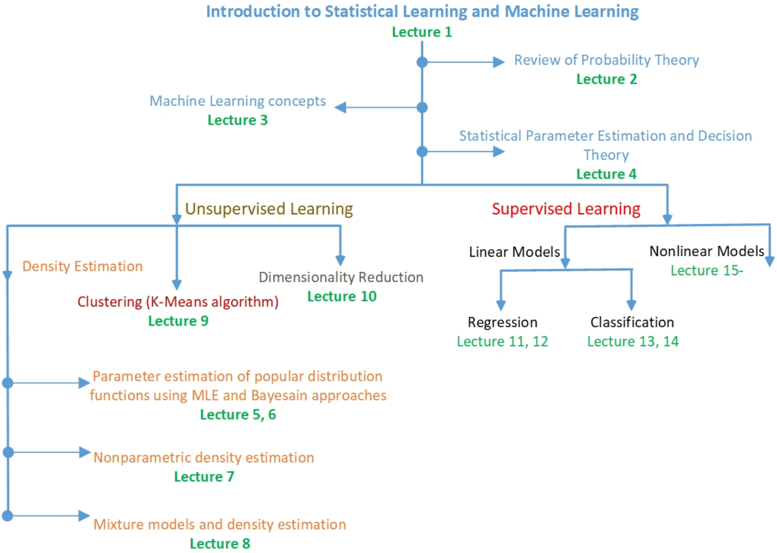
Statistical Learning and Machine Learning

Lecture 7 - Probability Distributions 3

September 18, 2021

Course overview and where do we stand



Introduction to Statistical Learning and Machine Learning
Lecture 1

Review of Probability Theory
Lecture 2

Machine Learning concepts
Lecture 3

Statistical Parameter Estimation and Decision Theory
Lecture 4

Unsupervised Learning

Supervised Learning

Density Estimation

Clustering (K-Means algorithm)
Lecture 9

Dimensionality Reduction
Lecture 10

Linear Models

Nonlinear Models
Lecture 15-

Regression
Lecture 11, 12

Classification
Lecture 13, 14

Parameter estimation of popular distribution functions using MLE and Bayesain approaches
Lecture 5, 6

Nonparametric density estimation
Lecture 7

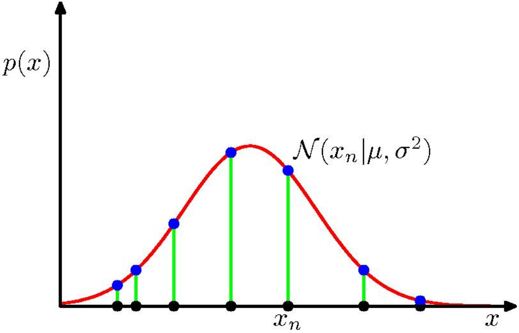
Mixture models and density estimation
Lecture 8

Objectives of the lecture

- Introduction to Gaussian mixture models
- Nonparametric density estimation techniques
    1. Histogram method
    2. Kernel density estimators
    3. Nearest-neighbour method

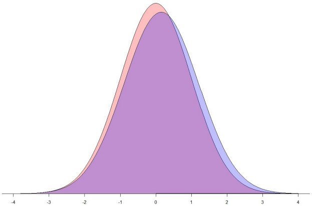
Density Estimation I

- Suppose that we have a set of data points $\{x_1, \ldots, x_N\}$ which correspond to (random) measurements of a process.
- Next, assume that the process is governed by the underlying probability density function $p(x)$. The value of $p(x)$ indicates how probable is to obtain the measurement x.



Density Estimation II

- The problem of estimating the probability distribution $p(x)$ given a set of data points $x_1, \ldots, x_N$ is called density estimation.
- The density estimation problem is ill-posed as there are infinitely many probability distributions that could given rise to the specific set of data points $\{x_1, \ldots, x_N\}$.



