Density Estimation in Supervised Learning

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Illustrate classification in Supervised Learning

Objective

Discuss basic density estimation techniques

Supervised Learning

- **The set-up:** the given training data consist of <sample, label> pairs, or (**x**, y); the objective of learning is to figure out a way to predict label y for any new sample **x**.

- Consider two types of problems:
  - Regression: y continuous
  - **Classification**: y is discrete, e.g., class labels.

Examples of Image Classification

The MNIST training images of hand-written digits

The Extended Yale B Face Images

How do we model the training images?

Parametric: each class of images (the feature vectors) may be modeled by a density function $p_{\boldsymbol{\theta}}(\mathbf{x})$ with parameter $\boldsymbol{\theta}$.

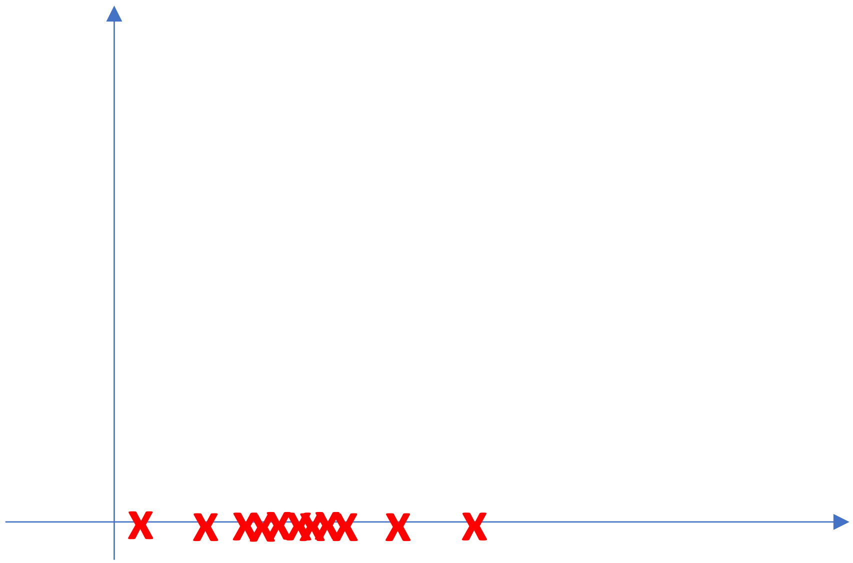
- To emphasize the density is for images from class/label *y*, we may write $p_{\boldsymbol{\theta}}(\mathbf{x}|y)$.
- We may also use the notation $p(\mathbf{x}|\boldsymbol{\theta})$, if the discussion is true for any y.

➔ How to estimate $\boldsymbol{\theta}$ from the training images?

Note: We may also consider **non-parametric** approaches.

MLE for Density Estimation (1/3)

- Given some training data; Assuming a parametric model $p(\mathbf{x}|\boldsymbol{\theta})$; What specific $\boldsymbol{\theta}$ will fit/explain the data best?
  - E.g., Consider a simple 1-D normal density with only a parameter μ (assuming the variance is known)



- Given a sample x_i, $p(x_i \mid \mu)$ gives an indication of how likely x_i is from $p(x_i \mid \mu)$

➔ the concept of the likelihood function.

MLE for Density Estimation (2/3)

- The likelihood function: the density function $p(\mathbf{x}|\boldsymbol{\theta})$ evaluated at the given data sample $\mathbf{x}_i$, and viewed as a function of the parameter $\boldsymbol{\theta}$.
  - Assessing how likely the parameter $\boldsymbol{\theta}$ (defining the corresponding $p(\mathbf{x}|\boldsymbol{\theta})$) gives arise to the sample $\mathbf{x}_i$.
  - We often use $L(\boldsymbol{\theta})$ to denote the likelihood function, and $l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta}))$ is called the log-likelihood.

- **Maximum Likelihood Estimation (MLE)**: Finding the parameter that maximizes the likelihood function

$$\widehat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, p(\mathbf{x}|\boldsymbol{\theta})$$

MLE for Density Estimation (3/3)

- How to consider *all* the given samples $D=\{\mathbf{x}_i, i=1,\ldots,n\}$?

- The concept of i.i.d. samples: the samples are assumed to be *independent* and *identically distributed*

- So, the data likelihood is given by

$$L(\boldsymbol{\theta}) = P(D|\boldsymbol{\theta}) =$$ `product[ p(x_i | theta) ]`

MLE Example 1

- Tossing a coin for n times, observing n_1 times for head.
  - Estimate the probability θ for head

- The likelihood function is:

$$L(\theta) = P(D|\theta) = \theta^{n_1}(1-\theta)^{n-n_1}$$

MLE Example 1 (cont'd)

- We want to find what θ maximizes this likelihood, or equivalently, the log-likelihood

$$l(\theta) = \log P(D|\theta) = \log(\theta^{n_1}(1-\theta)^{n-n_1})$$
$$= \cdots$$

Take the derivative and set to 0:

$$\frac{d}{d\theta} l(\theta) = 0$$

This will give us:

$$\hat{\theta} = \frac{n_1}{n}$$

MLE Example 2

- Given n i.i.d. samples $\{x_i\}$ from the 1-D normal distribution $N(\mu, \sigma^2)$, find the MLE for μ and σ^2

- The likelihood function is:

$$L(\mu, \sigma) = p(D|\mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- The log-likelihood is: $l(\mu, \sigma) = \log P(D|\mu, \sigma)$

$$= \log\left(\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right)$$

$$= -n\log\left(\sigma\sqrt{2\pi}\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

MLE Example 2 (cont'd)

The MLE solution for μ

$$\hat{\mu} = \text{argmax}_{\mu} l(\mu, \sigma)$$
$$= \text{argmax}_{\mu}\{-n \log\left(\sigma\sqrt{2\pi}\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\}$$

Set the derivative to 0: $$\frac{\partial}{\partial\mu} l(\mu, \sigma) = 0$$

The solution is: $$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

MLE Example 2 (cont'd)

The MLE solution for σ^2

$$\hat{\sigma} = \text{argmax}_{\sigma} l(\mu, \sigma)$$
$$= \text{argmax}_{\sigma}\{-n \log\left(\sigma\sqrt{2\pi}\right) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\}$$

Set the derivative to 0: $$\frac{\partial}{\partial\sigma} l(\mu, \sigma) = 0$$

The solution is: $$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$