Practical introduction to Pandas + Scikit-learn via Kaggle problems

https://github.com/savarin/pyconuk-introtutorial

$$F = \mathbf{E}[F] + \int_0^T \mathbf{E}\left[D_t F | \mathcal{F}_t\right] \, dW_t$$

HACKERSHIP

1. Preliminaries

2. What is Machine Learning?

3. Next Steps

Machine Learning

vs Big Data


NETFLIX

Python

vs R

1. Preliminaries

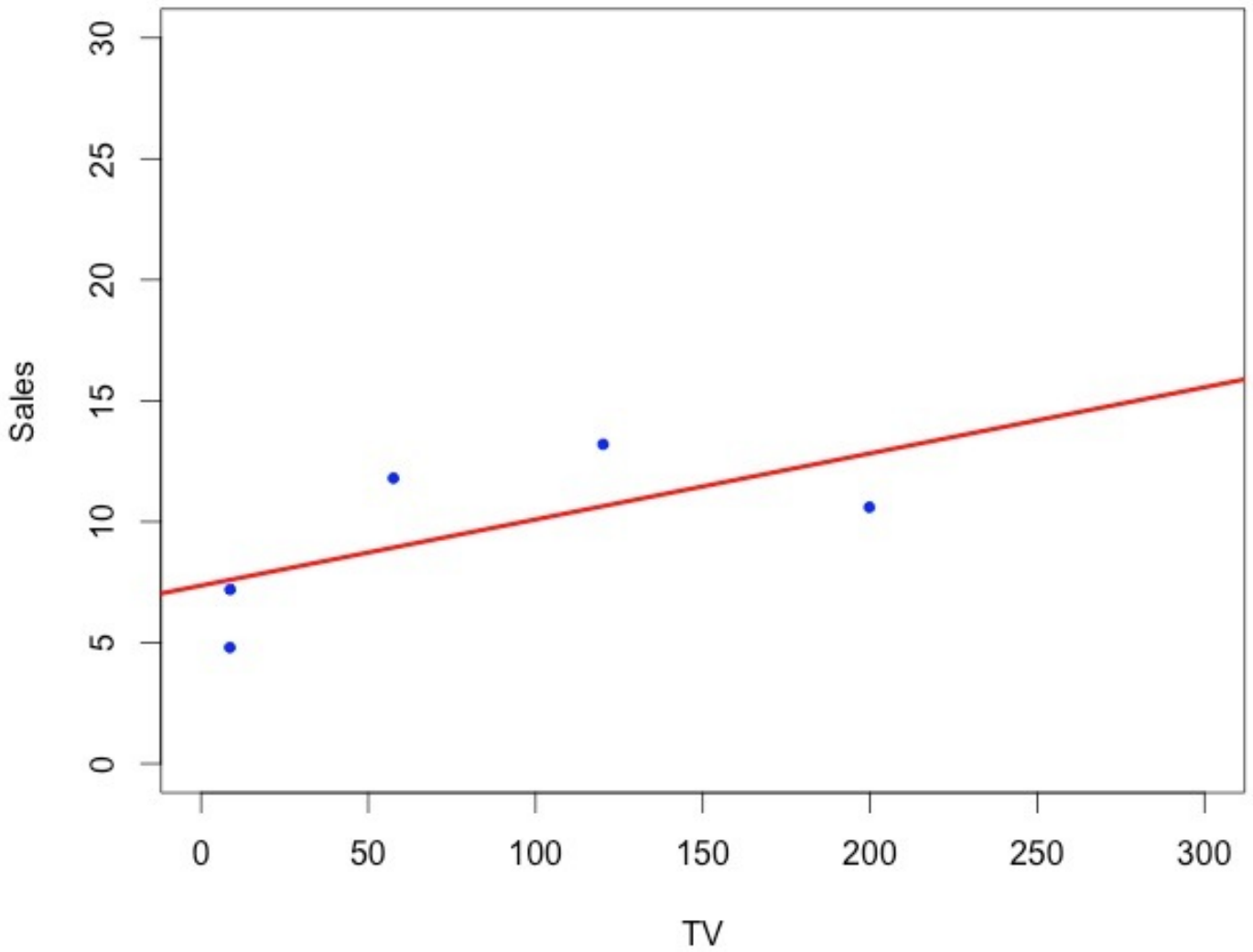
2. What is Machine Learning?

