
Machine Learning
Crash Course
Week 6

“

The key to artificial intelligence has always been the representation.” ~ Jeff Hawkins

Topics

- Machine Learning in Practice
  - Hypothesis Evaluation
  - Model Selection
  - Data Train and Test Sets
  - Bias & Variance
  - Bias, Variance, and Regularization

- System Design & Workflow
  - Task Breakdown
  - Data Discussion

1.
Machine Learning in Practice

Hypothesis Evaluation

Problem

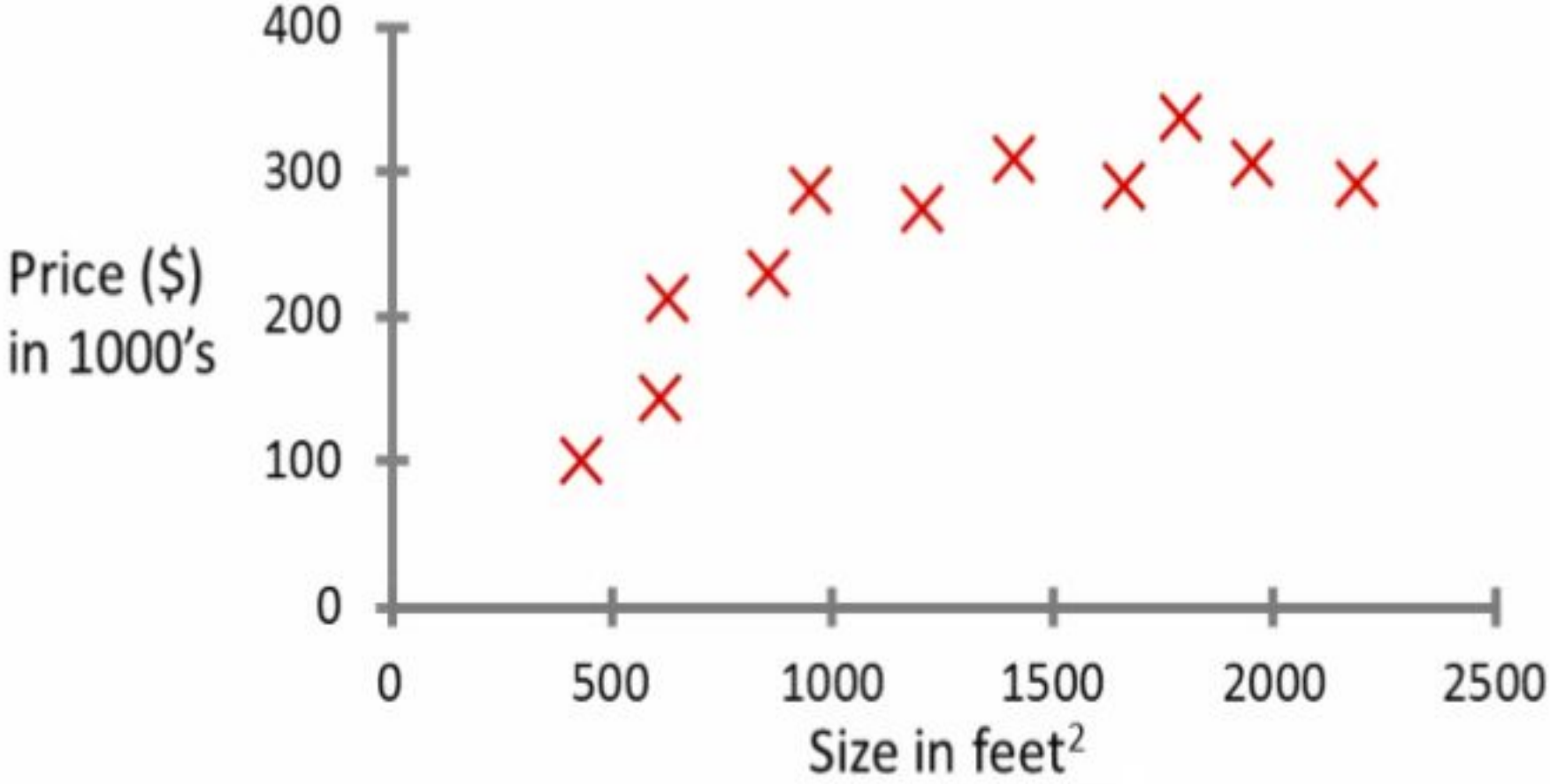
Suppose you have implemented regularized linear regression to predict housing prices.

- Now given a new test set of house data and a very large error.
- **What methodologies could you use to tackle this issue?**

Solution

1. More data, use more or less features, using polynomial features, or trying to increase and decrease lambda.
2. Design a Machine Learning Diagnostic

Housing price prediction.



Model Selection

- Consider creating an extra parameter for the degree of the polynomial
- Then measure the test error for each θ
- Then compare the the test error against how well the model generalizes
- **Problem:**
  - This is only an approximation of the generalization error
  - **DEMO**

Data Train and Test Sets

Solution:

- Partitioning your data into different sets
  - Training
  - Cross-Validation
  - Test
- **DEMO**

- 3 corresponding errors to calculate

$$\mathrm{J}_{train}(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)^2$$

$$\mathrm{J}_{cv}(\theta) = \frac{1}{2m_{cv}}\sum_{i=1}^{m_{cv}}(h_\theta(x^i_{cv}) - y^i_{cv})^2$$

$$\mathrm{J}_{test}(\theta) = \frac{1}{2m_{test}}\sum_{i=1}^{m_{test}}(h_\theta(x^i_{test}) - y^i_{test})^2$$

