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

**Objective:** estimate the model of the underlying probability distribution over variables \( \mathbf{X} \), \( p(\mathbf{X}) \), using examples in \( D \)

---

Density estimation

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

Independently drawn instances from the same fixed distribution
Density estimation

Types of density estimation:

**Parametric**
- the distribution is modeled using a set of parameters $\Theta$
  \[ \hat{p}(X) = p(X | \Theta) \]
- **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 $\Theta : \hat{p}(X | \Theta)$
- **Example**: Gaussian distribution with mean and variance parameters

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

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

**Model** \( \hat{p}(X) = p(X|\Theta) \)  
**Data** \( D = \{D_1, D_2, ..., D_n\} \)

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

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

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

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

\[
\hat{p}(X) = p(X|\Theta_{ML})
\]

---

**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:
  - Prior on a parameter
    \[
p(\Theta) + p(x \mid \Theta) = p(\Theta \mid D)
    \]
  - Data + \( p(x \mid \Theta) \)
  - Just one value

- ML Estimate
  \[
p(\Theta) + p(x \mid \Theta) = \text{Just one value}
  \]

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

Other criteria:

• **Maximum a posteriori probability (MAP)**
  
  maximize \( p(\Theta \mid D, \xi) \) (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**
  
  \( \hat{\Theta} = E(\Theta) \)  
  (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 H T T H T H H T H H T H H T T H 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?

\[ \tilde{\theta} = ? \]

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** $x_i$

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

*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

$$D = H H T H T H \quad \text{(encoded as } D = 110101)$$

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 \( D \):

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

\[
\text{Error}(D, \theta) = -P(D \mid \theta) 
\]
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
\[ \theta_{ML} = \arg \max_{\theta} P(D \mid \theta, \xi) \]

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

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

Maximum likelihood (ML) estimate.

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

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

\[
\theta_{ML}^{Head} = \frac{N_1}{N} = \frac{15}{25} = 0.6
\]
\[
(1 - \theta_{ML}^{Head}) = \frac{N_2}{N} = \frac{10}{25} = 0.4
\]
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):

Prior

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

(via Bayes theorem)

Likelihood of data

Posterior

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

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?

$p(\theta \mid \xi)$ - 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 $\Gamma(x) = (x-1)!$

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

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

Posterior distribution

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

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

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

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

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

---

**Maximum a posteriori probability (MAP)**

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

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

**Likelihood of data** \[p(\theta | D, \xi) = \frac{P(D | \theta, \xi)p(\theta | \xi)}{P(D | \xi)}\] (via Bayes rule)

• Selects the model of the posterior represented as a Beta distribution

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

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

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

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

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

**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 T H T H H 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?
MAP estimate example

- Assume the unknown and possibly biased coin
- Probability of the head is $\theta$
- **Data:**
  
  - Heads: 15
  - Tails: 10
- Assume $p(\theta | \xi) = Beta(\theta | 5,5)$

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:**
  
  - Heads: 15
  - Tails: 10
- Assume $p(\theta | \xi) = Beta(\theta | 5,5)$

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

$$p(\theta | \xi) = Beta(\theta | 5,20)$$

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

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

  \[
P(x = 1 \mid D, \xi) = \int P(x = 1 \mid \theta, \xi) p(\theta \mid D, \xi) d\theta
  \]

  Posterior density

\[
P(x = 1 \mid D, \xi) = \int_0^1 P(x = 1 \mid \theta, \xi) p(\theta \mid D, \xi) d\theta = E(\theta)
\]

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

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

Expected value of the parameter

**How to calculate the expected value of Beta?**

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

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

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

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

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

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

- **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:**
  \[ \hat{\theta} = E(\theta) \]

---

**Binomial distribution**

**Example problem:** a biased coin  
**Outcomes:** two possible values -- head or tail  
**Data:** a set of order-independent outcomes for N trials  
\[ N_1 \text{ - number of heads seen} \quad N_2 \text{ - number of tails seen} \]

**Model:** probability of a head \( \theta \)  
probability of a tail \( 1 - \theta \)

**Probability of an outcome**

\[ P(N_1 \mid N, \theta) = \binom{N}{N_1} \theta^{N_1} (1 - \theta)^{N - N_1} \quad \text{Binomial distribution} \]