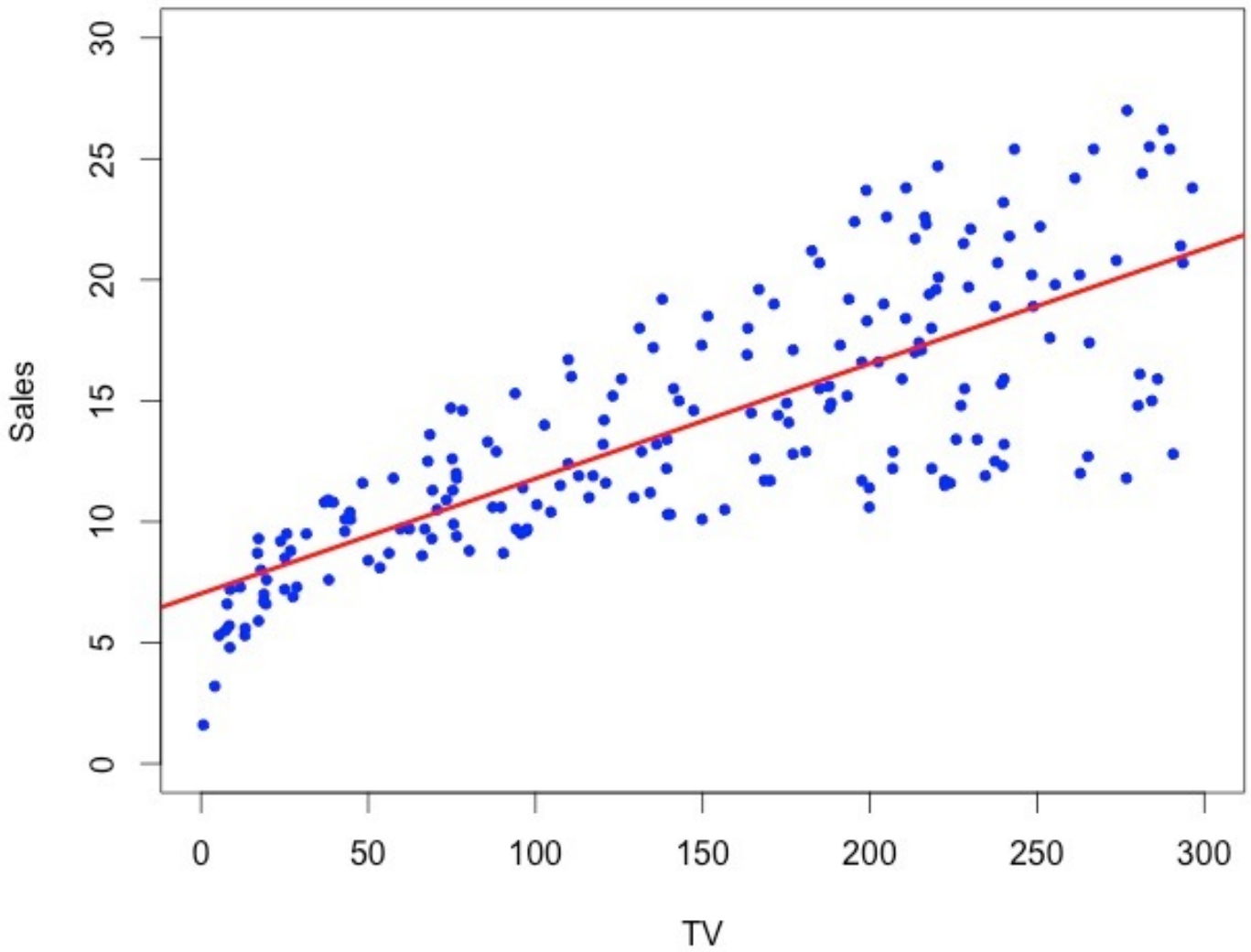
3. Next Steps

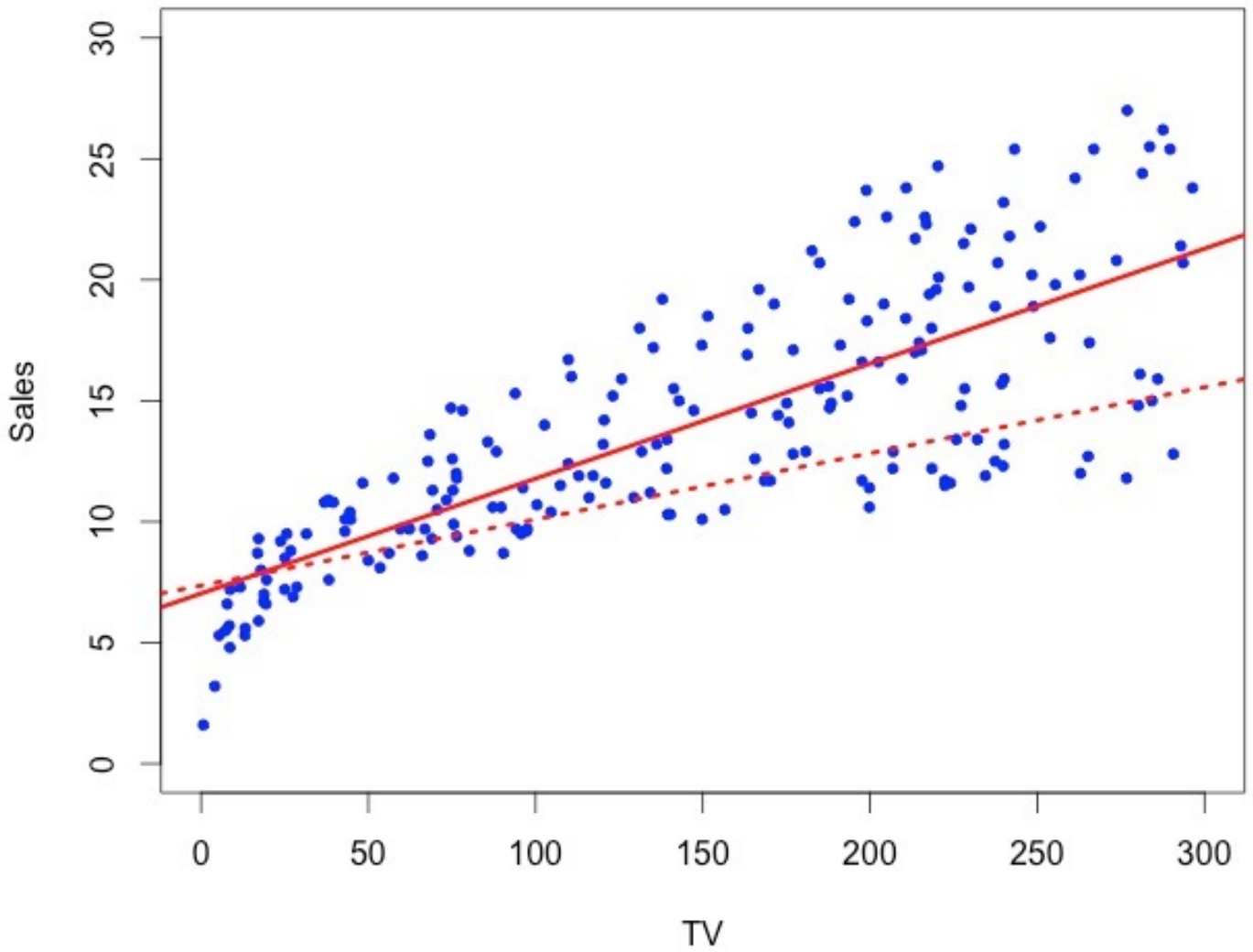
Machine Learning is about building programs with tunable parameters that are adjusted automatically so as to improve their behavior by adapting to previously seen data.

