Density estimation

Density estimation: is an unsupervised learning problem

- **Goal:** Learn a model that represent the relations among attributes in the data
  \[ D = \{ D_1, D_2, \ldots, D_n \} \]

**Data:** \( D_i = x_i \) 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 an underlying probability distribution model:
\[ p(X) = p(X_1, X_2, \ldots, X_d) \] 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 model of the underlying probability distribution over variables \( X \), \( p(X) \), using examples in \( D \)

---

**Density estimation**

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

**Independently drawn instances from the same fixed distribution**
Density estimation

Types of density estimation:

(1) Parametric
- the distribution is modeled using a set of parameters $\Theta$
  \[ \hat{p}(X) = p(X \mid \Theta) \]
- **Estimation**: find parameters $\Theta$ fitting the data $D$
- **Example**: estimate the mean and covariance of a normal distribution

\[ \hat{p}(x) = N(x \mid \mu, \sigma) \]

(2) Non-parametric
- The model of the distribution utilizes all examples in $D$
- As if all examples were parameters of the distribution
- \[ \hat{p}(X) = p(X \mid D) \]
- **Examples**:
  - histogram
  - Kernel density estimation
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

- How to measure the goodness of fit or alternative the error?

---

**ML Parameter estimation**

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

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

**Maximum likelihood (ML)**

- Find $\Theta$ that maximizes likelihood $p(D | \Theta, \xi)$

\[
P(D | \Theta, \xi) = P(D_1, D_2, \ldots, D_n | \Theta, \xi)
= P(D_1 | \Theta, \xi) P(D_2 | \Theta, \xi) \ldots P(D_n | \Theta, \xi)
= \prod_{i=1}^{n} P(D_i | \Theta, \xi)
\]

$\Theta_{ML} = \text{arg} \max_{\Theta} p(D | \Theta, \xi)$

**Independent examples**
Properties of log function: 

\[ \log x \]

\[ \max \arg (\log \Theta) \]

\[ \Theta^* = \arg \max \Theta f(\Theta) = \arg \max \Theta \log f(\Theta) \]
ML Parameter estimation

Model: \( \hat{p}(X) = p(X | \Theta) \)

Data: \( D = \{D_1, D_2, \ldots, D_n\} \)

- Maximum likelihood (ML)
  - Find \( \hat{\Theta} \) that maximizes likelihood \( p(D | \Theta, \xi) \)

\[
P(D | \Theta, \xi) = P(D_1, D_2, \ldots, D_n | \Theta, \xi)
= P(D_1 | \Theta, \xi)P(D_2 | \Theta, \xi) \ldots P(D_n | \Theta, \xi)
= \prod_{i=1}^{n} P(D_i | \Theta, \xi)
\]

\[
\log\text{-likelihood} \quad \log p(D | \Theta, \xi) = \sum_{i=1}^{n} \log P(D_i | \Theta, \xi)
\]

\[
\Theta_{ML} = \text{arg max } \Theta \quad p(D | \Theta, \xi) = \text{arg max } \Theta \quad \log p(D | \Theta, \xi)
\]

Bayesian parameter estimation

The ML estimate picks just one value of the parameter

- Problem: if there are two different parameter values that are close in terms of the likelihood, using only one of them may introduce a strong bias, if we use it, for example, for predictions.

Bayesian parameter estimation

- Remedies the limitation of one choice
- Uses the posterior distribution for parameters \( \Theta \)
- Posterior ‘covers’ all possible parameter values (and their “weights”)

\[
p(\Theta | D, \xi) = \frac{p(D | \Theta, \xi) p(\Theta | \xi)}{p(D | \xi)}
\]
Bayesian parameter estimation

What does it do?

• Prior and Posterior ‘covers’ all possible parameter values (and their “weights”)

Assume: we have a model of \( p(x \mid \Theta) \) with a parameter \( \Theta \)

• Bayesian parameter estimation:

\[
\begin{align*}
\text{Prior on a parameter} & \quad + \quad \text{Data} + \quad p(x \mid \Theta) \quad = \quad \text{Posterior on a parameter} \\
p(\Theta) & \quad + \quad p(x \mid \Theta) \quad = \quad p(\Theta \mid D)
\end{align*}
\]

• ML parameter estimate

\[
\begin{align*}
\text{Data} + \quad p(x \mid \Theta) \quad = \quad \text{Just one value}
\end{align*}
\]

Bayesian parameter estimation

Bayesian parameter estimation

– Uses the posterior distribution for parameters

– Posterior ‘covers’ all possible parameter values (and their “weights”)

Parameter posterior

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

• How to use the posterior for modeling \( p(X) \)?

\[
\hat{p}(X) = p(X \mid D) = \int_{\Theta} p(X \mid \Theta) p(\Theta \mid D, \xi) d\Theta
\]
Parameter estimation

Other criteria:

- **Maximum a posteriori probability (MAP)**
  
  maximize  \( p(\Theta \mid D, \xi) \) \hspace{1cm} \text{(mode of the posterior)}
  
  - Yields: one set of parameters \( \Theta_{MAP} \)
  
  - Approximation:
    \[ \hat{p}(X) = p(X \mid \Theta_{MAP}) \]

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

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 T H T T T H T H H H T H H T T 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 $\theta$
- **Data:**
  
  H H T T H H T T H T T T H T H H H T H H T T 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 $\theta$
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 $x_i$

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

- 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 \( D = H H T H T H \)
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?

