Introduction to Statistical Learning

Vahid Partovi Nia

YCBS 255: Applied Computational Statistics

23 October 2018

Outline

Statistical Learning

- Simple (predictive)
- Interpretable (transparent box)
- Fast to train (big data)
- Works in wide variety of real problems (practical)
- Easy to adapt (generalizable)
- Building block of neural networks (deep learning)

Learning

- 90% Supervised learning: relate a predicting variable y to some other measured variables $x_j, j = 1, \ldots, p$
- 10% Unsupervised learning: data grouping using some measured variables x.

Why Python?

- is popular
- codes are readable
- easy to learn
- open source
- combines machine learning with data analysis
- many IDEs specially jupyter

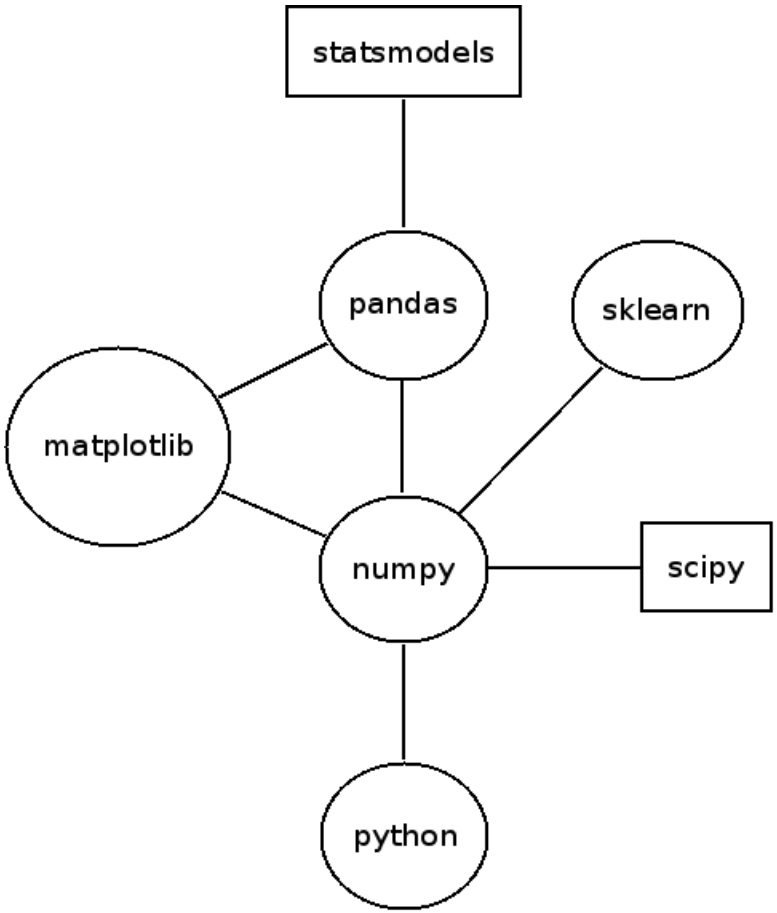
http://github.com/vahidpartovinia/ycbs255

Introduction to Python

Important libraries

- numpy: arrays, linear algebra, random numbers, numerical methods.
- pandas: data analysis, data visualization, data frames, statistics, basic statistical models.
- matplotlib: plots and data visualization.
- sklearn: machine learning algorithms.
- scipy: advanced numerical methods (extended version of numpy).
- statsmodels: advanced statistical models (extended version of pandas).

Schematic


statsmodels
pandas
sklearn
matplotlib
numpy
scipy
python

Numpy

Random vs Deterministic

Center as a predictor

We are willing to predict sales

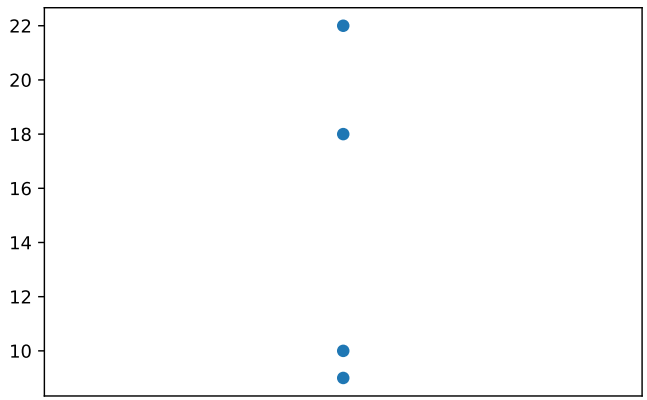
$$y_1 = 22, \quad y_2 = 10, \quad y_3 = 9, \quad y_4 = 18$$

For prediction a probabilistic model is required.

$$y_i = \beta_0 + \varepsilon_i.$$

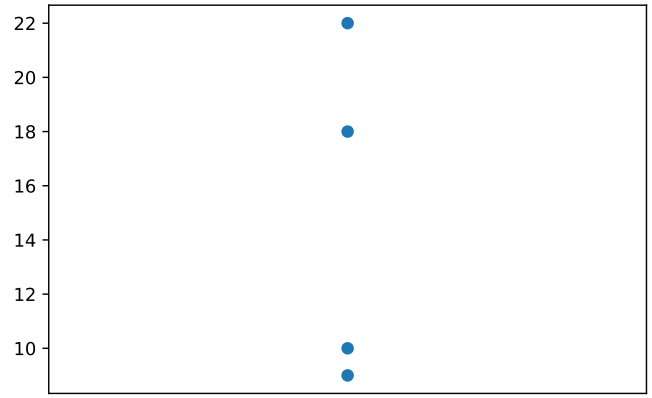
What is a good predictor?

$$y_1 = 22, \quad y_2 = 10, \quad y_3 = 9, \quad y_4 = 18$$



$$\begin{aligned} 22 &= \beta_0 + \varepsilon_1 \\ 10 &= \beta_0 + \varepsilon_2 \\ 9 &= \beta_0 + \varepsilon_3 \\ 18 &= \beta_0 + \varepsilon_4 \end{aligned}$$

Scale and location



$$S_1(\beta_0) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0)^2$$

$$S_2(\beta_0) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \beta_0|$$

What is mean?

$$
\begin{aligned}
4S_1(\beta_0) &= (22-\beta_0)^2 + (10-\beta_0)^2 \\
&\quad +(9-\beta_0)^2 + (18-\beta_0)^2 \\
\frac{dS_1(\beta_0)}{d\beta_0} &= -2(22-\beta_0) - 2(10-\beta_0) \\
&\quad -2(9-\beta_0) - 2(18-\beta_0) = 0
\end{aligned}
$$

What is mean?

$$
\begin{aligned}
4S_1(\beta_0) &= (22-\beta_0)^2 + (10-\beta_0)^2 \\
&\quad +(9-\beta_0)^2 + (18-\beta_0)^2 \\
\frac{dS_1(\beta_0)}{d\beta_0} &= -2(22-\beta_0) - 2(10-\beta_0) \\
&\quad -2(9-\beta_0) - 2(18-\beta_0) = 0 \\
\beta_0 &= \frac{1}{4}(22+10+9+18)
\end{aligned}
$$

```
> (22+10+9+18)/4
14.75
```

```
> import numpy as np
> y = np.array([22, 10, 9, 18])
> np.mean(y)
14.75
```

Which one?

Matplotlib Pandas