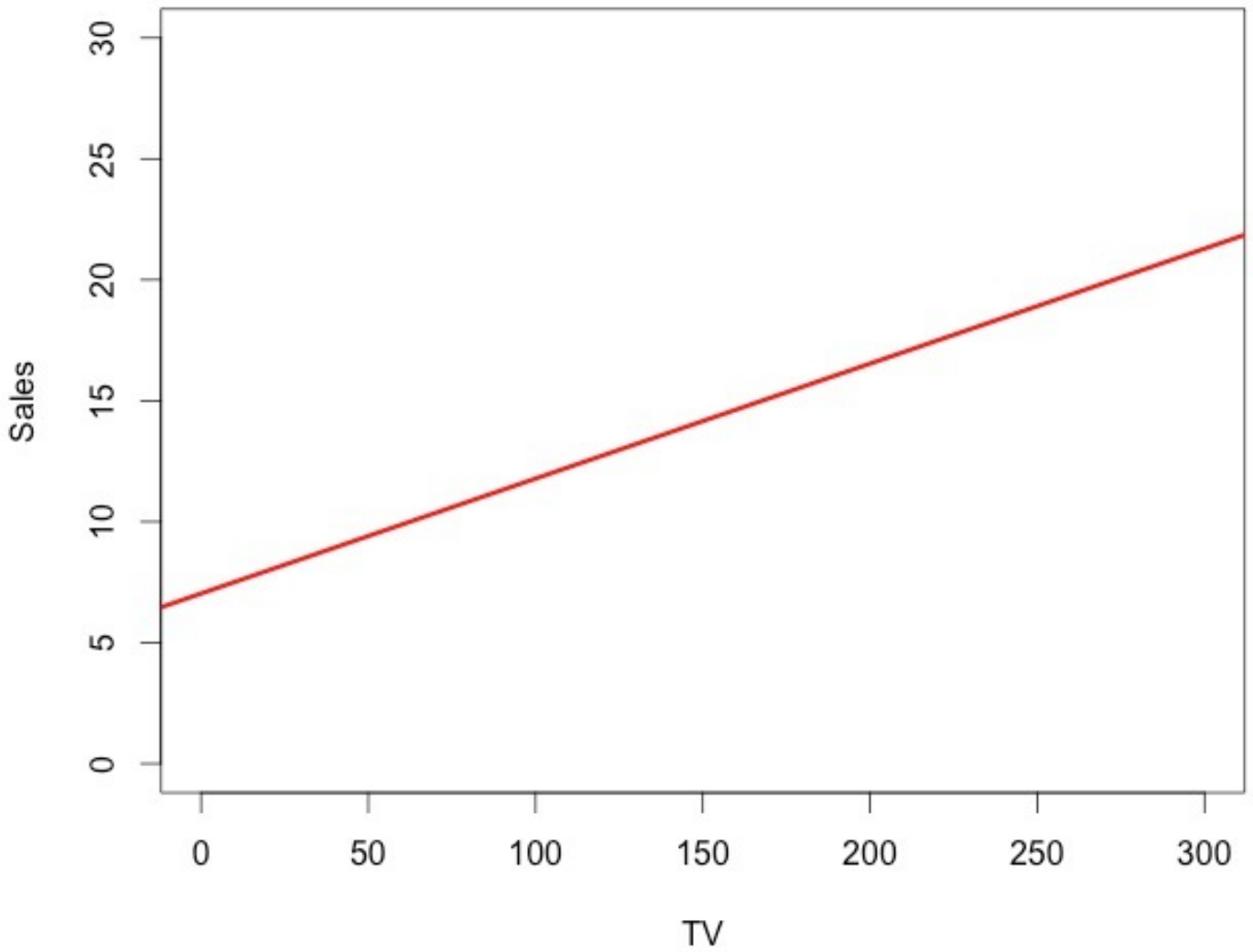
— Jake Vanderplas

> Machine Learning is about building programs with tunable parameters that are adjusted automatically so as to improve their behavior by adapting to previously seen data.
>
> — Jake Vanderplas


Sales
0
5
10
15
20
25
30
0
50
100
150
200
250
300
TV


Sales
0
5
10
15
20
25
30
0
50
100
150
200
250
300
TV


Sales
0
5
10
15
20
25
30
0
50
100
150
200
250
300
TV


Sales
0
5
10
15
20
25
30
0
50
100
150
200
250
300
TV

Active Competitions

American Epilepsy Society Seizure Prediction ...

