Density estimation

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

Announcements

Homework 2
• Due on Wednesday before the class
• Reports: hand in before the class
• Programs: submit electronically

Collaborations on homeworks:
• You may discuss material with your fellow students, but the report and programs should be written individually
Outline

Outline:
• Density estimation:
  – Maximum likelihood (ML)
  – Maximum a posteriori (MAP)
  – Bayesian
• Bernoulli distribution.
• Binomial distribution
• Multinomial distribution.
• Normal distribution.

Density estimation

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 values
  – Discrete values
E.g. \textit{blood pressure} with numerical values
  or \textit{chest pain} with discrete values
  [no-pain, mild, moderate, strong]

Underlying true probability distribution:
\( p(X) \)
Density estimation

Data: \( D = \{D_1, D_2, ..., D_n\} \)
\( D_i = x_i \) a vector of attribute values

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

Types of density estimation:

Parametric
- the distribution is modeled using a set of parameters \( \Theta \)
  \( p(X \mid \Theta) \)
- Example: mean and covariances of 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

Semi-parametric

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

CS 2750 Machine Learning
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 \( \hat{\Theta} \) that describe \( p(X | \Theta) \) 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 \( p(\Theta \mid D, \xi) \) \hspace{1em} \text{(mode of the posterior)}
  
  \begin{itemize}
  \item Yields: one set of parameters \( \Theta_{MAP} \)
  \item Approximation:
    \[ \hat{p}(X) = p(X \mid \Theta_{MAP}) \]
  \end{itemize}

• **Expected value of the parameter**
  \[ \hat{\Theta} = E(\Theta) \] \hspace{1em} \text{(mean of the posterior)}
  
  \begin{itemize}
  \item Expectation taken with regard to posterior \( p(\Theta \mid D, \xi) \)
  \item Yields: one set of parameters
  \item Approximation:
    \[ \hat{p}(X) = p(X \mid \hat{\Theta}) \]
  \end{itemize}

Parameter estimation. Coin example.

**Coin example:** we have a coin that can be biased

**Outcomes:** two possible values -- head or tail

**Data:** \( D \) \hspace{1em} 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:**
  - Heads: 15
  - Tails: 10

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

$\hat{\theta} = ?$

Solution: use frequencies of occurrences to do the estimate

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

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 $x_i$

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

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

\[ P(D \mid \theta) = ? \]
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 $\theta$
- 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 $D = \text{H H T H T H}$
encoded as $D = 110101$

What is the probability of observing a data sequence $D$:

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

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

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 $x_i$**

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

---

**Maximum likelihood (ML) estimate.**

**Likelihood of data:**

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

**Maximum likelihood estimate**

$$\theta_{ML} = \arg \max_{\theta} P(D \mid \theta, \xi)$$

**Optimize log-likelihood (the same as maximizing likelihood)**

$$l(D, \theta) = \log P(D \mid \theta, \xi) = \log \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1-x_i} =$$
$$\sum_{i=1}^{n} x_i \log \theta + (1-x_i) \log (1 - \theta) = \log \theta \sum_{i=1}^{n} x_i + \log (1 - \theta) \sum_{i=1}^{n} (1-x_i)$$

$N_1$ - number of heads seen  
$N_2$ - number of tails seen
Maximum likelihood (ML) estimate.

Optimize log-likelihood

\[ l(D, \theta) = N_1 \log \theta + N_2 \log(1 - \theta) \]

Set derivative to zero

\[ \frac{\partial l(D, \theta)}{\partial \theta} = \frac{N_1}{\theta} - \frac{N_2}{(1 - \theta)} = 0 \]

Solving

\[ \theta = \frac{N_1}{N_1 + N_2} \]

ML Solution: \[ \theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} \]

Maximum likelihood estimate. 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 T H T H H H T H H H T

  – Heads: 15
  – Tails: 10

  What is the ML estimate of the probability of a head and a tail?
Maximum likelihood estimate. Example

• Assume the unknown and possibly biased coin
• Probability of the head is $\theta$
• Data:
  
  H H T T H H T H T H T T H T H H T H H H H T H H T
  
  – Heads: 15
  – Tails: 10

What is the ML estimate of the probability of head and tail?

Head: $\theta_{ML} = \frac{N_1}{N} = \frac{15}{25} = 0.6$

Tail: $(1 - \theta_{ML}) = \frac{N_2}{N} = \frac{10}{25} = 0.4$

Maximum a posteriori estimate

Maximum a posteriori estimate
– Selects the mode of the posterior distribution

$\theta_{MAP} = \arg \max_{\theta} p(\theta \mid D, \xi)$

Likelihood of data

$p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) p(\theta \mid \xi)}{P(D \mid \xi)}$ (via Bayes rule)

Normalizing factor

$P(D \mid \theta, \xi) = \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{(1-x_i)} = \theta^{N_1} (1 - \theta)^{N_2}$

$p(\theta \mid \xi)$ - is the prior probability on $\theta$

How to choose the prior probability?
Prior distribution

Choice of prior: Beta distribution

\[ p(\theta \mid \xi) = \text{Beta}(\theta \mid \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1} \]

\[ \Gamma(x) \quad - \quad \text{A Gamma function} \]

For integer values of \( x \) \quad \Gamma(x) = x! \n
Why to use Beta distribution?