**Objective:**  
We would like to estimate the probability of a head \( \hat{\theta} \)
Binomial distribution

Example problem: $N$ coin flips, where each coin flip can have two results: head or tail

Outcome: $N_1$ - number of heads seen $N_2$ - number of tails seen in $N$ trials

Model: probability of a head $\theta$ probability of a tail $(1-\theta)$

Probability of an outcome:

$$P(N_1 \mid N, \theta) = \binom{N}{N_1} \theta^{N_1} (1 - \theta)^{N-N_1}$$

Binomial distribution:
- models order independent sequence of Bernoulli trials

![Binomial distribution graph](image)
Maximum likelihood (ML) estimate.

Likelihood of data:
\[
P(D | \theta) = \binom{N}{N_1} \theta^{N_1} (1 - \theta)^{N_2} = \frac{N!}{N_1! N_2!} \theta^{N_1} (1 - \theta)^{N_2}
\]

Log-likelihood
\[
l(D, \theta) = \log \left( \binom{N}{N_1} \theta^{N_1} (1 - \theta)^{N_2} \right) = \log \frac{N!}{N_1! N_2!} + N_1 \log \theta + N_2 \log(1 - \theta)
\]

Constant from the point of optimization !!!

ML Solution:
\[
\theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2}
\]
The same as for Bernoulli and \( D \) with iid sequence of examples

Posterior density

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

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

Likelihood
\[
P(D | \theta) = \frac{\Gamma(N_1 + N_2)}{\Gamma(N_1) \Gamma(N_2)} \theta^{N_1} (1 - \theta)^{N_2}
\]

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

MAP estimate
\[
\theta_{MAP} = \arg \max_{\theta} p(\theta | D, \xi)
\]
\[
\theta_{MAP} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}
\]
**Multinomial distribution**

**Example:** multiple rolls of a dice with 6 results

**Outcome:** counts of occurrences of \( k \) possible outcomes of \( N \) trials: \( N_i \) - a number of times an outcome \( i \) has been seen

\[
\sum_{i=1}^{k} N_i = N
\]

**Model parameters:** \( \Theta = (\theta_1, \theta_2, \ldots, \theta_k) \) s.t. \( \sum_{i=1}^{k} \theta_i = 1 \)

- \( \theta_i \) - probability of an outcome \( i \)

**Probability distribution:**

\[
P(N_1, N_2, \ldots, N_k \mid \Theta, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \prod_{i=1}^{k} \theta_i^{N_i} \]

**ML estimate:**

\[
\theta_{i, ML} = \frac{N_i}{N}
\]

---

**Posterior and MAP estimate**

**Choice of the prior:** Dirichlet distribution

\[
Dir(\Theta \mid \alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}
\]

Dirichlet is the conjugate choice for the multinomial sampling

\[
P(D \mid \Theta, \xi) = P(N_1, N_2, \ldots, N_k \mid \Theta, \xi) = \frac{N!}{N_1! N_2! \ldots N_k!} \prod_{i=1}^{k} \theta_i^{N_i} \]

**Posterior density**

\[
p(\Theta \mid D, \xi) = \frac{P(D \mid \Theta, \xi)Dir(\Theta \mid \alpha_1, \alpha_2, \ldots, \alpha_k)}{P(D \mid \xi)} = Dir(\Theta \mid \alpha_1 + N_1, \ldots, \alpha_k + N_k)
\]

**MAP estimate:**

\[
\hat{\theta}_{i, MAP} = \frac{\alpha_i + N_i - 1}{\sum_{i=1}^{k} (\alpha_i + N_i) - k}
\]
Dirichlet distribution

Dirichlet distribution:

\[ \text{Dir}(\theta | \alpha_1, \ldots, \alpha_k) = \frac{\Gamma(k) \prod_{i=1}^{k} \Gamma(\alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \cdots \theta_k^{\alpha_k-1} \]

Assume: \( k=3 \)

Other distributions

The same ideas can be applied to other distributions

- Typically we choose distributions that behave well so that computations lead to “nice” solutions