Predict seizures in intracranial EEG recordings

60 days
119 teams
$25,000

criteo

Display Advertising Challenge

Predict click-through rates on display ads

5.1 days
730 teams
$16,000

Africa Soil Property Prediction Challenge

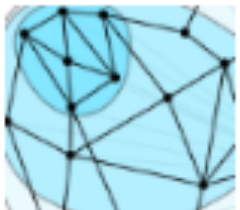
Predict physical and chemical properties of soil using spectral measurements

33 days
591 teams
$8,000

CIFAR-10 - Object Recognition in Images

Identify the subject of 60,000 labeled images

30 days
197 teams
Knowledge



Learning Social Circles in Networks

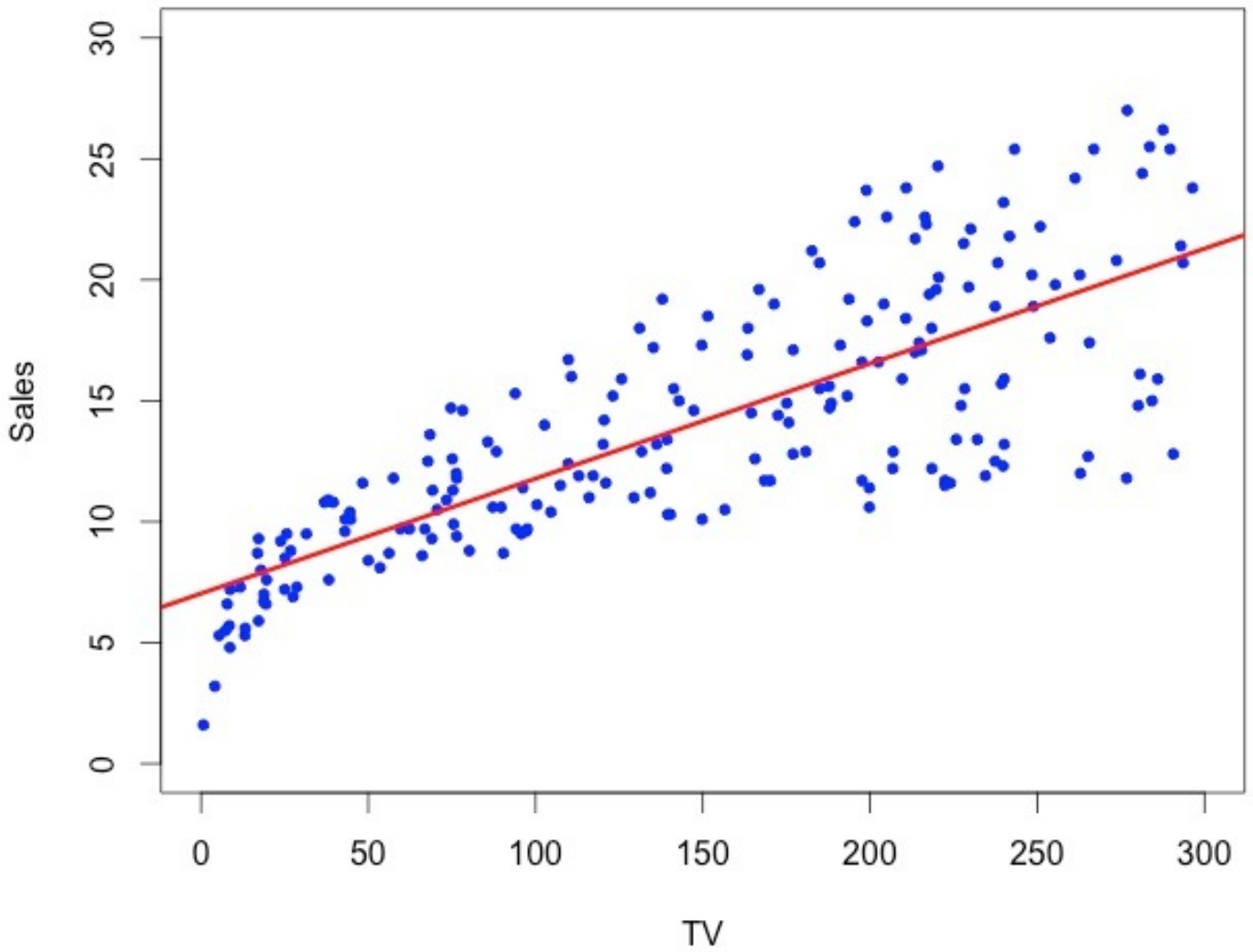
Model friend memberships to multiple circles