Criterion for the best fit: 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 :

\[
Error(D, \theta) = -P(D \mid \theta)
\]
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
\[ \Theta_{ML} = \arg \max_{\theta} P(D \mid \Theta, \xi) = \arg \max_{\theta} \log p(D \mid \Theta, \xi) \]

Optimize log-likelihood (the same as maximizing likelihood)
\[ l(D, \theta) = \log P(D \mid \theta, \xi) = \log \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{(1-x_i)} = \]
\[ = \sum_{i=1}^{n} x_i \log \theta + (1 - x_i) \log(1 - \theta) = \log \theta \sum_{i=1}^{n} x_i + (1 - \theta) \sum_{i=1}^{n} (1 - x_i) \]

\[ N_1 \text{ - number of heads seen} \quad N_2 \text{ - number of tails seen} \]

Maximum likelihood (ML) estimate.

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

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:**
  - Heads: 15
  - Tails: 10

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

\[
\hat{\theta} = \frac{N_1}{N} = \frac{15}{25} = 0.6
\]

\[
(1 - \hat{\theta}) = \frac{N_2}{N} = \frac{10}{25} = 0.4
\]
Bayesian parameter estimation

Uses the distributions (prior and posterior) over all possible values of the parameter \( \theta \) of the sampling distribution \( p(x \mid \theta) \) (Bernoulli):

\[
p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi)p(\theta \mid \xi)}{P(D \mid \xi)}
\]

(via Bayes theorem)

Prior

Posterior

Normalizing factor

We know that the likelihood is:

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

How to choose the prior probability?

\[
p(\theta \mid \xi) \quad - \text{is the prior probability on } \theta
\]

Prior distribution

Choice of prior: Beta distribution

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

\( \Gamma(x) \) - a Gamma function \( \Gamma(x) = (x-1)! \Gamma(x-1) \)

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

Why to use Beta distribution?

Beta distribution “fits” Bernoulli sample - 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

\[
p(\theta | \xi) = \text{Beta}(\theta | a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}
\]

Posterior distribution

\[
p(\theta | \mathcal{D}, \xi) = \frac{P(\mathcal{D} | \theta, \xi) \text{Beta}(\theta | \mu_{1}, \mu_{2})}{P(\mathcal{D} | \xi)} = \text{Beta}(\theta | \alpha_{1} + N_{1}, \alpha_{2} + N_{2})
\]
**Posterior distribution**

**Beta posterior**
- A conjugate prior to Bernoulli sample

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

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

---

**Maximum a posteriori probability (MAP)**

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

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

Likelihood of data

\[
p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi)p(\theta \mid \xi)}{P(D \mid \xi)} \quad \text{(via Bayes rule)}
\]

• The model of the posterior is represented by a Beta distribution

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

Maximum a posteriori estimate
– Selects the mode of the posterior distribution
– Assumes conjugate prior to Bernoulli sample

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

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

Mode of the posterior satisfies:
\[ \frac{\partial \log p(\theta \mid D, \xi)}{\partial \theta} = 0 \]

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

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
  – 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 H T H T T H T H T H H H H T H H H H T
  
  – Heads: 15
  – Tails: 10
• Assume $p(\theta \mid \xi) = \text{Beta}(\theta \mid 5,5)$

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 H T T H T H T H H H H T H H H H T
  
  – Heads: 15
  – Tails: 10
• Assume $p(\theta \mid \xi) = \text{Beta}(\theta \mid 5,5)$

$$
\theta_{\text{MAP}} = \frac{19}{33}
$$

• Assume $p(\theta \mid \xi) = \text{Beta}(\theta \mid 5,20)$

$$
\theta_{\text{MAP}} = \frac{19}{48}
$$
Bayesian framework

- **Predictive probability of an outcome** $x = 1$ in the next trial $P(x = 1 | D, \xi)$
  
  Posterior density
  
  $$P(x = 1 | D, \xi) = \int_0^1 P(x = 1 | \theta, \xi) p(\theta | D, \xi) d\theta$$
  
  $$= \int_0^1 \theta p(\theta | D, \xi) d\theta = E(\theta)$$

- **Equivalent to the expected value of the parameter**
  - expectation is taken with respect to the posterior distribution
  
  $$p(\theta | D, \xi) = \text{Beta}(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)$$

---

Expected value of the parameter

How to calculate the expected value of Beta?

$$E(\theta) = \int_0^1 \theta \text{Beta}(\theta \mid \eta_1, \eta_2) d\theta = \int_0^1 \theta \frac{\Gamma(\eta_1 + \eta_2)}{\Gamma(\eta_1) \Gamma(\eta_2)} \theta^{\eta_1-1} (1-\theta)^{\eta_2-1} d\theta$$

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

$$= \frac{\Gamma(\eta_1 + \eta_2)}{\Gamma(\eta_1) \Gamma(\eta_2)} \frac{\Gamma(\eta_1 + 1) \Gamma(\eta_2)}{\Gamma(\eta_1 + \eta_2 + 1)} \int_0^1 \text{Beta}(\eta_1 + 1, \eta_2) d\theta$$

$$= \frac{\eta_1}{\eta_1 + \eta_2}$$

**Note:** $\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$ for integer values of $\alpha$
Expected value of the parameter

• Substituting the results for the posterior:
  \[ p(\theta \mid D, \xi) = Beta(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2) \]

• 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:
  \[ \hat{\theta} = E(\theta) \]

\[ \Theta_{EV} = E_{p(\theta \mid D, \xi)}(\Theta) = \int \Theta p(\Theta \mid D, \xi) d\Theta \]