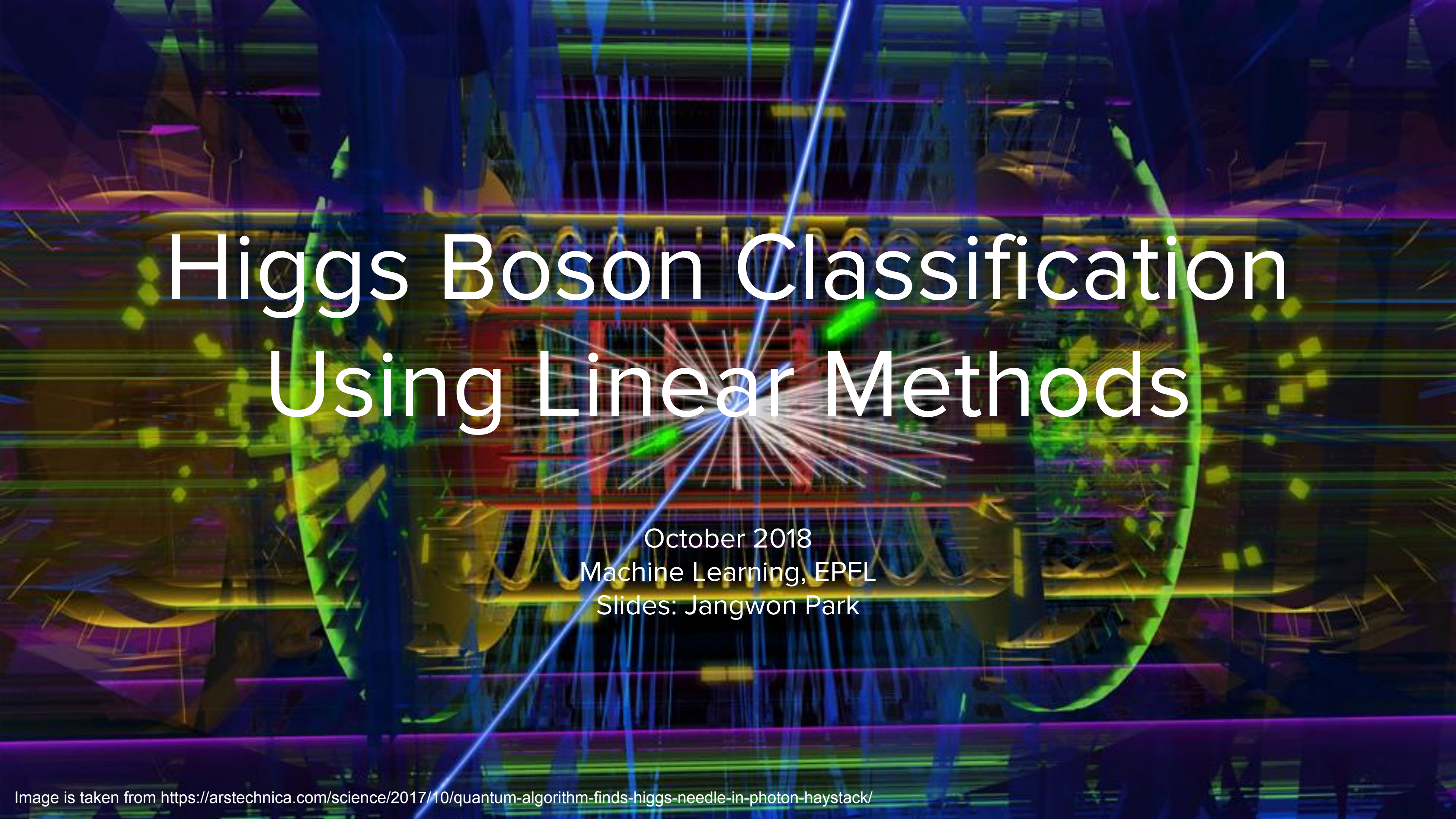


Higgs Boson Classification Using Linear Methods

October 2018
Machine Learning, EPFL
Slides: Jangwon Park

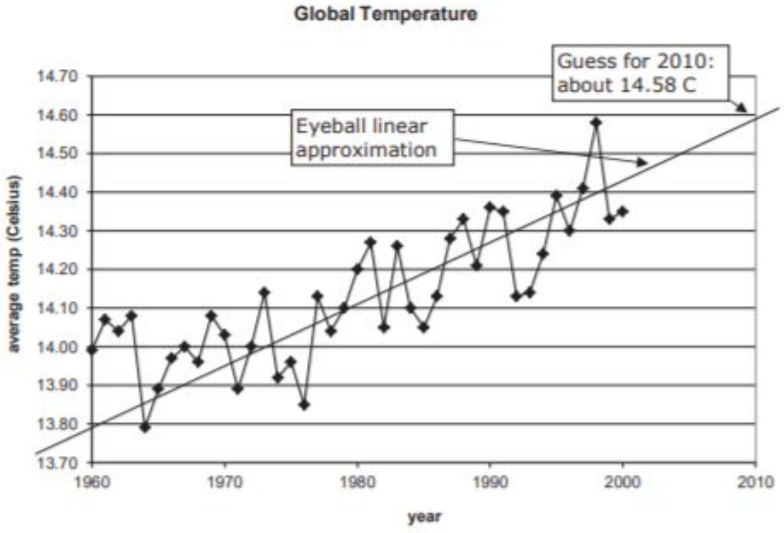
Image is taken from https://arstechnica.com/science/2017/10/quantum-algorithm-finds-higgs-needle-in-photon-haystack/

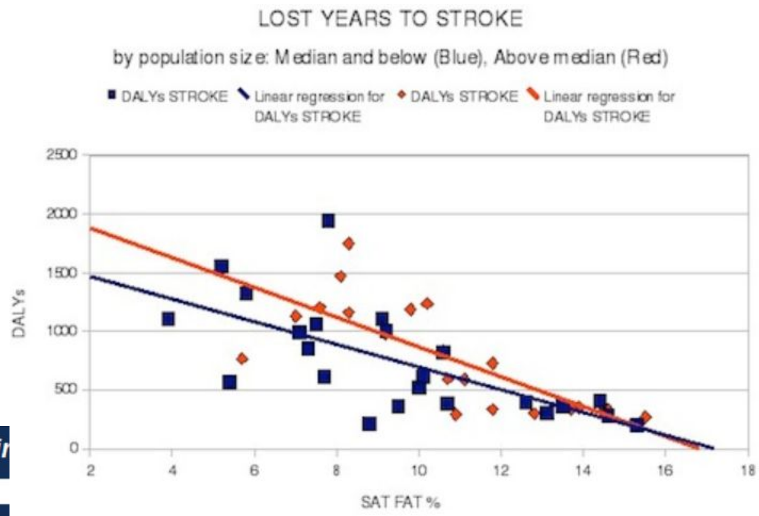
Introduction

- Identifying Higgs boson was first proposed as a Kaggle challenge in 2014.
- Re-introduced as an in-class project in a masters-level machine learning course at EPFL, Switzerland, in the context of **linear regression**.
- Linear regression is a fundamental to many more complex concepts in machine learning.