- To solve that problem, we take an *informed decision* regarding the probability distribution that generated the data $\{x_1, \ldots, x_N\}$.

Density Estimation III

- **Example**: A widely used probability distribution in machine Learning is the Gaussian distribution: $\mathrm{p(x)} = 1\frac{}{(2\pi\sigma^2)^{1/2}\; exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}}$ The Gaussian distribution, denoted by $\mathcal{N}(x|\mu,\sigma^2)$, is uniquely described by two parameters, the μ and the σ.
- We call distributions which are uniquely defined by a small set of parameters, as parametric distributions.

Frequenist (MLE) vs Bayesian Approach

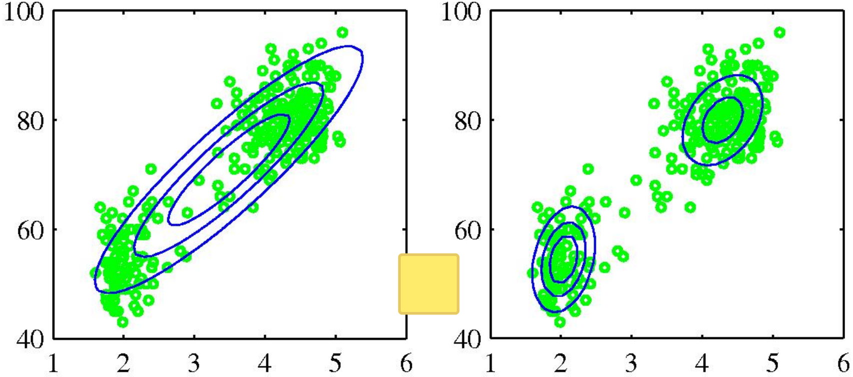
Two different ways to estimate the value of the parameter of the density function $p(x)$, e.g., μ

- Maximum Likelihood Estimation: Maximizing the (log-)likelihood function of a set of data points $\mathcal{D} = \{x_1, \ldots, x_N\}$ as a function of parameter(s) (μ); it is assumed that $\mathcal{D}$ correspond to the IID data.
- Bayesian Approach: Defining a prior distribution $p(\mu)$ for parameter (μ), and obtaining posterior distribution by using: $\text{posterior} \propto \text{likelihood} \times \text{prior}$

Note that the likelihood function plays a key role in both the approaches!

Mixture of Gaussian

Multi-modal data can be better described with mixtures of distributions rather than a single distribution.



Mixture of Gaussians

When using a superposition of K Gaussians, the resulting distribution has the form:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

and is called a mixture of Gaussians.

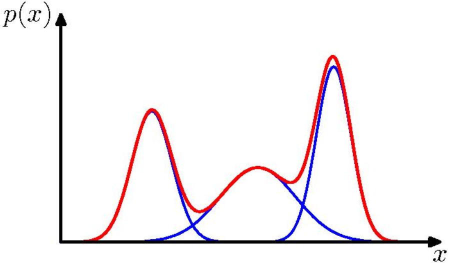
The parameters π_k are called mixing coefficients and satisfy:

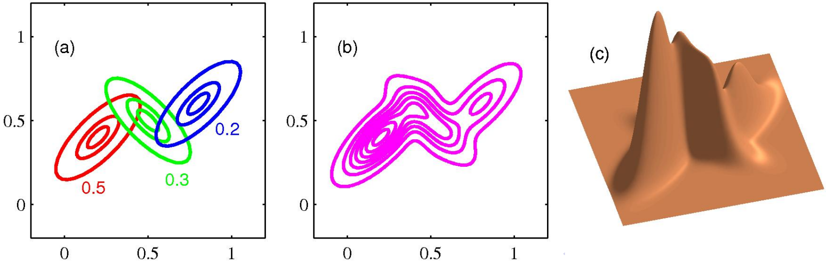
$$0 \leq \pi_k \leq 1, \qquad \sum_{k=1}^{K} \pi_k = 1.$$

Thus, the mixing coefficients have the form of probabilities ($\pi_k = p(k)$).