- Exponential family of distributions

  Conjugate choices for some of the distributions from the exponential family:
  - Binomial – Beta
  - Multinomial - Dirichlet
  - Exponential – Gamma
  - Poisson – Inverse Gamma
  - Gaussian - Gaussian (mean) and Wishart (covariance)
Gaussian (normal) distribution

- **Gaussian:** $x \sim N(\mu, \sigma)$
- **Parameters:**
  - $\mu$ - mean
  - $\sigma$ - standard deviation
- **Density function:**
  $$p(x | \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2} (x - \mu)^2 \right]$$
- **Example:**

![Gaussian distribution graph](image)

Parameter estimates

- **Loglikelihood**
  $$l(D, \mu, \sigma) = \log \prod_{i=1}^{n} p(x_i | \mu, \sigma)$$
- **ML estimates of the mean and variance:**
  $$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \hat{\sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$
  - ML variance estimate is biased
  $$E_n(\sigma^2) = E_n\left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \right) = \frac{n-1}{n} \sigma^2 \neq \sigma^2$$
- **Unbiased estimate:**
  $$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$
Multivariate normal distribution

- **Multivariate normal:** \( \mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma) \)
- **Parameters:**
  - \( \boldsymbol{\mu} \) - mean
  - \( \Sigma \) - covariance matrix
- **Density function:**
  \[
  p(\mathbf{x} | \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]
  \]
- **Example:**

Partitioned Gaussian Distributions

- **Multivariate Gaussian:**

- **Example:**

  - Precision matrix

  - What are the distributions for marginals and conditionals?

  \[
  p(x_a), \quad p(x_a | x_b)
  \]
Partitioned Conditionals and Marginals

- Conditional density:

- Marginal Density:
Parameter estimates

- Loglikelihood
  \[ l(D, \mu, \Sigma) = \log \prod_{i=1}^{n} p(x_i | \mu, \Sigma) \]

- ML estimates of the mean and covariances:
  \[
  \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T
  \]
  
  - Covariance estimate is biased
  \[
  E_n(\hat{\Sigma}) = E_n \left( \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T \right) = \frac{n-1}{n} \Sigma \neq \Sigma
  \]

  - Unbiased estimate:
  \[
  \hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T
  \]

Posterior of a multivariate normal

- Assume a prior on the mean \( \mu \) that is normally distributed:
  \[ p(\mu) \approx N(\mu_p, \Sigma_p) \]

- Then the posterior of \( \mu \) is normally distributed
  \[
  p(\mu | D) \approx \left( \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right] \right) \]
  \[
  \times \frac{1}{(2\pi)^{d/2} |\Sigma_p|^{1/2}} \exp \left[ -\frac{1}{2} (\mu - \mu_p)^T \Sigma_p^{-1} (\mu - \mu_p) \right]
  \]
  \[
  = \frac{1}{(2\pi)^{d/2} |\Sigma_n|^{1/2}} \exp \left[ -\frac{1}{2} (\mu - \mu_n)^T \Sigma_n^{-1} (\mu - \mu_n) \right]
  \]

CS 2750 Machine Learning
Posterior of a multivariate normal

- Then the posterior of $\mu$ is normally distributed

$$p(\mu \mid D) = \frac{1}{(2\pi)^{d/2}(\Sigma_n)^{1/2}} \exp \left[-\frac{1}{2}(\mu - \mu_n)^T \Sigma_n^{-1}(\mu - \mu_n)\right]$$

$$\Sigma_n^{-1} = n\Sigma^{-1} + \Sigma_p$$

$$\mu_n = \Sigma_p \left(\Sigma_p + \frac{1}{n} \Sigma\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) + \frac{1}{n} \Sigma \left(\Sigma_p + \frac{1}{n} \Sigma\right)^{-1} \mu_p$$

$$\Sigma_n = \Sigma_p \left(\Sigma_p + \frac{1}{n} \Sigma\right)^{-1} \frac{1}{n} \Sigma$$

Other distributions

**Gamma distribution:**

$$p(x \mid a, b) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-\frac{x}{b}} \quad \text{for } x \in [0, \infty]$$

**Exponential distribution:**

- A special case of Gamma for $a=1$

$$p(x \mid b) = \left(\frac{1}{b}\right) e^{-\frac{x}{b}}$$

**Poisson distribution:**

$$p(x \mid \lambda) = \frac{e^{-\lambda} \lambda^x}{x!} \quad \text{for } x \in \{0, 1, 2, \ldots\}$$
Other distributions

**Gamma distribution:**

\[ p(\lambda \mid a, b) = \frac{1}{\Gamma(a)b^a} \lambda^{a-1} e^{-\frac{\lambda}{b}} \quad \text{for} \quad \lambda \in [0, \infty] \]

Sequential Bayesian parameter estimation

- **Sequential Bayesian approach**
  - Under the iid the estimates of the posterior can be computed incrementally for a sequence of data points

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

- If we use a conjugate prior we get back the same posterior
- Assume we split the data D in the last element x and the rest

\[
p(D \mid \Theta) = p(x \mid \Theta)p(D_{n-1} \mid \Theta)
\]

- Then: A “new” prior

\[
p(\Theta \mid D, \xi) = \frac{p(x \mid \Theta)p(D_{n-1} \mid \Theta)p(\Theta \mid \xi)}{\int_{\Theta} p(x \mid \Theta)p(D_{n-1} \mid \Theta)p(\Theta \mid \xi) d\Theta}
\]

CS 2750 Machine Learning
Exponential family

Exponential family:
- all probability mass / density functions that can be written in the exponential normal form
  \[ f(x | \eta) = \frac{1}{Z(\eta)} h(x) \exp[\eta^T t(x)] \]
- \( \eta \) a vector of natural (or canonical) parameters
- \( t(x) \) a function referred to as a sufficient statistic
- \( h(x) \) a function of x (it is less important)
- \( Z(\eta) \) a normalization constant (a partition function)

Other common form:

\[ f(x | \eta) = h(x) \exp[\eta^T t(x) - A(\eta)] \quad \log Z(\eta) = A(\eta) \]

Exponential family: examples

- **Bernoulli distribution**
  \[ p(x | \pi) = \pi^x (1 - \pi)^{1-x} \]
  \[ = \exp \left\{ \log \left( \frac{\pi}{1 - \pi} \right) x + \log(1 - \pi) \right\} \]
  \[ = \exp \{ \log(1 - \pi) \} \exp \left\{ \log \left( \frac{\pi}{1 - \pi} \right) x \right\} \]

- **Exponential family**

\[ f(x | \eta) = \frac{1}{Z(\eta)} h(x) \exp[\eta^T t(x)] \]

- **Parameters**
  \[ \eta = ? \]
  \[ t(x) = ? \]
  \[ Z(\eta) = ? \]
  \[ h(x) = ? \]
Exponential family: examples

• Bernoulli distribution
  \[ p(x \mid \pi) = \pi^x (1 - \pi)^{1-x} \]
  \[ = \exp \left\{ \log \left( \frac{\pi}{1-\pi} \right) x + \log(1 - \pi) \right\} \]
  \[ = \exp \left\{ \log(1 - \pi) \right\} \exp \left\{ \log \left( \frac{\pi}{1-\pi} \right) x \right\} \]

• Exponential family
  \[ f(x \mid \eta) = \frac{1}{Z(\eta)} h(x) \exp \left[ \eta^T t(x) \right] \]

• Parameters
  \[ \eta = \log \frac{\pi}{1-\pi} \] (note \( \pi = \frac{1}{1 + e^{-\eta}} \)) \( t(x) = x \)
  \[ Z(\eta) = \frac{1}{1 - \pi} = 1 + e^\eta \]
  \[ h(x) = 1 \]

Exponential family: examples

• Univariate Gaussian distribution
  \[ p(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ - \frac{1}{2\sigma^2} (x - \mu)^2 \right] \]
  \[ = \frac{1}{2\pi} \exp \left( - \frac{1}{2\sigma^2} - \log \sigma \right) \exp \left\{ \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 \right\} \]

• Exponential family
  \[ f(x \mid \eta) = \frac{1}{Z(\eta)} h(x) \exp \left[ \eta^T t(x) \right] \]

