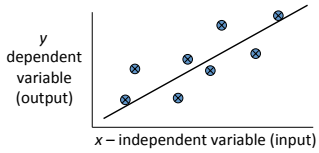
Week 7:
An End to End Regression Example

Regression

- In classification, the outputs are nominal
- In regression, the outputs are continuous values
- Many models could be used – Simplest is linear regression
  - Fit data with the best hyper-plane which "goes through" the points



California Housing Data

Frame the problem:

- What are the potential upstream and downstream systems for this project?
- What is the problem we are working at? Supervised or un-supervised? Regression or classification?
- What is the Performance Measure selected?

Working with Real Data

Understand the data
Discover and visualize the data to gain insights.

- How many features?
- Any missing values?
- Any categorical features?
- Have the data been pre-processed? Scaled or capped?

Create a Test Set

- What is the problem with randomly splitting the original data set into training and test sets?
  - How did we create test sets for MNIST?
- How do we create buckets for “median income” and enable stratified data splitting?
- What if we will add new data in the future but maintain the same samples in the test set?

Feature Engineering

- How of find out the most relevant features for learning?
- What new features can we introduce by combining existing features?

Preparing Data

- How to deal with missing values?
- How to handle categorical features?
- How to implement a Scikit-learn transformer through "duck typing"?
- Normalization (min-max scaling) v.s. Standardization?
- How to build a pipeline for preprocessing the numerical attributes:

Select Model

- What tells you when the errors produced by the training set are too small or too large?
- Why do we use cross-validation technique in training our models?

Fine Tune the Model

- What is the purpose of Scikit-Learn's GridSearchCV?
- GridSearchCV v.s. RandomizedSearchCV?

Model Evaluation

- What are the steps you can take to evaluate your system's performance?
- How to fix to the problem of system performance degradation?