Density Estimation

Chenghua Lin
Computing Science
University of Aberdeen

Outline

- Density estimation:
  - Maximum likelihood (ML)
  - Maximum a posteriori (MAP)
  - Bayesian framework

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

– **Data** $D = \{D_1, D_2, \dots, D_n\}$

Objective:

– find parameters $\widehat{\Theta}$ that best describe $p(\mathbf{X}|\Theta)$

Parameter estimation

- **Maximum likelihood (ML)**

  Maximize $p(\boldsymbol{D}|\Theta, \zeta)$

  – yields: **one set** of parameters Θ_{ML}

  – the target distribution is approximated as:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X}|\Theta_{ML})$$

Coin example

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 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

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

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

$P(x_i|\theta) = \theta^{x_i}(1-\theta)^{(1-x_i)}$ ← **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**:

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

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**:

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

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**:

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

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**:

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

Can be rewritten using the Bernoulli distribution:

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

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

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

Intuition:

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

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

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

Maximum likelihood estimate

- **Likelihood of data**

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

- **Maximum likelihood estimate**

$$\theta_{ML} = \arg\max_{\theta} P(D|\theta,\zeta)$$

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

$$l(D,\theta) = \log P(D|\theta,\zeta) = log \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{(1-x_i)}$$

$$= \sum_{i=1}^{n} x_i log\theta + (1-x_i)\log(1-\theta) = log\theta \underbrace{\sum_{i=1}^{n} x_i}_{N_1} + \log(1-\theta) \underbrace{\sum_{i=1}^{n}(1-x_i)}_{N_2}$$

N_1 - number of heads seen $\quad$ 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 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 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 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 is the ML estimate of the probability of a head and a tail?
  - **Heads**: $\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1+N_2} = \frac{15}{25} = 0.6$
  - **Tails**: $(1 - \theta_{ML}) = \frac{N_2}{N} = \frac{N_2}{N_1+N_2} = \frac{10}{25} = 0.4$

Parameter estimation

Other possible criteria:

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

  – Maximize $p(\mathbf{\Theta}|D,\zeta)$ (mode of the posterior)

  – yields: one set of parameters Θ_{MAP}

  – Approximation:

$$\hat{p}(\mathbf{X}) = p(\mathbf{X}|\Theta_{MAP})$$

Maximum a posteriori estimate

Maximum a posteriori estimate

Selects the mode of the **posterior distribution**

$$\theta_{MAL} = \arg\max_{\theta} P(D|\theta,\zeta)$$

Likelihood of data

prior

$$p(\theta|D,\zeta) = \frac{P(D|\theta,\zeta)p(\theta|\zeta)}{P(D|\zeta)}$$ **(via Bayes rule)**

Normalizing factor

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

$P(\theta|\zeta)$ is the prior probability on θ

How to choose the prior probability?

Maximum a posteriori estimate

How to choose the prior probability?

- Our prior belief in possible values for θ must reflect the fact that probability is zero for any θ outside the range [0, 1]
- Within [0, 1], we are free to specify our beliefs in any way we wish
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in [0,1]
- Convenience of mathematical calculation, e.g., conjugate prior

Prior distribution

Choice of prior: Beta distribution

$$p(\theta|\zeta) = Beta(\theta|\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$$

$\Gamma(x)$ - A Gamma function, i.e., a generalization of factorial to real numbers

For integer values of x $\quad \Gamma(x) = x!$

Why to use Beta distribution?

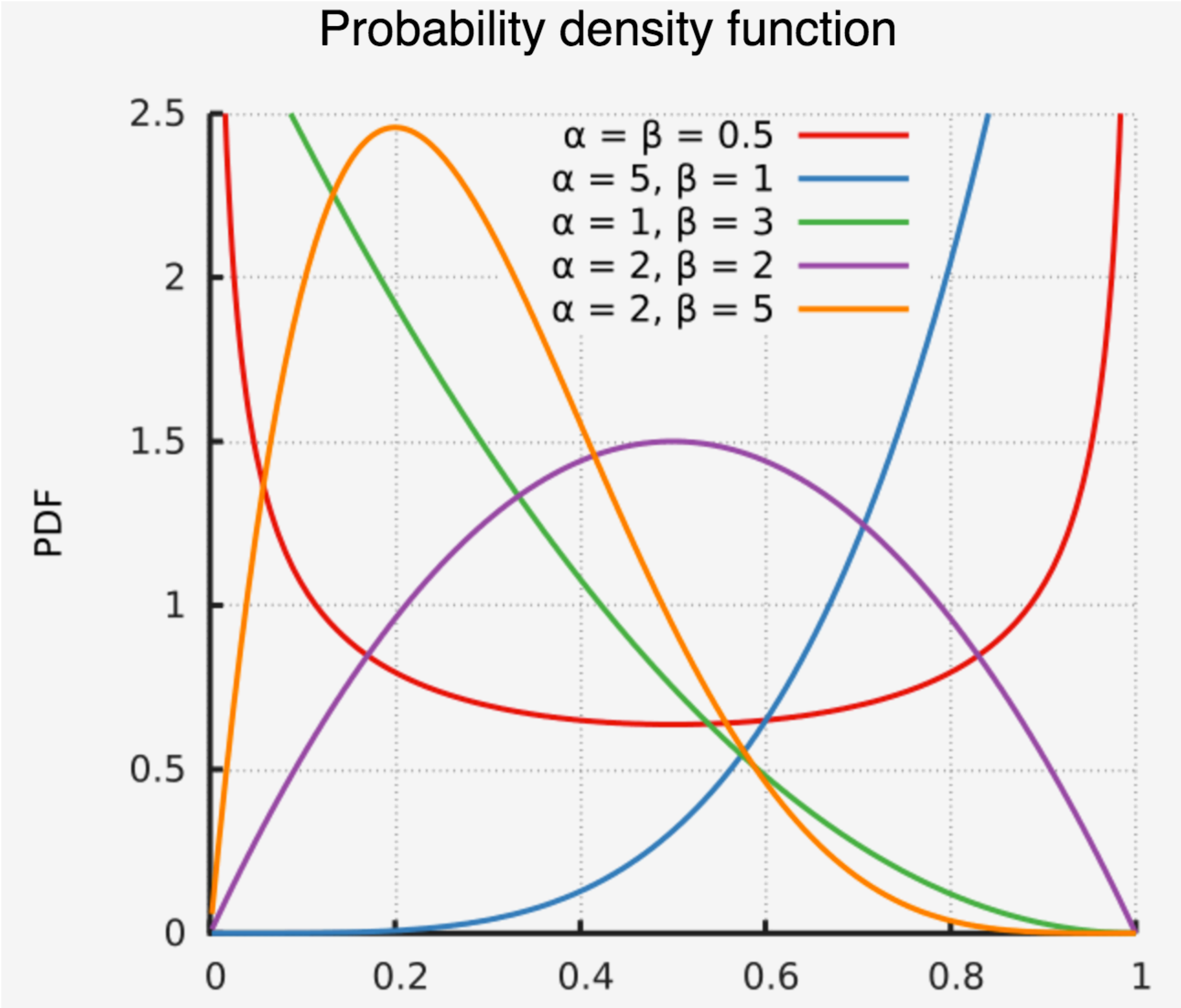
Beta distribution "**fits**" Bernoulli trials - **conjugate choices**

