CS 2750 Machine Learning
Lecture 4

Density estimation

Milos Hauskrecht
milos@pitt.edu
5329 Sennott Square

Density estimation

Density estimation: is an unsupervised learning problem

- **Goal:** Learn a model that represent the relations among attributes in the data

$$D = \{D_1, D_2, .., D_n\}$$

Data: $D_i = \mathbf{x}_i$ a vector of attribute values

Attributes:

- modeled by random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ with
  - **Continuous or discrete valued variables**

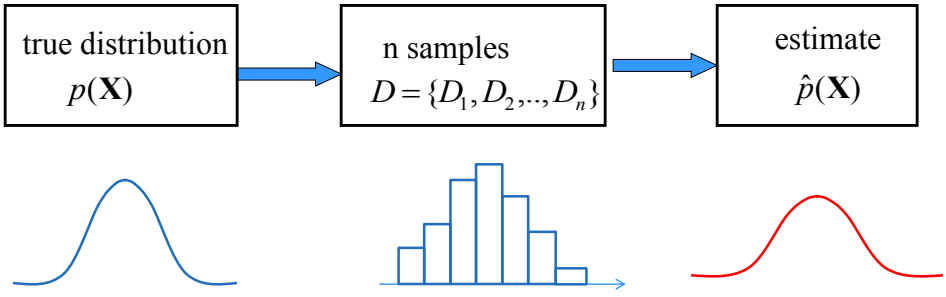
Density estimation: learn an underlying probability distribution model : $p(\mathbf{X}) = p(X_1, X_2, \ldots, X_d)$ **from D**

Density estimation

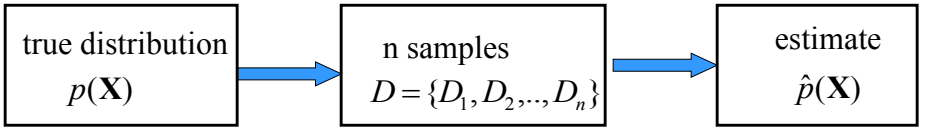
Data: $D = \{D_1, D_2, .., D_n\}$

$D_i = \mathbf{x}_i$ a vector of attribute values

Objective: estimate the model of the underlying probability distribution over variables $\mathbf{X}$, $p(\mathbf{X})$, using examples in D

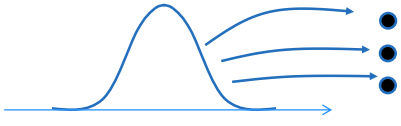


Density estimation



Standard (iid) assumptions: Samples

- **are independent of each other**
- **come from the same (identical) distribution (fixed $p(\mathbf{X})$)**



Independently drawn instances from the same fixed distribution

Density estimation

Types of density estimation:

Parametric

- the distribution is modeled using a set of parameters Θ

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \Theta)$$

- **Example:** mean and covariances of a multivariate normal
- **Estimation:** find parameters Θ describing data D

Non-parametric

- The model of the distribution utilizes all examples in D
- As if all examples were parameters of the distribution
- **Examples:** Nearest-neighbor

Learning via parameter estimation

In this lecture we consider **parametric density estimation**

Basic settings:

- A set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$
- **A model of the distribution** over variables in $\boldsymbol{X}$ with parameters Θ : $\hat{p}(\mathbf{X} \mid \Theta)$

Example: Gaussian distribution with mean and variance parameters

- **Data** $D = \{D_1, D_2, .., D_n\}$

Objective: find parameters Θ such that $p(\mathbf{X} \mid \Theta)$ fits data D the best

---

ML Parameter estimation

Model $\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \boldsymbol{\Theta})$ **Data** $D = \{D_1, D_2, .., D_n\}$

- **Maximum likelihood (ML)**
  - Find Θ that maximizes the likelihood $p(D \mid \Theta, \xi)$

$$P(D \mid \Theta, \xi) = P(D_1, D_2, .., D_n \mid \Theta, \xi)$$
$$= P(D_1 \mid \Theta, \xi) P(D_2 \mid \Theta, \xi) \ldots P(D_n \mid \Theta, \xi)$$
$$= \prod_{i=1}^{n} P(D_i \mid \Theta, \xi)$$

Independent examples

log-likelihood $\log p(D \mid \Theta, \xi) = \sum_{i=1}^{n} \log P(D_i \mid \Theta, \xi)$

$$\Theta_{ML} = \arg\max_{\Theta} p(D \mid \Theta, \xi) = \arg\max_{\Theta} \log p(D \mid \Theta, \xi)$$

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \boldsymbol{\Theta}_{ML})$$

Bayesian parameter estimation

Bayesian parameter estimation

- uses the posterior distribution over possible parameters
- Yields: all possible settings of (and their "weights")
- The target distribution is approximated as:

Parameter posterior ← $p(\Theta \mid D, \xi) = \dfrac{p(D \mid \Theta, \xi) p(\Theta \mid \xi)}{p(D \mid \xi)}$ → Data Likelihood, Parameter prior

- How to use the posterior for modeling p(X)?

$$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid D) = \int_{\boldsymbol{\Theta}} p(X \mid \boldsymbol{\Theta}) p(\boldsymbol{\Theta} \mid D, \xi) d\boldsymbol{\Theta}$$

Parameter estimation

Other criteria:

- **Maximum a posteriori probability (MAP)**

  maximize $p(\boldsymbol{\Theta} \mid D, \xi)$ (mode of the posterior)

  – Yields: one set of parameters $\boldsymbol{\Theta}_{MAP}$

  – Approximation:

  $$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \boldsymbol{\Theta}_{MAP})$$

- **Expected value of the parameter**

  $\hat{\boldsymbol{\Theta}} = E(\boldsymbol{\Theta})$ (mean of the posterior)

  – Expectation taken with regard to posterior $p(\boldsymbol{\Theta} \mid D, \xi)$

  – Yields: one set of parameters

  – Approximation:

  $$\hat{p}(\mathbf{X}) = p(\mathbf{X} \mid \hat{\boldsymbol{\Theta}})$$