40 days
101 teams
Knowledge

Data Files

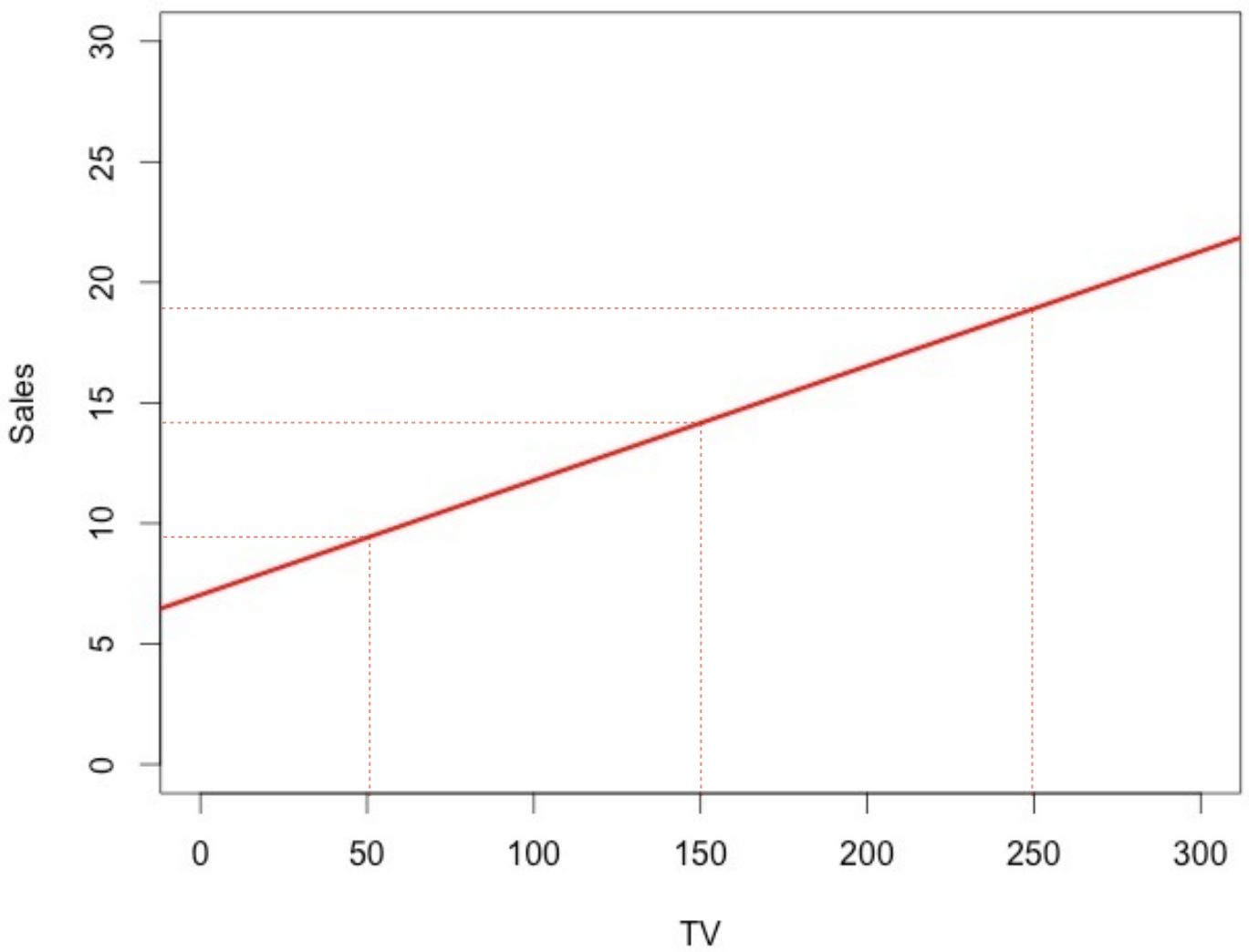
File Name	Available Formats
train	.csv (59.76 kb)
gendermodel	.csv (3.18 kb)
genderclassmodel	.csv (3.18 kb)
test	.csv (27.96 kb)
gendermodel	.py (3.58 kb)
genderclassmodel	.py (5.63 kb)
myfirstforest	.py (3.99 kb)

