Density estimation

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Announcements

Homework 1:
• due on Thursday, January 23 before the class

You should submit:
• A hardcopy of the report (before the lecture)
• Programs (if we ask for them) in electronic form
  – Instructions for program submissions are on the course web site
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**: \( D = \{ D_1, D_2, \ldots, D_n \} \)
\( D_i = x_i \) a vector of attribute values

**Attributes**:
• modeled by random variables \( X = \{ X_1, X_2, \ldots, X_d \} \) with
  – Continuous or discrete valued variables

**Density estimation**: learn the underlying probability distribution: \( p(X) = p(X_1, X_2, \ldots, X_d) \) from \( D \)
Density estimation

Data: \( D = \{D_1, D_2, \ldots, D_n\} \)
\[ D_i = x_i \quad \text{a vector of attribute values} \]

Objective: estimate the underlying probability distribution over variables \( X \), \( p(X) \), using examples in \( D \)

Standard (iid) assumptions: Samples
- are independent of each other
- come from the same (identical) distribution (fixed \( p(X) \))

Types of density estimation:

Parametric
- the distribution is modeled using a set of parameters \( \Theta \)
  \[ p(X | \Theta) \]
- Example: mean and covariances of a multivariate normal
- Estimation: find parameters \( \Theta \) 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 $X = \{X_1, X_2, \ldots, X_d\}$
- A model of the distribution over variables in $X$ with parameters $\Theta : \hat{p}(X | \Theta)$
- Data $D = \{D_1, D_2, \ldots, D_n\}$

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

---

## Parameter estimation

- **Maximum likelihood (ML)**
  - maximize $p(D | \Theta, \xi)$
  - yields: one set of parameters $\Theta_{ML}$
  - the target distribution is approximated as:
    \[ \hat{p}(X) = p(X | \Theta_{ML}) \]
- **Bayesian parameter estimation**
  - uses the posterior distribution over possible parameters
    \[ p(\Theta | D, \xi) = \frac{p(D | \Theta, \xi) p(\Theta | \xi)}{p(D | \xi)} \]
  - Yields: all possible settings of $\Theta$ (and their “weights”)
  - The target distribution is approximated as:
    \[ \hat{p}(X) = p(X | D) = \int p(X | \Theta) p(\Theta | D, \xi) d\Theta \]
Parameter estimation

Other possible criteria:

• **Maximum a posteriori probability (MAP)**
  
  maximize \( p(\Theta | D, \xi) \) (mode of the posterior)
  
  – Yields: one set of parameters \( \Theta_{MAP} \)
  
  – Approximation:
  
  \[ \hat{p}(X) = p(X | \Theta_{MAP}) \]

• **Expected value of the parameter**
  
  \( \hat{\Theta} = E(\Theta) \) (mean of the posterior)
  
  – Expectation taken with regard to posterior \( p(\Theta | D, \xi) \)
  
  – Yields: one set of parameters
  
  – Approximation:
  
  \[ \hat{p}(X) = p(X | \hat{\Theta}) \]

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 \( \theta \) 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 $\theta$
• Data:
  H H T T H H T H T T H T H T H H 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?

$\tilde{\theta} = ?$

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

Assume: we know the probability $\theta$

Probability of an outcome of a coin flip $x_i$

$P(x_i | \theta) = \theta^{x_i} (1-\theta)^{(1-x_i)} \quad \text{Bernoulli distribution}$

- Combines the probability of a head and a tail
- So that $x_i$ is going to pick its correct probability
- Gives $\theta$ 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 $\theta$
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 $D$:

$P(D | \theta) =$ ?
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 $\theta$
- probability of a tail $(1 - \theta)$

**Assume:** a sequence of coin flips $D = H \ H \ T \ H \ T \ H$

encoded as $D = 110101$

What is the probability of observing a 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 \( \theta \)
probability of a tail \( 1 - \theta \)

Assume: a sequence of coin flips \( D = H H T H T H \)
encoded as \( D = 110101 \)

What is the probability of observing a data sequence \( D \):

\[
P(D \mid \theta) = \theta \theta (1 - \theta)(1 - \theta)\theta
\]

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

Can be rewritten using the Bernoulli distribution:

The goodness of fit to the data

Learning: we do not know the value of the parameter \( \theta \)

Our learning goal:
- Find the parameter \( \theta \) 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 $\theta$

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