

CS 2710 Foundations of AI
Lecture 24

Learning probability distributions

Milos Hauskrecht

milos@cs.pitt.edu

5329 Sennott Square

CS 2710 Foundations of AI

Unsupervised learning

- **Data:** $D = \{D_1, D_2, \dots, D_n\}$
 $D_i = \mathbf{x}_i$ a vector of attribute values
 - e.g. the description of a patient
 - no specific target attribute we want to predict (no output y)
- **Objective:**
 - learn (describe) relations between attributes, examples

Types of problems:

- **Clustering**
 - Group together “similar” examples
- **Density estimation**
 - Model probabilistically the population of examples

CS 2710 Foundations of AI

Density estimation

Data: $D = \{D_1, D_2, \dots, D_n\}$
 $D_i = \mathbf{x}_i$ a vector of attribute values

Attributes:

- modeled by random variables $\mathbf{X} = \{X_1, X_2, \dots, X_d\}$ with:
 - Continuous values
 - Discrete values

E.g. *blood pressure* with numerical values

or *chest pain* with discrete values

[no-pain, mild, moderate, strong]

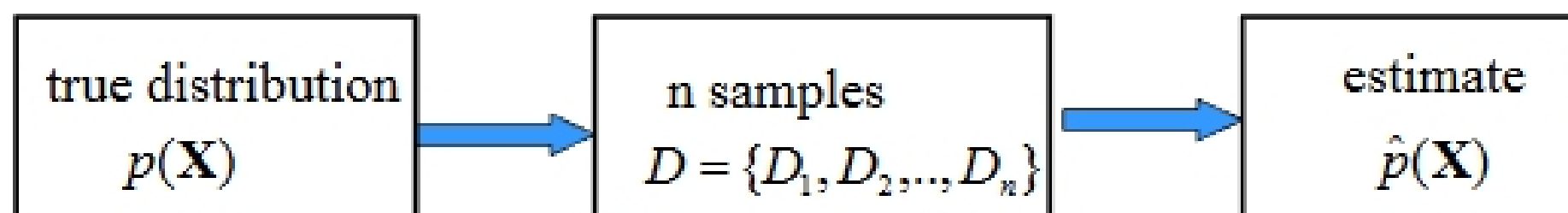
Underlying true probability distribution:

$$p(\mathbf{X})$$

Density estimation

Data: $D = \{D_1, D_2, \dots, D_n\}$
 $D_i = \mathbf{x}_i$ a vector of attribute values

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



Standard (iid) assumptions: Samples

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

Learning via parameter estimation

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

Basic settings:

- A set of random variables $\mathbf{X} = \{X_1, X_2, \dots, X_d\}$
- **A model of the distribution** over variables in \mathbf{X} with parameters Θ
- **Data** $D = \{D_1, D_2, \dots, D_n\}$

Objective: find parameters $\hat{\Theta}$ that fit the data the best, or in other words reduce the misfit between the data and the model

- What is the best set of parameters?
 - There are various criteria one can apply here.

Parameter estimation. Basic criteria.

- **Maximum likelihood (ML) criterion**

$$\arg \max_{\Theta} p(D | \Theta, \xi) \leftarrow \text{Likelihood of data}$$

ξ - represents prior (background) knowledge

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

$$\arg \max_{\Theta} p(\Theta | D, \xi) \leftarrow \text{Posterior probability}$$

MAP selects the mode of the posterior

$$p(\Theta | D, \xi) = \frac{p(D | \Theta, \xi) p(\Theta | \xi)}{p(D | \xi)}$$