$$P(D|\theta,\zeta) = \theta^{N_1}(1-\theta)^{N_2}$$

Posterior distribution is again a Beta distribution

$$p(\theta|D,\zeta) = \frac{P(D|\theta,\zeta)Bata(\theta|\alpha_1,\alpha_2)}{P(D|\zeta)} = Bata(\theta|\alpha_1 + N_1, \alpha_2 + N_2)$$

Beta distribution



Maximum a posterior probability

Maximum a posteriori estimate

- Selects the mode of the **posterior distribution**

$$p(\theta|D,\zeta) = \frac{P(D|\theta,\zeta)Bata(\theta|\alpha_1,\alpha_2)}{P(D|\zeta)} = Bata(\theta|\alpha_1 + N_1, \alpha_2 + N_2)$$

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

Notice that parameters of the prior

act likes smoothing 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 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
- Assume $p(\theta|\zeta) = Bata(\theta|5,5)$
- What is the MAP estimate?

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

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

  – **Heads**: 15
  – **Tails**: 10
- Assume
  – $p(\theta|\zeta) = Bata(\theta|5{,}5)$ $\theta_{MAP} = 19/33$
  – $p(\theta|\zeta) = Bata(\theta|5{,}20)$ $\theta_{MAP} = 19/48$

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 $p(\theta|\zeta) = Bata(\theta|\alpha_1 + N_1, \alpha_2 + N_2)$
- **The posterior can be used to define $\hat{p}(\mathbf{X})$:**

$$\hat{p}(\mathbf{X}) = \mathrm{p}(\mathbf{X}|\mathrm{D}) = \int_{\mathbf{\Theta}} p(X|\mathbf{\Theta})\, p(\mathbf{\Theta}|D,\zeta) d\mathbf{\Theta}$$

Bayesian framework

- **Predictive probability of an outcome $x = 1$ in the next trial**

$$P(x = 1|D, \zeta)$$

$$P(x = 1|D, \zeta) = \int_0^1 P(x = 1|\theta, \zeta)\, \overbrace{p(\theta|D, \zeta)}^{\text{Posterior density}} d\theta$$

$$= \int_0^1 \theta\, p(\theta|D, \zeta) d\theta = E(\theta)$$

- **Equivalent to the expected value of the parameter**
  - expectation is taken with regard to the posterior distribution

$$p(\theta|D, \zeta) = Bata(\theta|\alpha_1 + N_1, \alpha_2 + N_2)$$

Expected value of the parameter

- **How to obtain the expected value?**

$$E(\theta) = \int_0^1 \theta \; Beta(\theta | a_1, a_2) d\theta = \int_0^1 \theta \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \theta^{a_1 - 1} (1 - \theta)^{a_2 - 1} d\theta$$

$$= \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \int_0^1 \theta^{a_1} (1 - \theta)^{a_2 - 1} d\theta$$

$$= \frac{\Gamma(a_1 + a_2)}{\Gamma(a_1)\Gamma(a_2)} \frac{\Gamma(a_1 + 1)\Gamma(a_2)}{\Gamma(a_1 + a_2 + 1)} \int_0^1 \underbrace{Bata(a_1 + 1, a_2) d\theta}_{1} = \frac{a_1}{a_1 + a_2}$$

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

Expected value of the parameter

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

$$p(\theta|D,\zeta) = Bata(\theta|\alpha_1 + N_1, \alpha_2 + N_2)$$

- **We get**

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

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

$$\hat{\theta} = E(\theta)$$

Comparison

- **Bayesian Learning**
  - Assumes a prior over model parameters.
  - Find posterior of parameters.
- **Maximum a posterior learning**
  - Assumes a prior over model parameters.
  - Chooses the parameters that maximise the posterior $P(\theta|D)$
- **Maximum likelihood learning:**
  - No prior over model parameters
  - Chooses the parameters that maximises the likelihood of the data, $P(D|\theta)$