Notes for ECE 20875 - Python for Data Science

Shubham Saluja Kumar Agarwal

January 21, 2025

These are lecture notes for Fall 2025 ECE 20875 by professor Aristides Carrillo at Purdue. Modify, use, and distribute as you please.

Contents

Introduction

Data has a multitude of definitions, but it originated from information, and is used to generate something, be it knowledge or beliefs. There are also several kinds of data, such as quantitative and qualitative, or physical and digital.
Data analysis helps us make decisions and take actions.
Data science has several branches, such as collecting and organizing data, making observations, visualizing trends, identifying similarities, making predictions, prescribing courses of action, and developing and accelerating algorithms.

Histograms

Count: number of elements in each bin of the histogram, and is denoted by x_k.

$$\sum_{k=1}^{n} x_k = m$$

where the sum of all bins is the total number of samples m.
Probability: probability of the occurrence of each bin, and is denoted by $\hat{p_k} = \frac{x_k}{\sum_l x_l}$.

$$\sum_k \hat{p_k} = 1$$

that is, the sum of all probabilities is 1.
Density: normalization of probability and bin width, denoted by $\hat{d_k} = \frac{\hat{p_k}}{w}$.

$$\sum_k w\hat{d_k} = 1$$

that is, the area under the probability curve is 1.
Note: the term "frequency" can be applied to both "count" and "probability".

Number of Bins

There are several parameters defining a histogram, including number of bins n, the width w, and number of samples m.
Bins do not need to all have the same width.
The selection of n and w can be done through a number of methods:

- $n = \sqrt{m}$
- $n = \lceil \log m \rceil + 1$
- $n = \lceil 2m^{1/3} \rceil$
- $w = 3.5\frac{\hat{\sigma}}{m^{1/3}}$

All of these methods have different considerations.
w should be selected to minimize the error of estimating a point. The Integrated Square Error (ISE), or the smoothing parameter is:

$$L(w) = \int (\hat{f_m}(x) - f(x)^2) dx$$

where $\hat{f_m}(x)$ is the density estimate of the histogram with m samples, while $f(x)$ is the true but unknown model.
Since there are unknowns in the above equation, we estimate as follows:

$$L(w) \approx J(w) + K = \frac{2}{(m-1)w} - \frac{m+1}{(m-1)w}(\hat{p_1}^2 + \hat{p_2}^2 + \cdots + \hat{p_n}^2) + K$$

where $\hat{p_k}$ are the individual bin probabilities and K is a constant.
Bin width is optimized by selecting the value that minimizes $J(w)$.
This can be approached through brute force methods. However, since n is finite, it can be better to run through values of n instead (using grid search).

Probability

Conduct an experiment, get an outcome. The outcome has a probability of occurring between 0 and 1.
The set of all possible outcomes is the sample space.
Sum of all probabilities of all outcomes is 1.
An event is a set of possible outcomes.