Mixture of Gaussians

The resulting distribution can take complex forms:





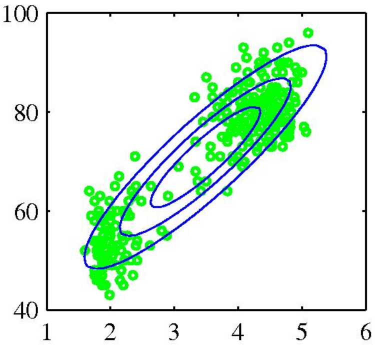
Parameter Estimation in Mixture of Gaussian

The form of the Gaussian mixture is governed by the parameters: $\boldsymbol{\pi}$, $\boldsymbol{\mu}_k$ and Σ.

1. How can maximum likelihood estimation procedure be applied to obtain the above parameters, given N IID data points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{\boldsymbol{n}}\}$?
2. What problems do you see in this approach?

Nonparametric Density Estimation

- Nonparametric density estimation approaches make very few assumptions about the form of the distribution
- Parametric methods: efficient (dependence on few parameters) but not flexible (e.g., poor performance on multimodal data)
- Nonparametric methods: no assumptions on data hence are flexible; may be computationally expensive.



Histogram Method

For 1-dimensional data, we can use histogram density models. Standard histograms are obtained by:

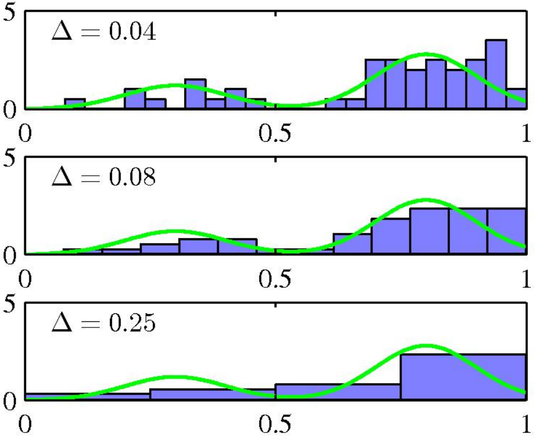
- partitioning $\mathbb{R}$ into distinct bins of width Δ_i
- counting the number of data points n_i falling in bin i
- normalizing using N and Δ_i to obtain a probability density: $\mathrm{p}_i = \frac{n_i}{N\Delta_i}$.

This gives a model for the density $p(x)$ that is constant over the width of each bin (piece wise constant model).

Histogram Method

Bin-width as smoothing parameter

Histogram-based density estimation of the distribution following the green line using $N = 50$ data points for a varying number of (equal-length) bins.



Histogram Method: Advantages and disadvantages

- Advantages
  1. Provides quick visualization of data (as approximate density estimate)
  2. Suitable for large data: data can be discarded once histogram is computed
  3. amenable for sequential processing
- Disadvantages
  1. Discontinuity at bin edges
  2. Curse of dimensionality

Foundations for nonparametric density estimation methods

Let us assume that observations are drawn from an unknown probability density $p(\boldsymbol{x})$ in some D-dimensional space $\boldsymbol{x} \in \mathbb{R}^D$. We wish to estimate $p(\boldsymbol{x})$ from data.

Consider a small region $\mathcal{R}$ containing $\boldsymbol{x}$. The probability mass within $\mathcal{R}$ is $P = \int_{\mathcal{R}} p(\boldsymbol{x}) d\boldsymbol{x}$. Suppose we collect N observations drawn from $p(\boldsymbol{x})$. Since each data point has probability P of being within $\mathcal{R}$, the total number K of points that lie within $\mathcal{R}$ follow binomial distribution.

- For $N \to \infty$ data points drawn from $p(\boldsymbol{x})$, the approximate number of data points falling inside $\mathcal{R}$ is NP.
- For small $\mathcal{R}$, we can assume that $p(\boldsymbol{x})$ is constant over $\mathcal{R}$ leading to $P \sim p(\boldsymbol{x})V$ for $\boldsymbol{x} \in \mathcal{R}$.
- Combining the above two cases we have:

$$p(\boldsymbol{x}) = \frac{K}{NV}. \tag{3}$$

Nonparametric density estimation

Two approaches for nonparametric density estimation based on (3)

- Kernel density estimation: Fix V and determine K from the data
- K nearest-neighbour: Fix K and determine V from the data

Kernel Density Estimation

Let $\mathcal{R}$ be a small hypercube centered at the point x. To count the number K of points inside $\mathcal{R}$ we define the function:

$$k(\boldsymbol{u}) = \begin{cases} 1, & |u_i| \leq 1/2, \quad i = 1, \dots, D \\ 0, & \text{otherwise} \end{cases}$$

which expresses a hypercube centered at the origin. $k(\boldsymbol{u})$ is called kernel function (or *Parzen window*).

The quantity $k\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right)$ will be 1 if the data point $\boldsymbol{x}_n$ lies inside cube of side h centered at $\boldsymbol{x}$.

The number of points inside the hypercube of side h centered at $\boldsymbol{x}$:

$$K = \sum_{n=1}^{N} k\left(\frac{\boldsymbol{x}-\boldsymbol{x}_n}{h}\right)$$

Kernel Density Estimation

Remember that $p(\mathrm{x}) = \frac{K}{NV}$. Thus by substituting $V = h^D$ and K from the above:

$$p(\mathrm{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} k\left(\frac{\mathrm{x} - \mathrm{x}_n}{h}\right).$$

Reinterpret the above not as single hypercube centered at $\boldsymbol{x}$ but N hypercubes centered at $\boldsymbol{x}_n$.

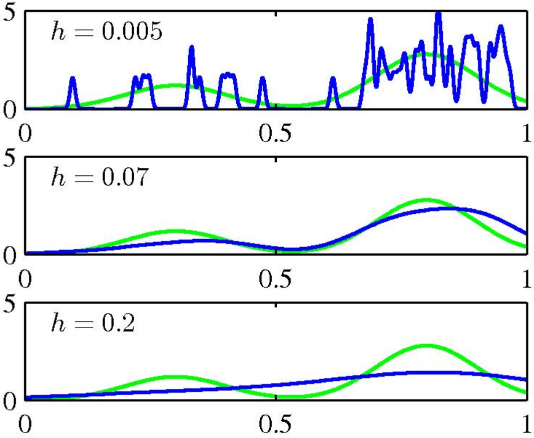
To avoid artificial discontinuities (at the borders of the hypercube), we can use a Gaussian kernel instead:

$$p(\mathrm{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{1/2}} exp\left\{-\frac{\|\mathrm{x} - \mathrm{x}_n\|^2}{2h^2}\right\}$$

where h is the standard deviation of the Gaussian components.

Kernel Density Estimation

Illustration of Kernel density model applied on real data; note the role of *h* as smoothing parameter



Nearest-neighbour methods

In kernel density estimators, the use of a fixed parameter h causes problems

- for regions with high data density, we need h to be small; large h will lead to over-smoothing
- but for regions with low density of points, small h will lead to noisy/spikier estimates
- h should depend on data (adaptive)

Nearest-neighbour method fixes the number of neighbours K (instead of fixing V). Then:

$$p(\mathrm{x}) = \frac{K}{NV} \tag{4}$$

where V corresponds to the volume of the space around $\boldsymbol{x}$ containing K neighbours.

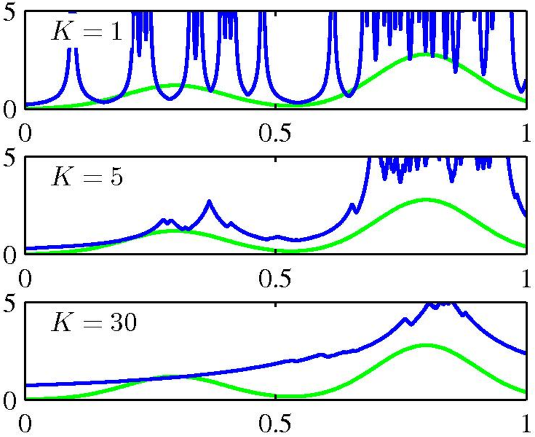
Nearest-neighbour methods

Operation:

- Consider a small sphere around $\boldsymbol{x}$ where we need to calculate $p(\boldsymbol{x})$
- Allow the radius of the sphere to grow until it contains precisely K data points
- The estimate of $p(\boldsymbol{x})$ is then given by (4) with V set to the volume of the sphere

Nearest-neighbour methods

Example: Illustration of Nearest-neighbour density model applied on real data.



K-NN method for classification

Let us assume that we have a dataset formed by N_k points in class $\mathcal{C}_k$, and $N = \sum_k N_k$ points in total. To classify a new point $\boldsymbol{x}$:

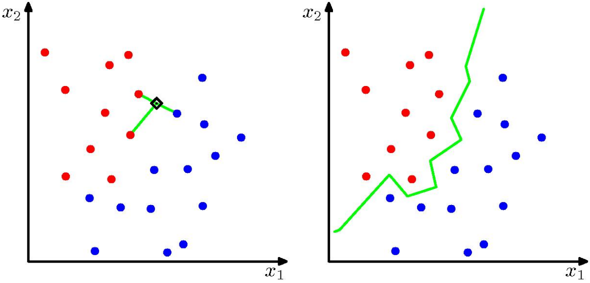
- Draw a sphere centered at $\boldsymbol{x}$ containing exactly K (labeled) points irrespective of their class
- The sphere has a volume V and contains K_k points from class $\mathcal{C}_k$
- Then we define the following probabilities:
  - $p(\boldsymbol{x}|\mathcal{C}_k) = K_k/(N_k V)$
  - $p(\boldsymbol{x}) = K/(NV)$
  - $p(\boldsymbol{C}_k) = N_k/N$
- Using the Bayes formula:

$$p(\mathcal{C}_k|\mathrm{x}) = \frac{p(\mathrm{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathrm{x})} = \frac{K_k}{K}$$

To minimize the misclassification probability, a test point $\boldsymbol{x}$ is assigned to the class having largest posterior probability.

K-NN method for classification

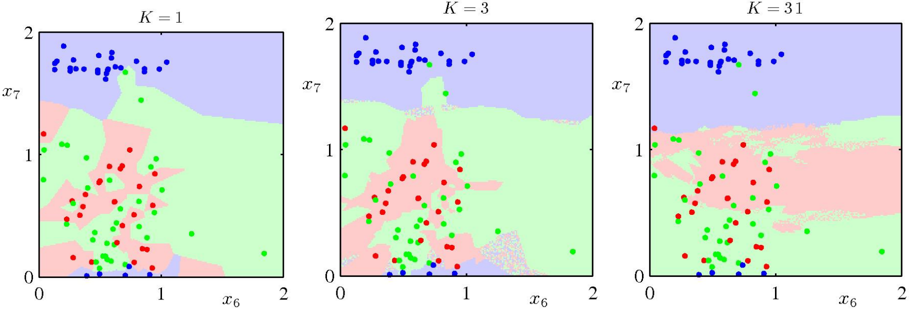
A test point is assigned according to the majority class membership of the K nearest training data points. (left) K=3 and (right) K=1.



(a) (b)

K-NN method for classification

Illustration of K as controlling the degree of smoothness of the estimated density



$K = 1$
$K = 3$
$K = 31$
x_7
x_6
0
1
2