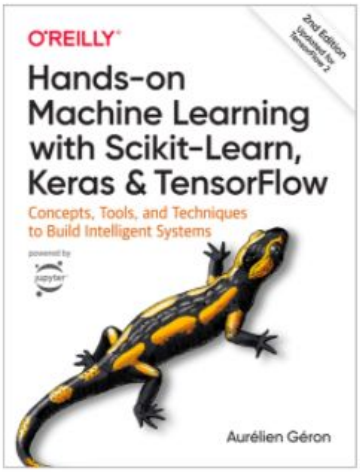
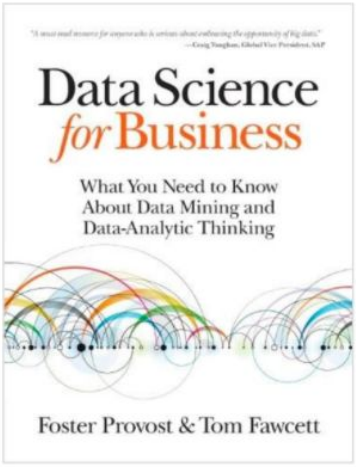
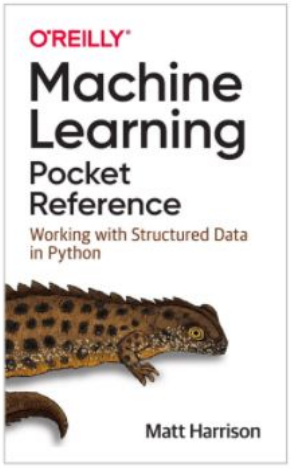
Day1_Links







Introduction to Data Science:

1. What is Data Science: https://youtu.be/xC-c7E5PK0Y
2. What is ML: https://youtu.be/HcqpanDadyQ
3. Reading:
    a. HOML, Ch. 1, The Machine Learning Landscape (33 pages)
    b. FES, Ch. 1, Introduction (20 pages)
    c. DSB, Ch. 1, Introduction to Data Analytic Thinking (18 pages)
    d. DSB, Ch. 2, Business Problems and Data Science Solutions (22 pages)
4. Mathematics for Data Science:
    a. https://www.dataquest.io/blog/math-in-data-science/
    b. https://towardsdatascience.com/mathematics-for-data-science-e53939ee8306

The Data Science Lifecycle

1. Reading:
    a. MLPR, Ch. 2, Overview of the Machine Learning Process
    b. FES, Ch. 3, A Review of the Predictive Modeling Process
2. Project Examples
    a. HOML, Ch. 2, End-to-End Machine Learning Project (code: link)
    b. FES, Ch. 2, Illustrative Example: Predicting Risk of Ischemic Stroke
    c. MLPR, Ch. 3, Classification Walkthrough: Titanic (code: link)
3. Cross Validation Techniques
    a. https://machinelearningmastery.com/k-fold-cross-validation/
    b. FES, Sections 3.4.1-2
    c. https://scikit-learn.org/stable/modules/cross_validation.html

Discussion Question:

- How do we calculate correlation and covariance?
- In what ways are they similar? In what ways to they differ?
- What is each metric actually measuring?
- Outline a scenario in which the use of a covariance metric would prove to be preferable to the use of a correlation metric.

Before discussing the differences between the two concepts, it's important to mention that their is in fact a strong relationship borne out by the mathematical equation that expresses them as merely being a variation of another. Specifically, $\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$. So while, the correlation is more generally thought of as a measure of the relationship between X and Y, giving us a standardized bound between -1 and +1 that describes the linearity, and the R^2 (coefficient of determination), the formula shows that it is very much tied to covariance.

Covariance, unlike correlation, is not standardized and changes as the magnitude changes. This can be seen from the formula: $Cov(aX, bY) = ab\,Cov(X, Y)$. Also, from the typical notation, σ_{XY} it should be clear that it is very much not only a measure of a joint distribution and the relationship thereof, but that it is very similar to the standard deviation, which is a measure of how much change, or variance, a variable has.

The simplest way that I usually think about covariance, is by looking at the var-cov matrix that might be used in data science, or, with a specific application in mind, in finance. When optimizing a portfolio, one uses the var-cov matrix to find optimal portfolio weights. Typically, there are several variations such as the single-index model and the constant correlation model, which highlight the nuanced relationship we are dealing with. That is, the diagonal of the matrix is just the variance, while all other positions included the covariance i.e. they are a measure of "the relationship of variance with respect to a joint distribution".

The calculation of covariance is $Cov(X, Y) = E[(X - EX)(Y - EY)]$, which again, shows how closely related to the variance this is, with the formula essentially the same except for the second part being a Y instead of an X.

Both covariance and correlation measure the relationship between the variables, but if the variables were on different scales, we would prefer correlation, as it is a standardized measure, but less in tune with the specific nuances of the data in terms of change, versus the covariance which works well with data that is on a similar scale. So while an optimized portfolio would naturally do well with a var-cov matrix, it is otherwise difficult to compare covariance, as scale and other factors vary greatly.