fit



TV	Sales
230.1	22.1
44.5	10.4
17.2	9.3
151.2	18.5
180.8	12.9
…	…

predict



TV	Sales
50	?
150	?
250	?
…	…

Presentation Format

The tutorial will start with data manipulation using pandas - loading data, and cleaning data. We'll then use scikit-learn to make predictions. By the end of the session, we would have worked on the Kaggle Titanic competition from start to finish, through a number of iterations in an increasing order of sophistication. We'll also have a brief discussion on cross-validation and making visualisations.

- Section 1-0 - First Cut.ipynb
- Section 1-1 - Filling-in Missing Values.ipynb
- Section 1-2 - Creating Dummy Variables.ipynb
- Section 1-3 - Parameter Tuning.ipynb
- Appendix A - Cross-Validation.ipynb
- Appendix B - Visualisation.ipynb

Time-permitting, we would cover the following additional materials.

- Section 1-4 - Building Pipelines.ipynb
- Section 1-5 - Final Checks.ipynb
- Section 2-1 - Support Vector Machines.ipynb
- Section 2-2 - SVM with Parameter Tuning.ipynb

A Kaggle account would be required for the purposes of making submissions and reviewing our performance on the leaderboard.

1. Preliminaries

2. What is Machine Learning?

3. Next Steps

Table Of Contents

10 Minutes to pandas

This is a short introduction to pandas, geared mainly for new users. You can see more complex recipes in the *Cookbook*

Customarily, we import as follows

```
In [1]: import pandas as pd

In [2]: import numpy as np

In [3]: import matplotlib.pyplot as plt
```

Object Creation

See the *Data Structure Intro section*

Creating a `Series` by passing a list of values, letting pandas create a default integer index

```
In [4]: s = pd.Series([1,3,5,np.nan,6,8])

In [5]: s
Out[5]:
0     1
1     3
2     5
3   NaN
4     6
5     8
dtype: float64
```

Agile Tools for Real-World Data

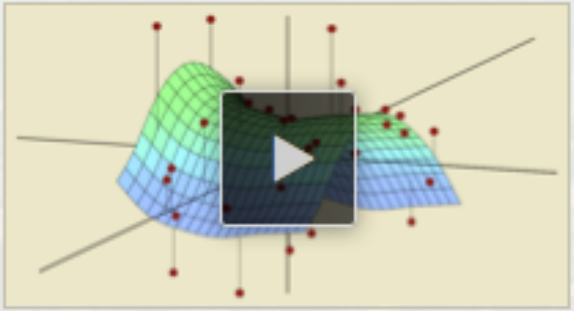
Python for Data Analysis

O'REILLY®

Wes McKinney

Statistical Learning

REGISTER FOR STATLEARNING



overview

ABOUT THIS COURSE

This is an introductory-level course in supervised learning, with a focus on regression and classification methods. The syllabus includes: linear and polynomial regression, logistic regression and linear discriminant analysis; cross-validation and the bootstrap, model selection and regularization methods (ridge and lasso); nonlinear models, splines and generalized additive models; tree-based methods, random forests and boosting; support-vector machines. Some unsupervised learning methods are discussed: principal components and clustering (k-means and hierarchical).

This is not a math-heavy class, so we try and describe the methods without heavy reliance on formulas and complex mathematics. We focus on what we consider to be the important elements of modern data analysis. Computing is done in R. There are lectures devoted to R, giving tutorials from the ground up, and progressing with more detailed sessions that implement the techniques in each chapter.

