

CS 3750 Machine Learning

Lecture 2

Density estimation

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

CS 3750 Advanced Machine Learning

Outline

Outline:

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

CS 3750 Advanced Machine Learning

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]

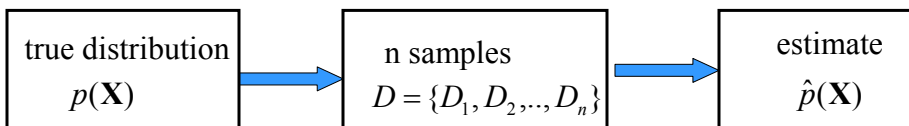
Underlying true probability distribution:

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

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 $p(\mathbf{X})$)

Density estimation

Types of density estimation:

Parametric

- the distribution is modeled using a set of parameters Θ
$$p(\mathbf{X}|\Theta)$$
- **Example:** mean and covariances of a multivariate normal
- **Estimation:** find parameters Θ 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

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 : $\hat{p}(\mathbf{X}|\Theta)$
- **Data** $D = \{D_1, D_2, \dots, D_n\}$

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

Parameter estimation

- **Maximum likelihood (ML)**

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

- yields: one set of parameters Θ_{ML}
- the target distribution is approximated as:

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

- **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 Θ (and their “weights”)
- The target distribution is approximated as:

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

Parameter estimation

Other possible criteria:

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

maximize $p(\Theta | D, \xi)$ (mode of the posterior)

- Yields: one set of parameters Θ_{MAP}
- Approximation:

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

- **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:

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

Parameter estimation. Coin toss 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 choice 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

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

- 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 coin flips $D = H H T H T H$
encoded as $D = 110101$

What is the probability of observing a data sequence D :

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

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 coin flips $D = H H T H T H$
encoded as $D = 110101$

What is the probability of observing a data sequence D :

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


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 coin flips $D = \text{H H T H T H}$
encoded as $D = 110101$

What is the probability of observing a data sequence D :

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

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

Can be rewritten using the Bernoulli distribution:

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

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

Maximum likelihood (ML) estimate.

Likelihood of data:

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

Maximum likelihood estimate

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

Optimize log-likelihood (the same as maximizing likelihood)

$$\begin{aligned} l(D, \theta) &= \log P(D | \theta, \xi) = \log \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{(1-x_i)} = \\ &= \sum_{i=1}^n x_i \log \theta + (1 - x_i) \log(1 - \theta) = \log \theta \sum_{i=1}^n x_i + \log(1 - \theta) \sum_{i=1}^n (1 - x_i) \end{aligned}$$

N_1 - number of heads seen N_2 - number of tails seen

CS 3750 Advanced Machine Learning

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

$$\theta = \frac{N_1}{N_1 + N_2}$$

$$\text{ML Solution: } \theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}$$

CS 3750 Advanced Machine Learning

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 head and tail ?

Head: $\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} = \frac{15}{25} = 0.6$

Tail: $(1 - \theta_{ML}) = \frac{N_2}{N} = \frac{N_2}{N_1 + N_2} = \frac{10}{25} = 0.4$

Posterior density

Bayesian and MAP approaches rely on the posterior density

$$p(\theta | D, \xi)$$

Can be calculated as:

Likelihood of data \swarrow **prior** \swarrow

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

\swarrow **Normalizing factor**

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

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

How to choose the prior probability?

Prior distribution

Choice of prior: **Beta distribution**

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

$\Gamma(x)$ - A Gamma function

For integer values of x $\Gamma(n) = (n-1)!$

Why to use Beta distribution?

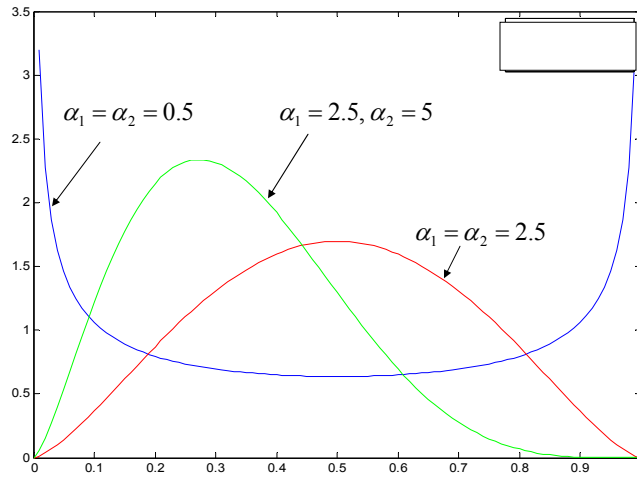
Beta distribution “fits” Bernoulli trials - **conjugate choice**

