ECE595 Homework 1 Supplementary Notes

Guanzhe Hong

2019 Spring

Introduction

This document discusses some of the theoretical details of the tools you used in homework 1, including the concepts of random sampling, the histogram, and the kernel density estimator. We hope that you can gain a deeper understanding of the tools you used or were at least curious about in the homework.

Random Sampling

Consider a probability space $(\mathcal{S}, \mathcal{F}, P)$ which models a certain random experiment, where $\mathcal{S}$ is the sample space, $\mathcal{F}$ is the collection of outcomes (events), and P is the probability measure associated with $\mathcal{F}$. Moreover, consider a random variable $\boldsymbol{X} : \mathcal{F} \to \boldsymbol{X}(\mathcal{F})$, where $\boldsymbol{X}(\mathcal{F}) = \{x \in \mathbb{R} | \, x = \boldsymbol{X}(\xi), \xi \in \mathcal{F}\}$, i.e., $\boldsymbol{X}$ quantizes every outcome of the random experiment modelled by $(\mathcal{S}, \mathcal{F}, P)$ by assigning a real number to each of them. As you have learned before, the probability distribution of $\boldsymbol{X}$ is described by the probability mass function (PMF) if $\mathcal{F}$ is discrete, and by the probability density function (PDF) if $\mathcal{F}$ is continuous; these functions are commonly denoted as $f_{\boldsymbol{X}} : \boldsymbol{X}(\mathcal{F}) \to [0, 1]$, and described by

$$f_{\boldsymbol{X}}(x) = P(\boldsymbol{X} = x) = P(\{\xi \in \mathcal{F} | \boldsymbol{X}(\xi) = x\}) \tag{1}$$

(Sometimes the PMF is denoted as $p_{\boldsymbol{X}}$ instead of $f_{\boldsymbol{X}}$ to distinguish it from the PDF)

To obtain one sample from the probability distribution specified by $\boldsymbol{X}$, is to perform an instance of the random experiment that is associated with $\boldsymbol{X}$, and obtain a quantized outcome $x \in X(\mathcal{F})$ whose probability of occurrence is $f_{\boldsymbol{X}}(x)$.

Histograms

Let $X_1, X_2, ..., X_n$ be iid observations taking values in $[0, 1]$ with probability density function f. Note that the restriction to [0, 1] is not crucial in this discussion; we can always rescale the data to be on this interval, so feel free to use whatever interval your feel comfortable with for your data later on in this project.

To simplify our discussion below, we adopt the following notations and definitions:

1. Define the **bins** as the intervals

$$B_1 = [0, \frac{1}{m}), \, B_2 = [\frac{1}{m}, \frac{2}{m}), ..., \, B_m = [\frac{m-1}{m}, 1]$$

2. Define the **bin width** as $h = 1/m$
3. Define v_i as the number of observations falling in the i^{th} bin
3. Define $\widehat{p_i} = \frac{v_i}{n}$

We define the **histogram estimator** as:

$$\widehat{f}_n(x) = \begin{cases} \widehat{p}_1/h & \text{if } x \in B_1 \\ ... & \\ \widehat{p}_m/h & \text{if } x \in B_m \end{cases} \tag{2}$$

It can be written more concisely as

$$\widehat{f}_n(x) = \sum_{i=1}^{m} (\widehat{p}_i/h)\mathbb{I}_{B_i} \tag{3}$$

To gain some intuition about why the histogram is defined this way, suppose $x \in B_i$ and h is small,

$$\mathbb{E}[\widehat{f}_n(x)] = \frac{\mathbb{E}[\widehat{p}_i]}{h} = \frac{\mathbb{E}[v_i]/n}{h} = \frac{(n\int_{B_i} f(t)dt)/n}{h} \approx \frac{f(x)h}{h} = f(x) \tag{4}$$

(Reason for the third equality: recall the expectation value of a binomial random variable)

Note that we should distinguish between the histogram estimator and other common histograms we plot in practice; for instance, we often see the following unnormalized histogram in real life:

$$\widehat{g}_n = \sum_{i=1}^{m} v_i \mathbb{I}_{B_i}$$

which does not satisfy equality (4).

If you are curious about whether there exists a smooth (instead of discontinuous) estimator of the density f, please read on.

Kernel Density Estimation [1]

For this section, assume a set of iid observations $X_1, ..., X_n$ taking values in $\mathbb{R}$, sampled from a probability distribution with density f.

Recall that the histogram estimator is a discontinuous estimator of the true probability density of the available data; the kernel density estimator (kde) also estimates that probability density, but it is smooth. We present a few necessary definitions first.

Define a **kernel** as any smooth function $K : \mathbb{R} \to \mathbb{R}$ such that $K(x) \geq 0$ on $\mathbb{R}$, $\int_{\mathbb{R}} K(x)dx = 1$, $\int_{\mathbb{R}} xK(x)dx = 0$, $\sigma_K^2 \stackrel{\text{def}}{=} \int_{\mathbb{R}} x^2 K(x)dx > 0$. An example of a kernel is the standard Gaussian $K(x) = (2\pi)^{-1/2}\exp\{-x^2/2\}$.

Given kernel K, constant $h > 0$, we define the **kernel density estimator** (kde) of the dataset $\{X_i\}_{i=1}^n$ as the following:

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K(\frac{x - X_i}{h}) \tag{5}$$

Finally, we introduce the **integrated square error** function $L(\widehat{f}, f)$ that serves as a measure of how good the estimator $\widehat{f}$ approximate the true density f. It is defined as:

$$L(\widehat{f}, f) = \int_{\mathbb{R}} [f(x) - \widehat{f}(x)]^2 dx \tag{6}$$

There are a few important things to note:

(1) Notice that h controls the "width" of the kernel (the smaller the h, the thinner the kernel is). We call this number the **bandwidth** of the kde.

[1] This section is based on chapter 20 of *All of Statistics: A Concise Course in Statistical Inference* by L. Wasserman

(2) The kde $\widehat{f}$ is normalized, i.e., $\int_{\mathbb{R}} \widehat{f}(x)dx = 1$ (this can be shown with a change of variable)

(3) It can be shown that with the optimal choice of h, $\mathbb{E}[L(\widehat{f}, f)] \approx \frac{C}{n^{4/5}}$. For the histogram estimator with the optimal choice of bin width, $\mathbb{E}[L(\widehat{f}_{hist}, f)] \approx \frac{D}{n^{2/3}}$. This means the kde converges to the true density f faster than the histogram estimator does.

In practice, to estimate the optimal bandwidth h for the kde, same as the case for histograms, we make use of the cross-validation estimator of risk derived from $L(\widehat{f}, f)$:

$$\widehat{J}(h) \approx \frac{1}{hn^2} \sum_i \sum_j K^*(\frac{X_i - X_j}{h}) + \frac{2}{hn} K(0) \tag{7}$$

where $K^*(x) = K^{(2)}(x) - 2K(x)$, $K^{(2)}(z) = \int_{\mathbb{R}} K(z-y)K(y)dy$. Under the assumption that the true density f is bounded, and denoting h_n as the bandwidth chosen that minimizes $\widehat{J}(h)$, the following can be proven:

$$\frac{L(\widehat{f}_{h_n}, f)}{\inf_h L(\widehat{f}_h, f)} \xrightarrow{P} 1 \tag{8}$$

In Python, there are implementations of the kde method for univariate and multivariate data. See for example `scipy.stats.gaussian_kde` from the Scipy library, and `seaborn.distplot` from the Seaborn library. For further understanding on the topic, you can read chapter 20 of *All of Statistics: A Concise Course in Statistical Inference* by L. Wasserman.