- Linear regression can be adapted for binary classification.
  - Predicted result >= 0.5: force the outcome to be 1
  - Otherwise 0.

Introduction

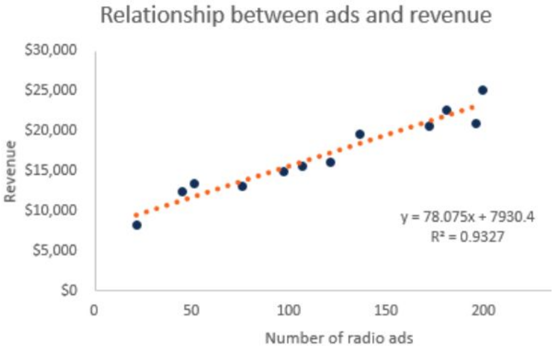
- Linear regression is still among the most used technique with many applications -- mastering it is a prerequisite for a data scientist.





hod #3: Simple Lir

	Radio ads	Revenue
	21	$8,350.0
	180	$22,755.0
	50	$13,455.0
	195	$21,100.0
	96	$15,000.0
	44	$12,500.0
	171	$20,700.0
	135	$19,722.0
	120	$16,115.0
	75	$13,100.0
	106	$15,670.0
Dec	198	$25,300.0
Totals	**1,391**	**$203,767.0**
Average	**116**	**$16,980.6**



Climate, business, epidemiology, ...

Dataset

- 30-feature particle accelerator data from CERN
- Training data: 250,000 events
- Test data: 550,000 events
- Labels = {Higgs boson = s, background = b}
- Evaluation criterion: accuracy in %

Problem Statement

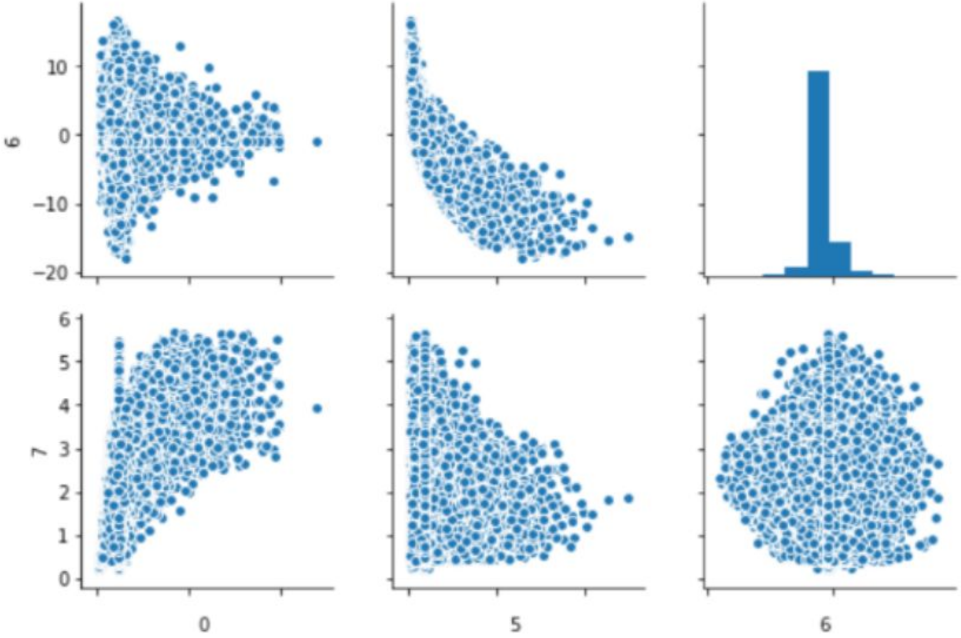
- Using linear methods, classify whether an event is Higgs boson (label = 1) or background (label = 0).

Exploratory Data Analysis

- Some features have a constant value of -999, indicating that they are “undefined”.
- Turns out that these features are sometimes “undefined” depending on the value of a common feature called *PRI-jet-num*.
  - *PRI-jet-num* is a discrete feature that takes <u>only four values</u> {0, 1, 2, 3}
- Manual feature clustering can be done to **filter** the entire dataset into <u>four</u> instances:
  - One for each of the instance of *PRI-jet-num*.

Exploratory Data Analysis

- Possible interactions between some features as evident in pairwise scatter plot



Feature Processing Outline

1. Feature Clustering
2. Fifth Order Degree Expansion
3. Backward Selection
4. Interaction Terms
5. Forward Selection

Feature Clustering

- Create four **mutually exclusive** subsets of the dataset based on the instance of *PRI-jet-num*.
- Real number of features is less than 30.

