Generate Handwritten Digits using Kernel Density Estimation

Hoang Tung Lam

December 2019

1 Introduction

In this assignment, you will have to implement a simple Kernel Density Estimator from scratch (using numpy) to learn a generative model for the MNIST digits dataset. With this generative model, new samples can be generated

2 Density Estimation

The relationship between the outcomes of a random variable and its probability is referred to as the probability density, or simply the "density.". Given a random variable, we are interested in the density of its probabilities. We might want to know things like the shape of the probability distribution. The problem is, we may not know the probability distribution for a random variable. We rarely do know the distribution because we don't have access to all possible outcomes for a random variable.

This problem is referred to as probability density estimation, or simply "density estimation," as we are using the observations in a random sample to estimate the general density of probabilities beyond just the sample of data we have available.

You probably know we can summarize density with histogram and identify a common probability distributions from it. Once identified, you can attempt to estimate the density of the random variable with a chosen probability distribution. This can be achieved by estimating the parameters of the distribution from a random sample of data. This process is refer as parametric density estimation.

But in some cases, a data sample may not resemble a common probability distribution or cannot be easily made to fit the distribution. This is often the case when the data has two peaks (bimodal distribution) or many peaks (multimodal distribution).

In this case, parametric density estimation is not feasible and alternative methods can be used that do not use a common distribution. Instead, an algorithm is used to approximate the probability distribution of the data without a pre-defined distribution, referred to as a nonparametric method. The most common nonparametric approach for estimating the probability density function of a continuous random variable is called kernel smoothing, or kernel density estimation, KDE for short.

3 Kernel Density Estimation

Kernel Density Estimation is a nonparametric method for using a dataset to estimating probabilities for new points. Let (x1, x2, . . . , xn) be a univariate independent and identically distributed sample drawn from some distribution with an unknown density f. We are interested in estimating the shape of this function f. Its kernel density estimator is

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - x_i\right) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

where K is the kernel — a non-negative function — and $h > 0$ is a smoothing parameter called the bandwidth. K_h is called the scaled kernel and defined as $K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right)$. Intuitively one wants to choose h as small as the data will allow; however, there is always a trade-off between the bias of the estimator and its variance.

A range of kernel functions are commonly used: uniform, triangular, biweight, triweight, Epanechnikov, normal, and others. The Epanechnikov kernel is optimal in a mean square error sense, though the loss of efficiency is small for the kernels listed previously, and due to its convenient mathematical properties, the normal kernel is often used, which means $K(x) = \phi(x)$, where ϕ is the standard normal density function.

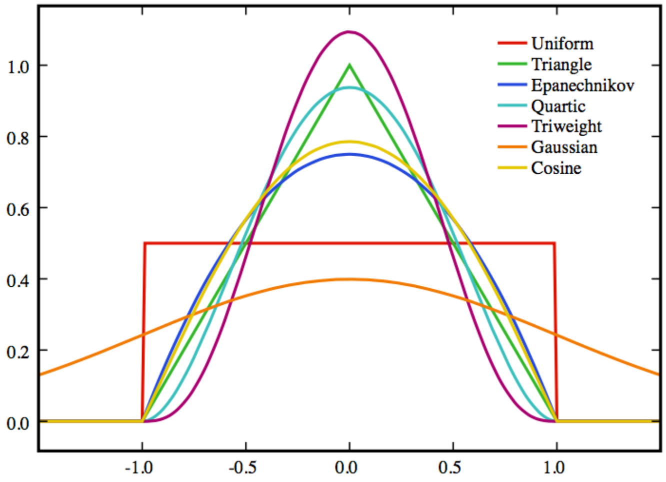

Figure 1: Visualization of different kernel functions

Since the choice of kernel function and the bandwidth of the kernel exhibits a strong influence on the resulting estimate. Please do your own research and experiment to choose the best kernel and bandwidth for this assignment and explain your reason.

4 Handwritten Digits Generator

Given the MNIST dataset of handwritten digits, you could use a KDE to learn a non-parametric generative model of the dataset in order to efficiently draw new samples from this generative model.

How do you pick the bandwidth for your KDE? What kernel do you think is suitable for this problem? Justify your answer.

5 Implementation

In this section you will have to implement a KDE from scratch with some form of bandwith estimator to learn a non-parametric generative model of the MNIST dataset.

5.1 Data

Please read and download the data from this link: `http://yann.lecun.com/exdb/mnist/`

5.2 Requirements

You can use tools to model the distribution but not the KDE itself (like the case of sklearn).

5.3 Task

- Implement a Kernel Density Estimator
- Implement a bandwidth estimator
- Learn some generative models with different kernel and bandwidth
- Sample learned KDE to generate sample, compare result of different setting.