Bias & Variance

Underperformance Issues

- (**1**) Usually related to high bias (underfit) or high variance (overfit)
- (**2**) **How would we distinguish a high bias or variance issue?**

Solution

- (**1**) Plot the error against the degree of the polynomial **(DEMO)**
- **NOTE:**
  - Training error decreases as the model progresses from (**underfit**) to (**overfit**)

Bias, Variance, & Regularization

Underperformance Issues

- Usually related to high λ (underfit) and small λ (overfit)
- **DEMO**

Solution

- Given your hypothesis, cost function, optimization goals w/o reg.
- **Experiment with different λ in the cost functions**
- **DEMO**

SUMMARY

W.R.T. J_{train}

- **Small λ:** Small Reg. term, better fit hypothesis, low J_{train}
- **Large λ:** Large Reg. term, worse fit hypothesis, high J_{train}

W.R.T. J_{cv}

- **Small λ:** Large Reg. term, underfit, low J_{cv}
- **Large λ:** Small Reg. term, overfit, low J_{cv}

2.
System Design & Workflow

Task Breakdown **w/DEMO**

Data Analysis

Model Design

Evaluation & Optimization

Data Analysis

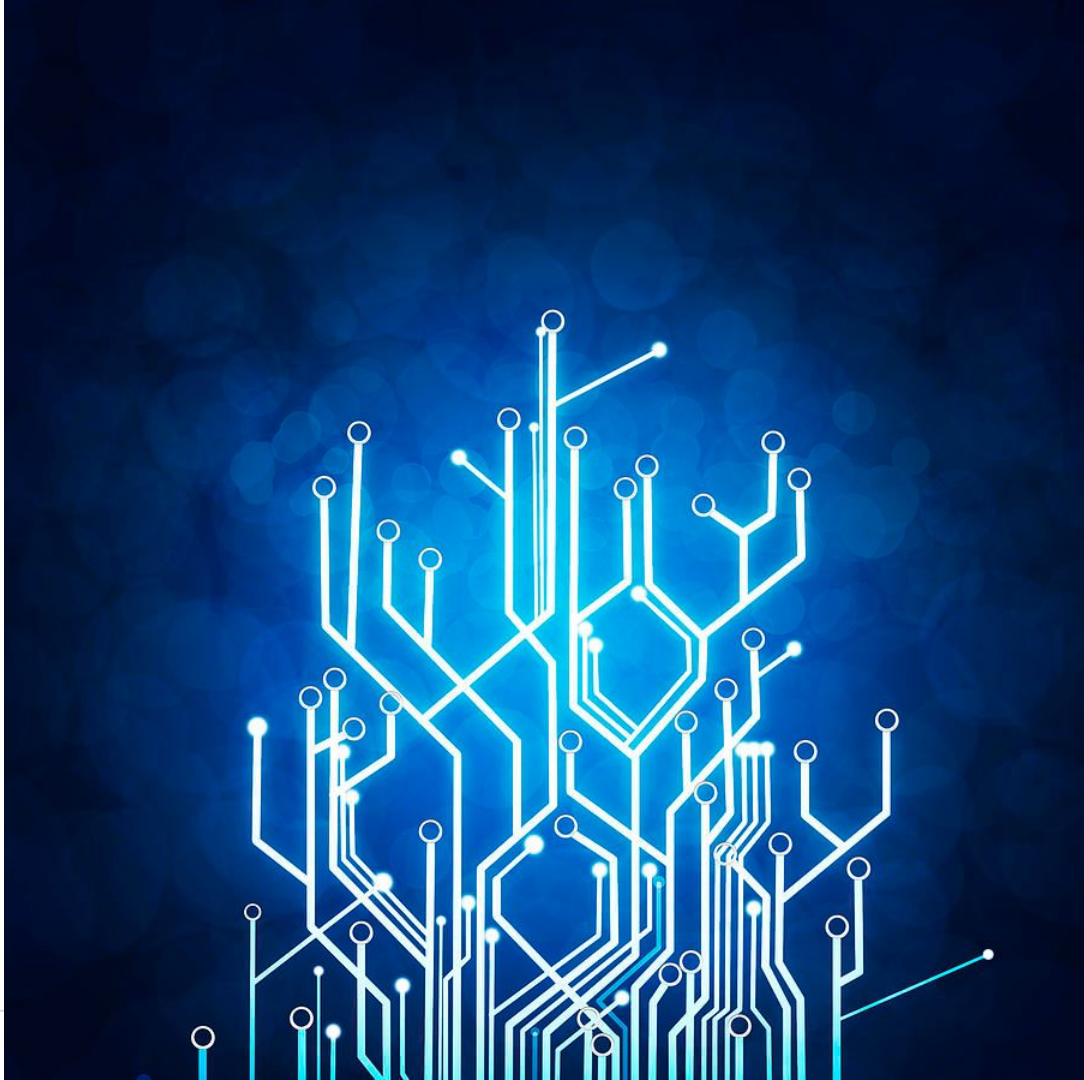
You should check ahead of time if the data has redundant features and/or enough features to learn from.

Thanks!

Any questions?

You can find me at:

- cs.oswego.edu/~kzeller



SOURCES

- https://towardsdatascience.com/machine-learning-in-practice-what-are-the-steps-a4b15ee18546
- https://www.geeksforgeeks.org/getting-started-machine-learning/
- https://www.ritchieng.com/
- https://machinelearningmastery.com/machine-learning-in-python-step-by-step/
- https://www.datacamp.com/community/tutorials/kaggle-tutorial-machine-learning
- https://www.kaggle.com/dansbecker/your-first-machine-learning-model