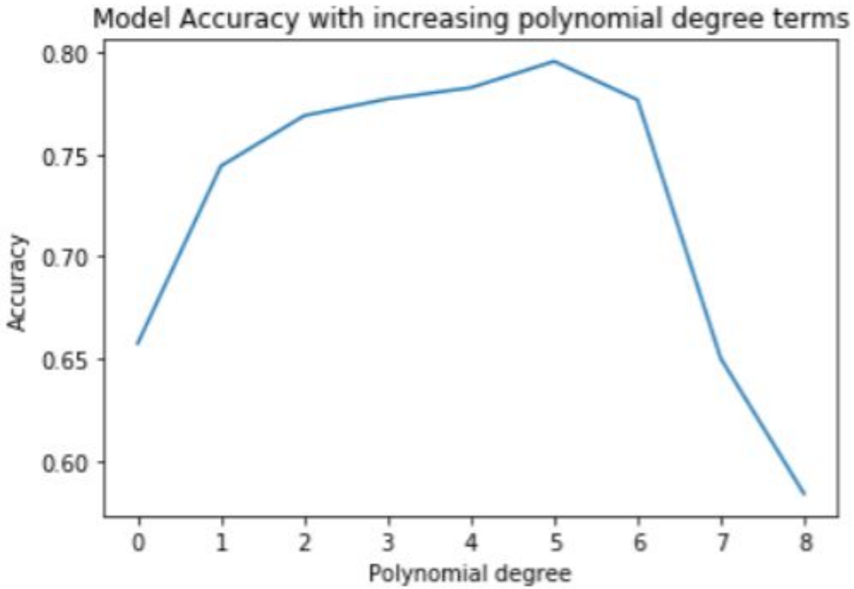
EventId	DER_de	DER_m	DER_pr	DER_le	PRI_jet_num	PRI_jet_	PRI_jet_	PRI_jet_	PRI_jet_	PRI_jet_	PRI_jet_	Label
100003	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	b
100004	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	b
100008	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	b
100010	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	b
100013	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	b
100014	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	b
100015	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	s
100017	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	s
100018	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	b
100019	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	b
100020	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	b
100021	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	b
100022	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	b
100024	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	b
100025	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	b
100036	-999	-999	-999	-999	0	-999	-999	-999	-999	-999	-999	s

Example:
All undefined features for *PRI-jet-num* = 0

For this subset, real number of features is 30 - 10 = 20 features!

Fifth Order Degree Expansion

- Degree expansion raises every feature to some power to capture non-linear relationships.
- **Bias-variance curve** shows that including up to fifth order terms improves model accuracy, but any more leads to overfitting.



Backward Selection

- New number of features is potentially: 5*30 = 150 features.
- Not all may be useful.
- Backward selection runs 10-fold cross validation while removing **one feature at a time**.
  - If the accuracy increases without the current feature in question, then it is removed.
  - Otherwise, it is kept.

Interaction Terms

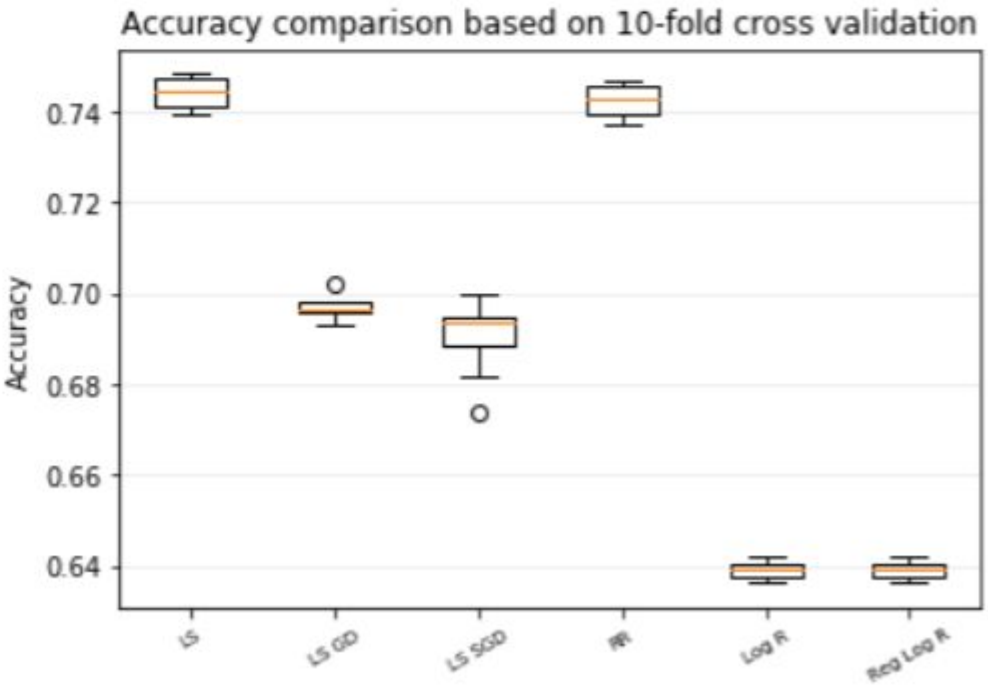
- It was clear from exploratory data analysis that certain features were non-linearly associated with each other. Which ones are useful?
- From the features from backward selection, we have **over 4,000** second order interaction terms => too many!
- Performing backward selection by first adding all thousands of terms will increase our computational cost exponentially.
- How can we add each interaction term progressively?

Forward Selection

- Add an interaction term if and only if the **model including it** has a **higher** accuracy based on 10-fold cross validation.
- Both backward and forward selection are <u>guaranteed to avoid overfitting</u>!

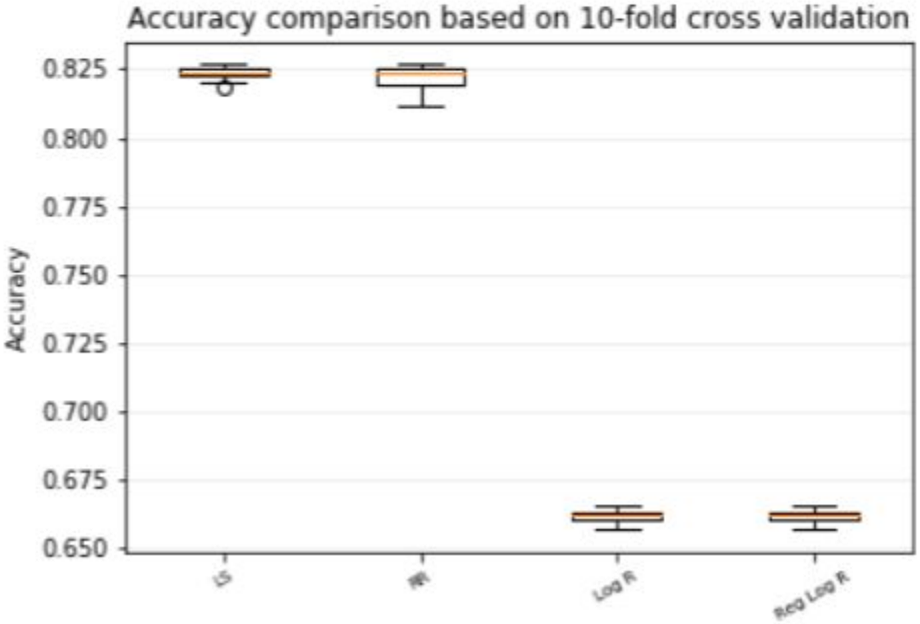
Initial Results

- The following shows the results of various models **prior** to any feature processing steps.
- Legend:
  - LS: least squares solution
  - LS GD: by gradient descent
  - LS SGD: by stochastic gradient descent
  - RR: ridge regression
  - Log R: logistic regression
  - Reg Log R: regularized logistic regression



Final Results

- Improvement from feature processing is evident.



Discussion

- Optimal model is least squares model with accuracy of 82.321%.
- Feature augmentation led to a nearly 8% improvement in accuracy.
- Ridge regression did not add any benefit compared to least squares solution.
  - Verifies that backward/forward selection is guaranteed to avoid overfitting.
- Logistic model did not appear to be a good choice in this particular application.
  - Even after tailored feature processing, it is much worse than linear regression.
- NOT performing feature clustering results in lower accuracy by about 2%.
- Though linear regression is a powerful tool, non-parametric methods such as neural networks will probably work better.