This book “Python for Probability, Statistics and Machine Learning” ’s purpose is to give some python programmers a tutorial on how to use python libraries like numpy, matplotlib, Panda, scipy, etc to perform probability evaluation and statistical tasks to reinforce machine learning. It can be break up into four parts: it starts with a brief introduction of some main scientific python libraries, Jupiter Notebook and developing environments. Next, it brings in random variables, various distribution and sampling methods in python. Then, it includes some python statistical module and gives a tutorial on how to test the hypothesis, evaluate confidence levels and perform linear regression in python. At last, it gives an overview of machine learning concepts like decision tree, neuron networks and deep learning, general steps in building a machine learning model with some simple neuron networks, and ways to train and test models.

This book used python3.6+ for its examples. The author expects the reader’s knowledge of simple loop, list, basic algebraic operation, matrix operation, python math operations, and input/output and import system are some python built-ins. Since python 3 has good documentation online, it is not very hard to pick up these parts. The book introduces the use of a more advanced python library by small blocks of codes with brief comments and its output. Overall, it is a book for intermediate or advanced python programmer.

For the math part, since the author is trying to give a brief background on every probability and statistic concepts appear, this book is dense with math formulas. Like the fact that it is not a book just focus on probability and statistic, there is no space for giving more examples of datasets and problem scenarios, and therefore the math interpretation is not very easy to understand and applied if the reader is not very familiar with advanced statistic methods and linear algebra.

However, besides a lot of complicated formulas given, each variable involved within the formula is clearly defined. And the alphabetical notion in the formula in the math formula is used to name the variable, as a result, increase the readability of the example codes. After a result is printed, the author usually clarifies what the number means, which makes sense in the context. However, there is still some use of constants that can be improved. For example, in the illustration of Chi- square distribution with cumulative distribution function in python, after calculating out the z score according to the formula, the author writes:

>>>.... >>> 1-statis.chi2(2).cdf(z)

Based on the previous example and information, it can be inferred that the “1-statis.chi2(2).cdf(z)” instead of “statis.chi2(2).cdf(z)” is probably seeking for a tail bound of Chi-square distribution with cumulative distributed function, but it can be clarified more in a comment. Moreover, since it passes a “2” as a parameter in the Chi-square method, according to the dataset it provides, there are 3 categories, so I assume this “2” is the degree of freedom. Again, it would be easier for readers to understand if the author adds one more line before:

>>> degree\_of\_freedom = 3-1 >>>1- statis.chi2(degree\_of\_freedom).cdf(z) #tail bound

For the machine learning part, the author makes a lot of good visual representation to explain the concepts. The mathematical methods mentioned in this part are exclusive rather than deep related to the previous chapters. Some general principles, such as linear regression may apply to some middle steps like calculating gradient descent, but different libraries and methods are used. It is good in the aspect for the reader to read the catalog and directly study the part needed, but if the reader is expecting to learn how to do a machine learning study integrating with various statistical methods, this book may not provide enough examples. The machine learning part involves some famous dataset, such as minst and make\_moon dataset, which is good for readers to find more related work for themselves, too. Generally, I feel the machine learning part is great for a beginner in machine learning and can be independently read from other parts.