Real Data

0.1.1 Popular Places to get datasets

UC irvine ML Repo
Kaggle, Amazon AWS, Data Portals, Quandl, Wikipedia, Quora, reddit

0.1.2 This chapter's dataset

California Housing Prices (1990 California census)

0.2 Look at the Big Picture

Task: use California census data to build a model of housing prices in the state.
Data includes: population, median income, median housing price for each block group
Block group: the smallest geographical unit for which the UC Census Bureau publishes sample data "distrcts"

0.2.1 Frame the Problem

Ask what the business objective is
How does the company expect to use and benefit from this model?
This Answer: model's output will be fed to another ML system.
Signal: a piece of info fed to a ML sys
Ask what the current solution is (if any)
Frame the problem:
is it supervised, unsupervised, RL (Reinforcement Learning) ?
is it classification, regression, or something else?
batch learning or online learning?
Answer: since given labeled training examples, it is a supervised learning task. regression because predict value. (multiple regression, univariate regression)

0.2.2 Select a Performance Measure

a typical performance measure for regression is RMSE

0.2.3 Pipeline

a secuence of data processing components. components usually run ASYNC.