Business Intelligence Techniques and Applications

Session 2. Supervised Learning and Linear Regression

Renyu (Philip) Zhang

1

Data Visualization

- Purpose of data visualization:
  - Explore
  - Communicate

2

2

Anscombe's Quartet

X1	Y1	X2	Y2	X3	Y3	X4	Y4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

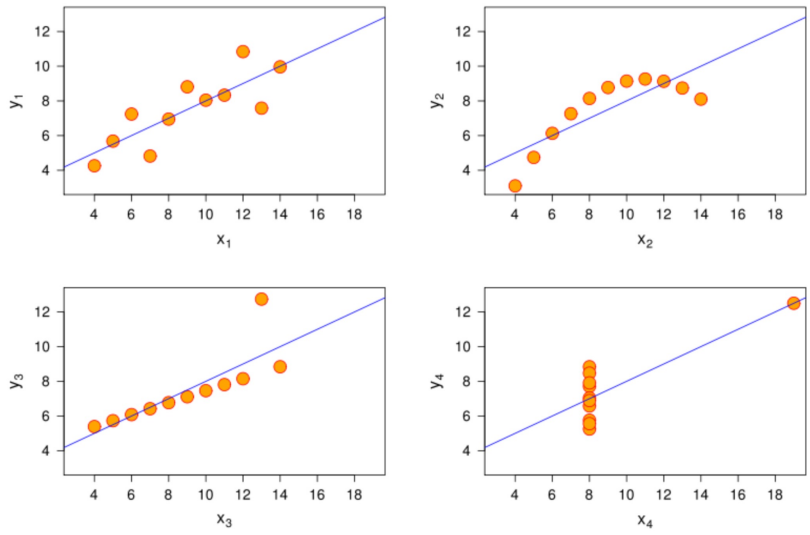
- Mean of X: 9.0
- Variance of X: 11.0
- Mean of Y: 7.5
- Variance of Y: 4.12
- Correlation between X and Y: 0.816
- Linear regression result:

  Y = 3.0 + 0.5*X

3

3

Anscombe's Quartet: Visualization



4

4

Module 1: Predictive Modeling (Applied Machine Learning)

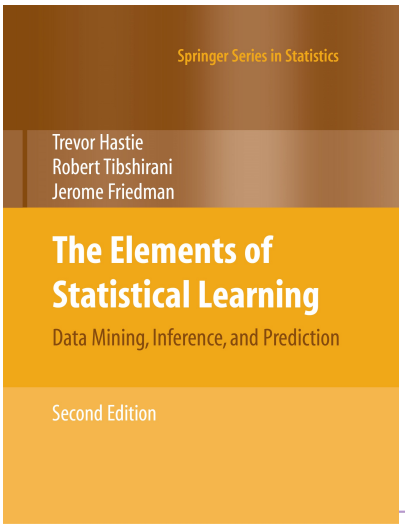
- Supervised learning
  - Regression: Linear regression, regression tree/forest
  - Classification: Logistic regression, naïve Bayes, k-nearest neighbors, tree-based models, neural nets
  - Bias-variance tradeoff, generalization error
- Unsupervised learning
  - Clustering: k-means, hierarchical clustering
  - Dimensionality reduction: Principal component analysis
- Deployment and implementation in comprehensive applications
  - Alchemy vs. science vs. engineering: Automated machine learning

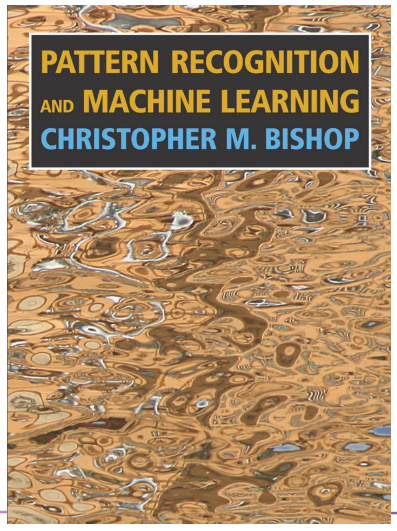
5

5

References

Both available at: https://github.com/DSME6756/BA-W2021/tree/main/Analytics%20References





6

6

Homework

- Problem Set 2, due at 9:30AM, December 23, Thursday
  - Submit the solutions with code in a Jupyter Notebook on Blackboard.

- Next week: Linear regression and logistic regression.