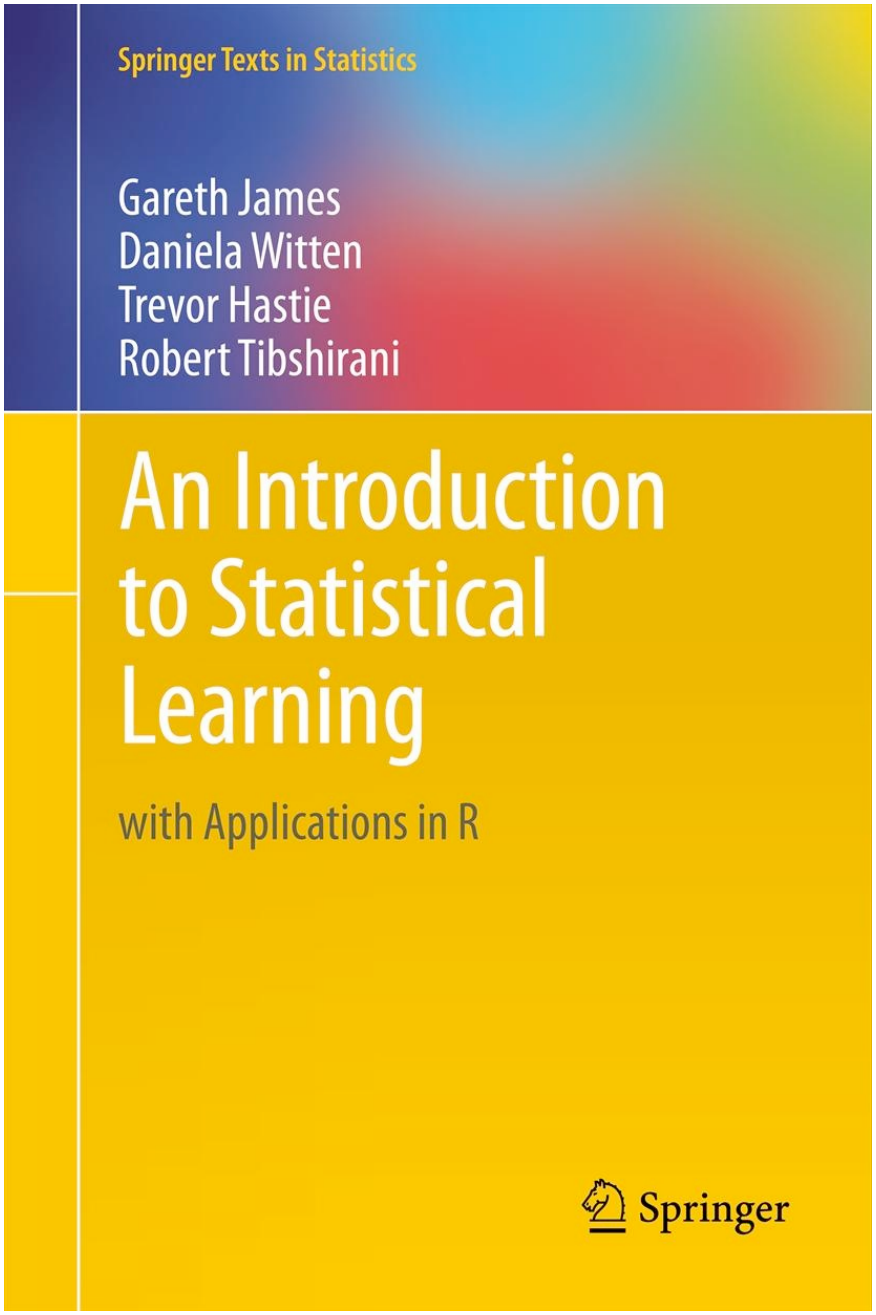
The lectures cover all the material in An Introduction to Statistical Learning, with Applications in R by James, Witten, Hastie and Tibshirani (Springer, 2013). As of January 5, 2014, the pdf for this book will be available for free, with the consent of the publisher, on the book website.

Course Number	**StatLearning**
Classes Start	**Jan 20, 2014**
Classes End	**Apr 04, 2014**

OUR RESEARCH COMMUNITY

Stanford University pursues the science of learning. Online learners are important participants in that pursuit. The information we gather from your engagement with our instructional offerings makes it possible for faculty, researchers, designers and engineers to continuously improve their work and, in that process, build learning science.

By registering as an online learner, you are also participating in research...


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer