Machine Learning

Lab 2: Density estimation

1 Overview

In this lab we will consider the problem of estimating, using a collection of observations, the probability of generating a particular future observation. You will use two approaches to this problem: parametric and non-parametric density estimation. In the former case, you will look at the salaries for employees of the city of Chicago. Assuming that the salaries within each department are sampled from a Gaussian distribution, you will learn the underlying model. The second data set we will consider is crime in the same city. You will generate a crime "heat map" of the city for different types of crimes and in different years to better understand where, and when, crime takes place. All data used here is publicly available at `data.cityofchicago.org`.

2 Provided Resources

- `gaussfit.m` - A placeholder function which will learn the parameters for a Gaussian density function fit to the provided data.
- `kde.m` - A placeholder function which will evaluate the density function for data using Gaussian kernel density estimation.
- `kdemap.m` - Helper function which calls `kde.m` in order to generate a map of crime over the city of Chicago. For example, `imagesc(kdemap(lat, lon, 0.01, 100))` will return a 100-by-100 entry grid sampling of the kernel density estimate using all crimes and a Gaussian kernel of standard deviation 0.01. Note that there are over 5 million crimes, so don't expect to be able to run this using all the data.
- `crimes.mat` - Data file containing crimes in the city of Chicago with latitude, longitude, and the type of crime. Contains: `lat` and `lon` - The latitude and longitude of each of 5.6 million crimes. `type` - For each crime, which category it falls into. `year` - The year in which each crime took place. `types` - Text names for each type of crime, and the index it corresponds with.
- `employees.mat` - Data file containing the salaries for employees of the city of Chicago, along with the department in which they work. Contains: `dept` - integers indicating the department for which each employee works, one for 32,160 total employees. `depts` - a struct which has text names for each of 35 departments. `sal` - the annual salary in dollars for each employee.

3 Guide

3.1 Gaussian MLE

Consider a multivariate Gaussian distribution with *isotropic* covariance matrix, $\Sigma = \sigma^2 I$. Note its expression is derived from the standard multivariate Gaussian distribution by assuming isometric covariance.
For a collection of i.i.d. data points x_i, the joint probability is computed as the product of individual probabilities:

$$\prod_i \frac{1}{(2\pi\sigma^2)^{D/2}} e^{-\frac{||x_i-\mu||_2^2}{2\sigma^2}}.$$

Given a data set, the *maximum likelihood estimators* for this distribution can be worked out exactly, and the solution is given by

$$\hat{\mu} = \frac{1}{N}\sum_i x_i$$

and

$$\hat{\sigma} = \sqrt{\frac{1}{DN}\sum_i ||x_i - \hat{\mu}||_2^2}.$$

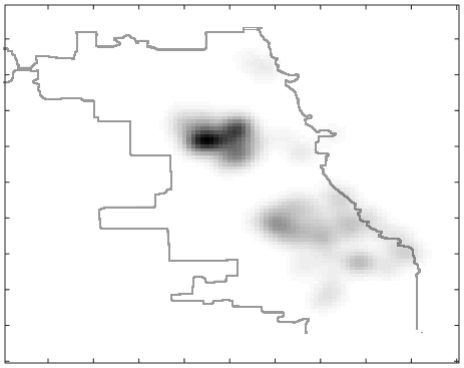
1. Complete and turn in the function `gaussfit.m`. This function should learn the maximum likelihood estimates for mean μ and variance σ^2 of a collection of data points, assuming the isotropic Gaussian model, above.
2. Load the file `employees.mat` containing employee salaries and their respective department assignment.
3. For each department use `gaussfit.m` to learn the parameters of the assumed underlying Gaussian distribution (in this case, the observation space is 1D). Hint: the salaries for individuals only in the health department (number 13) can be found using `sal(dept == 13)`.
4. Which departments have salary distributions with the highest and lowest mean salary?
5. Which departments have the greatest and least variance in salary? Anything special happening?

3.2 Kernel density estimation

1. Complete and turn in the function `kde.m`. This uses kernel density estimation with a Gaussian kernel of given standard deviation to sample an estimated probability distribution. Given the N observation vectors x_i and a standard deviation σ, this function should calculate for each sampling point z the quantity

$$p(z) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{||x_i - z||_2^2}{2\sigma^2}\right).$$

2. Load the file `crimes.mat` containing years of criminal activity in the city of Chicago.
3. Using the helper function `kdemap.m`, test your code by generating a heatmap of gambling crimes (crime index 15) for the year 2014 with Gaussian kernels of standard deviation $\sigma = 0.01$. The latitude of these crimes can be found for example using `lat(type == 15 & year == 2014)` (similarly for longitude). Your result should look similar to this (the approximate city boundary was added here to demonstrate the region shown by `kdemap.m`):



4. Look at gambling crimes from the year 2001 to 2014.
5. How has the distribution of gambling crime changed over time?
6. Generate/turn in a Gaussian kernel density estimate for interference with an officer (crime type 1) in the year 2014.

4 Extra

1. The modeling of salaries above allows for the city of Chicago to theoretically charge people to work for them! What does this say about the model? How could we change the model to fix this?

2. Can you think of an analogous issue the kernel density estimation suffers from? Hint: Chicago rests on the perimeter of lake Michigan.