Beta distribution “fits” Bernoulli trials - conjugate choices

\[ P(D \mid \theta, \xi) = \theta^{N_1}(1-\theta)^{N_2} \]

Posterior distribution is again a Beta distribution

\[ p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi)\text{Beta}(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = \text{Beta}(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2) \]

Beta distribution

\[ \Gamma(x) \quad - \quad \text{A Gamma function} \]

For integer values of \( x \) \quad \Gamma(x) = x! \n
Why to use Beta distribution?

Beta distribution “fits” Bernoulli trials - conjugate choices

\[ P(D \mid \theta, \xi) = \theta^{N_1}(1-\theta)^{N_2} \]

Posterior distribution is again a Beta distribution

\[ p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi)\text{Beta}(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = \text{Beta}(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2) \]
Maximum a posterior probability

**Maximum a posteriori estimate**
- Selects the mode of the posterior distribution

\[
p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) \text{Beta}(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = \text{Beta}(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)
\]

\[
= \frac{\Gamma(\alpha_1 + \alpha_2 + N_1 + N_2)}{\Gamma(\alpha_1 + N_1)\Gamma(\alpha_2 + N_2)} \theta^{N_1 + \alpha_1 - 1}(1 - \theta)^{N_2 + \alpha_2 - 1}
\]

**Notice** that parameters of the prior act like counts of heads and tails (sometimes they are also referred to as prior counts)

**MAP Solution:**

\[
\theta_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}
\]

MAP estimate example

- Assume the unknown and possibly biased coin
- Probability of the head is \( \theta \)
- **Data:**
  H H T T H H T H T H T T H T H H H T H H T T
  - **Heads:** 15
  - **Tails:** 10
- Assume \( p(\theta \mid \xi) = \text{Beta}(\theta \mid 5, 5) \)

What is the MAP estimate?
MAP estimate example

- Assume the unknown and possibly biased coin
- Probability of the head is $\theta$
- Data:
  
  H H T T H T H T T T H T H H H T H H H T T
  - Heads: 15
  - Tails: 10

- Assume $p(\theta | \xi) = \text{Beta}(\theta | 5,5)$

What is the MAP estimate?

$$\theta_{\text{MAP}} = \frac{N_1 + \alpha_1 - 1}{N - 2} = \frac{N_1 + \alpha_1 - 1}{N_1 + N_2 + \alpha_1 + \alpha_2 - 2} = \frac{19}{33}$$

---

MAP estimate example

- Note that the prior and data fit (data likelihood) are combined
- The MAP can be biased with large prior counts
- It is hard to overturn it with a smaller sample size
- Data:
  
  H H T T H H T T H T T T H T H H H T H H H T T
  - Heads: 15
  - Tails: 10

- Assume $p(\theta | \xi) = \text{Beta}(\theta | 5,5)$
  $\theta_{\text{MAP}} = \frac{19}{33}$

- Assume $p(\theta | \xi) = \text{Beta}(\theta | 5,20)$
  $\theta_{\text{MAP}} = \frac{19}{48}$
Bayesian framework

- Both ML or MAP estimates pick one value of the parameter
  - Assume: there are two different parameter settings that are close in terms of their probability values. Using only one of them may introduce a strong bias, if we use them, for example, for predictions.

- Bayesian parameter estimate
  - Remedies the limitation of one choice
  - Uses all possible parameter values
  - Where \( p(\theta \mid D, \xi) \approx Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2) \)

- The posterior can be used to define \( \hat{p}(X) : \)
  \[
  \hat{p}(X) = p(X \mid D) = \int p(X \mid \Theta) p(\Theta \mid D, \xi) d\Theta
  \]

Bayesian framework

- Predictive probability of an outcome \( x = 1 \) in the next trial
  \[
  P(x = 1 \mid D, \xi) = \frac{1}{\int_0^1 P(x = 1 \mid \theta, \xi) p(\theta \mid D, \xi) d\theta}
  \]
  \[
  = \int_0^1 \theta p(\theta \mid D, \xi) d\theta = E(\theta)
  \]

- Equivalent to the expected value of the parameter
  - expectation is taken with regard to the posterior distribution
  \[
  p(\theta \mid D, \xi) = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)
  \]
Expected value of the parameter

How to obtain the expected value?

\[
E(\theta) = \int_0^1 \theta \beta(\theta | \eta_1, \eta_2) d\theta = \int_0^1 \frac{\theta^{\eta_1+\eta_2}}{\Gamma(\eta_1)\Gamma(\eta_2)} (1-\theta)^{\eta_2-1} d\theta
\]

\[
= \frac{\Gamma(\eta_1+\eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \int_0^1 \theta^{\eta_1} (1-\theta)^{\eta_2-1} d\theta
\]

\[
= \frac{\Gamma(\eta_1+\eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \frac{\Gamma(\eta_1+1)\Gamma(\eta_2)}{\Gamma(\eta_1+\eta_2+1)} \int_0^1 \beta(\eta_1+1, \eta_2) d\theta
\]

\[
= \frac{\eta_1}{\eta_1+\eta_2}
\]

**Note:** \( \Gamma(\alpha + 1) = \alpha \Gamma(\alpha) \) for integer values of \( \alpha \)

---

**Expected value of the parameter**

- **Substituting the results for the posterior:**

\[
p(\theta | D, \xi) = \beta(\theta | \alpha_1 + N_1, \alpha_2 + N_2)
\]

- **We get**

\[
E(\theta) = \frac{\alpha_1 + N_1}{\alpha_1 + N_1 + \alpha_2 + N_2}
\]

- **Note that the mean of the posterior is yet another “reasonable” parameter choice:**

\[
\hat{\theta} = E(\theta)
\]