Lecture-3
Density estimation

CS 277:
Machine Learning
and
Data Science

By:
Dr. Joydeep Chandra
Associate Professor
Dept. of CSE, IIT Patna

Outline

Outline:

- **Density estimation:**
  - **Maximum likelihood (ML)**
  - **Bayesian parameter estimates**
  - **MAP**
- **Bernoulli distribution**
- **Binomial distribution**
- **Multinomial distribution**
- **Normal distribution**

Density estimation

Density estimation: is an unsupervised learning problem

- **Goal:** Learn relations among attributes in the data

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

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

Attributes:

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

Density estimation: learn the underlying probability distribution:

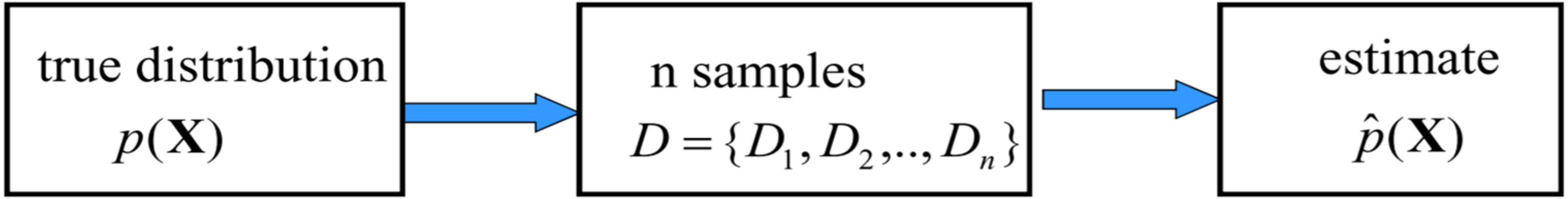
$p(\mathbf{X}) = p(X_1, X_2, ..., X_d)$ **from D**

Density estimation

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

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

Objective: estimate the underlying 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** $p(\mathbf{X})$ **)**

Density estimation

Types of density estimation:

Parametric

- the distribution is modeled using a set of parameters $\boldsymbol{\Theta}$

$$p(\mathbf{X}|\boldsymbol{\Theta})$$

- **Example:** mean and covariances of a multivariate normal
- **Estimation:** find parameters $\boldsymbol{\Theta}$ 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 **X** with parameters $\Theta \;:\; \hat{p}(\mathbf{X} \mid \Theta)$
- **Data** $D = \{D_1, D_2, .., D_n\}$

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

Parameter estimation

- **Maximum likelihood (ML)**

  maximize $p(D \mid \Theta, \xi)$

  – yields: one set of parameters Θ_{ML}

  – the target distribution is approximated as:

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

- **Bayesian parameter estimation**

  – uses the posterior distribution over possible parameters

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

  – Yields: all possible settings of Θ (and their "weights")

  – The target distribution is approximated as:

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

Parameter estimation

Other possible criteria:

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

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

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

  – Approximation:

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

- **Expected value of the parameter**

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

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

  – Yields: one set of parameters

  – Approximation:

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

Parameter estimation: Coin example.

Coin example: we have a coin that can be biased

Outcomes: two possible values -- head or tail

Data: *D* a sequence of outcomes x_i such that

- **head $x_i = 1$**
- **tail $x_i = 0$**

Model: probability of a head Θ

probability of a tail $(1-\Theta)$

Objective:

We would like to estimate the probability of a **head** $\hat{\theta}$ from data

Parameter estimation: Example

Assume the unknown and possibly biased coin

- Probability of the head is $\boldsymbol{\Theta}$
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T

  - **Heads:** 15
  - **Tails:** 10

What would be your estimate of the probability of a head ?

$$\tilde{\theta} = ?$$

Parameter estimation: Example

Assume the unknown and possibly biased coin

- Probability of the head is Θ
- **Data:**

  H H T T H H T H T H T T T H T H H H H T H H H H T

  - **Heads:** 15
  - **Tails:** 10

What would be your estimate of the probability of a head ?

Solution: use frequencies of occurrences to do the estimate

$$\tilde{\theta} = \frac{15}{25} = 0.6$$

This is **the maximum likelihood estimate** of the parameter Θ

Probability of an outcome

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ
probability of a tail $(1-\theta)$

Assume: we know the probability θ
Probability of an outcome of a coin flip x_i

$$P(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{(1-x_i)}$$ ⟵ **Bernoulli distribution**

- Combines the probability of a head and a tail
- So that x_i is going to pick its correct probability
- Gives θ for $x_i = 1$
- Gives $(1-\theta)$ for $x_i = 0$

Probability of a sequence of outcomes

Data: *D* a sequence of outcomes x_i such that

- **head $x_i = 1$**
- **tail $x_i = 0$**

Model: probability of a head Θ

probability of a tail $(1-\Theta)$

Assume: a sequence of independent coin flips

D = H H T H T H (encoded as D= 110101)

What is the probability of observing the data sequence **D:**

$$P(D \mid \theta) = ?$$

Probability of a sequence of outcomes

Data: *D* a sequence of outcomes x_i such that

- **head $x_i = 1$**
- **tail $x_i = 0$**

Model: probability of a head Θ

probability of a tail $(1-\Theta)$

Assume: a sequence of independent coin flips

D = H H T H T H encoded as D= 110101

What is the probability of observing the data sequence **D:**

$$P(D \mid \theta) = \theta\theta(1-\theta)\theta(1-\theta)\theta$$

Probability of a sequence of outcomes

Data: *D* a sequence of outcomes $\boldsymbol{x_i}$ such that

- **head $x_i = 1$**
- **tail $x_i = 0$**

Model: probability of a head $\boldsymbol{\Theta}$

probability of a tail *(1-$\boldsymbol{\Theta}$)*

Assume: a sequence of independent coin flips

D = H H T H T H encoded as D= 110101)

What is the probability of observing the data sequence **D:**

$$P(D \mid \theta) = \theta\theta(1-\theta)\theta(1-\theta)\theta$$

likelihood of the data

Probability of a sequence of outcomes

Data: *D* a sequence of outcomes x_i such that

- **head $x_i = 1$**
- **tail $x_i = 0$**

Model: probability of a head Θ

probability of a tail $(1-\Theta)$

Assume: a sequence of independent coin flips

D = H H T H T H encoded as D= 110101)

What is the probability of observing the data sequence **D:**

$$P(D \mid \theta) = \theta\theta(1-\theta)\theta(1-\theta)\theta$$

$$P(D \mid \theta) = \prod_{i=1}^{6} \theta^{x_i}(1-\theta)^{(1-x_i)}$$

Can be rewritten using the Bernoulli distribution:

The goodness of fit to the data

Learning: **we do not know the value of the parameter** θ

Our learning goal:

- Find the parameter θ that fits the data D the best?

One solution to the "best": Maximize the likelihood

$$P(D \mid \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{(1-x_i)}$$

Intuition:

- more likely are the data given the model, the better is the fit

Note: Instead of an error function that measures how bad the data fit the model we have a measure that tells us how well the data fit :

$$Error\,(D,\theta) = -P(D \mid \theta)$$

Example: Bernoulli distribution

Coin example: we have a coin that can be biased

Outcomes: two possible values -- head or tail

Data: D a sequence of outcomes x_i such that

- **head** $x_i = 1$
- **tail** $x_i = 0$

Model: probability of a head θ

probability of a tail $(1-\theta)$

Objective:

We would like to estimate the probability of a **head** $\hat{\theta}$

Probability of an outcome x_i

$$P(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{(1-x_i)}$$ **Bernoulli distribution**

Any Questions??