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

Posterior distribution is again a Beta distribution

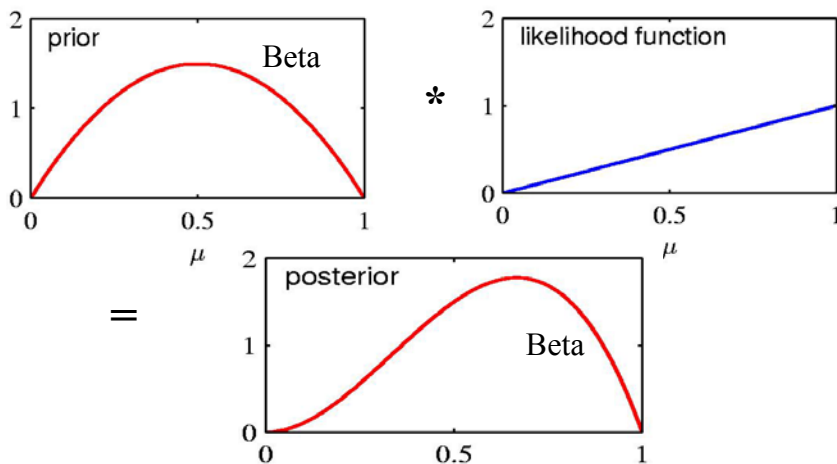
$$p(\theta | D, \xi) = \frac{P(D | \theta, \xi) \text{Beta}(\theta | \alpha_1, \alpha_2)}{P(D | \xi)} = \text{Beta}(\theta | \alpha_1 + N_1, \alpha_2 + N_2)$$

Beta distribution



CS 3750 Advanced Machine Learning

Posterior distribution



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

CS 3750 Advanced Machine Learning

Bayesian framework

The ML estimate picks 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
- Keeps all possible parameter values
- Where $p(\theta | D, \xi) \approx \text{Beta}(\theta | \alpha_1 + N_1, \alpha_2 + N_2)$
- **The posterior can be used to define $p(A | D)$:**

$$p(A | D) = \int_{\Theta} p(A | \Theta) p(\Theta | D, \xi) d\Theta$$

Bayesian framework

- **A probability of an outcome $x=1$ in the next trial**
 $P(x=1 | D, \xi)$

$$\begin{aligned} P(x=1 | D, \xi) &= \int_0^1 \overbrace{P(x=1 | \theta, \xi) p(\theta | D, \xi)}^{\text{Posterior density}} d\theta \\ &= \int_0^1 \theta p(\theta | D, \xi) d\theta = E(\theta) \end{aligned}$$

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

$$p(\theta | D, \xi) = \text{Beta}(\theta | \alpha_1 + N_1, \alpha_2 + N_2)$$

Expected value of the parameter

How to obtain the expected value?

$$\begin{aligned} E(\theta) &= \int_0^1 \theta \text{Beta}(\theta | \eta_1, \eta_2) d\theta = \int_0^1 \theta \frac{\Gamma(\eta_1 + \eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \theta^{\eta_1-1} (1-\theta)^{\eta_2-1} d\theta \\ &= \frac{\Gamma(\eta_1 + \eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \int_0^1 \theta^{\eta_1} (1-\theta)^{\eta_2-1} d\theta \\ &= \frac{\Gamma(\eta_1 + \eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \frac{\Gamma(\eta_1 + 1)\Gamma(\eta_2)}{\Gamma(\eta_1 + \eta_2 + 1)} \underbrace{\int_0^1 \text{Beta}(\eta_1 + 1, \eta_2) d\theta}_1 \\ &= \frac{\eta_1}{\eta_1 + \eta_2} \end{aligned}$$

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

Expected value of the parameter

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

$$p(\theta | D, \xi) = \text{Beta}(\theta | \alpha_1 + N_1, \alpha_2 + N_2)$$

- **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:

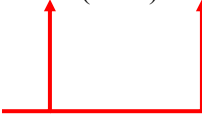
$$\hat{\theta} = E(\theta)$$

Maximum a posteriori probability

Maximum a-posteriori estimate

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

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

$$p(\theta | D, \xi) = \frac{P(D | \theta, \xi) \text{Beta}(\theta | \alpha_1, \alpha_2)}{P(D | \xi)} = \text{Beta}(\theta | \alpha_1 + N_1, \alpha_2 + N_2)$$
$$= \frac{\Gamma(\alpha_1 + \alpha_2 + N_1 + N_2)}{\Gamma(\alpha_1 + N_1) \Gamma(\alpha_2 + N_2)} \theta^{N_1 + \alpha_1 - 1} (1 - \theta)^{N_2 + \alpha_2 - 1}$$


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

MAP Solution:

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

CS 3750 Advanced Machine Learning

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

What is the MAP estimate?

CS 3750 Advanced Machine Learning

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

What is the MAP estimate ?

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

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

– **Heads:** 15

– **Tails:** 10

- Assume

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

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