• Parameters
  \[ \eta = ? \]
  \[ t(x) = ? \]
  \[ Z(\eta) = ? \]
  \[ h(x) = ? \]
Exponential family: examples

- **Univariate Gaussian distribution**
  
  \[ p(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[ -\frac{1}{2\sigma^2} (x - \mu)^2 \right] \]
  
  \[ = \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{\mu}{2\sigma^2} - \log \sigma \right) \exp\left\{ \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 \right\} \]

- **Exponential family**
  
  \[ f(x \mid \eta) = \frac{1}{Z(\eta)} h(x) \exp[\eta^T t(x)] \]

- **Parameters**
  
  \[ \eta = \begin{bmatrix} \mu / 2\sigma^2 \\ -1 / 2\sigma^2 \end{bmatrix} \]
  
  \[ t(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \]

  \[ Z(\eta) = \exp\left\{ \frac{\mu^2}{2\sigma^2} + \log \sigma \right\} = \exp\left\{ -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2} \log\left( -2\eta_2 \right) \right\} \]

  \[ h(x) = 1 / \sqrt{2\pi} \]

---

Exponential family

- For iid samples, the likelihood of data is

  \[ P(D \mid \eta) = \prod_{i=1}^{n} p(x_i \mid \eta) = \prod_{i=1}^{n} h(x_i) \exp[\eta^T t(x_i) - A(\eta)] \]

  \[ = \prod_{i=1}^{n} h(x_i) \exp \left[ \sum_{i=1}^{n} \eta^T t(x_i) - A(\eta) \right] \]

  \[ = \prod_{i=1}^{n} h(x_i) \exp \left[ \eta^T \left( \sum_{i=1}^{n} t(x_i) \right) - nA(\eta) \right] \]

- **Important:**
  
  - the dimensionality of the sufficient statistic remains the same with the number of samples
Exponential family

• The log likelihood of data is

\[ l(D, \eta) = \log \left[ \prod_{i=1}^{n} h(x_i) \right] \exp \left[ \eta^T \left( \sum_{i=1}^{n} t(x_i) \right) - nA(\eta) \right] \]

\[ = \log \left[ \prod_{i=1}^{n} h(x_i) \right] + \left[ \eta^T \left( \sum_{i=1}^{n} t(x_i) \right) - nA(\eta) \right] \]

• Optimizing the loglikelihood

\[ \nabla_\eta l(D, \eta) = \left( \sum_{i=1}^{n} t(x_i) \right) - n \nabla_\eta A(\eta) = 0 \]

• For the ML estimate it must hold

\[ \nabla_\eta A(\eta) = \frac{1}{n} \left( \sum_{i=1}^{n} t(x_i) \right) \]

Exponential family

• Rewritting the gradient:
Exponential family

- **Rewritting the gradient:**

\[
\nabla_\eta A(\eta) = \nabla_\eta \log Z(\eta) = \nabla_\eta \log \int h(x) \exp \left\{ \eta^T t(x) \right\} dx
\]

\[
\nabla_\eta A(\eta) = \frac{\int t(x) h(x) \exp \left\{ \eta^T t(x) \right\} dx}{\int h(x) \exp \left\{ \eta^T t(x) \right\} dx}
\]

\[
\nabla_\eta A(\eta) = \int t(x) h(x) \exp \left\{ \eta^T t(x) - A(\eta) \right\} dx
\]

\[
\nabla_\eta A(\eta) = E(t(x))
\]

- **Result:**

\[
E(t(x)) = \frac{1}{n} \left( \sum_{i=1}^{n} t(x_i) \right)
\]

- For the ML estimate the parameters  $\eta$ should be adjusted such that the expectation of the statistic $t(x)$ is equal to the observed sample statistics.

Moments of the distribution

- **For the exponential family**
  - The k-th moment of the statistic corresponds to the k-th derivative of $A(\eta)$
  - If $x$ is a component of $t(x)$ then we get the moments of the distribution by differentiating its corresponding natural parameter

- **Example: Bernoulli** $p(x \mid \pi) = \exp \left\{ \log \left( \frac{\pi}{1-\pi} \right) x + \log(1 - \pi) \right\}$

\[
A(\eta) = \log \frac{1}{1-\pi} = \log(1 + e^\eta)
\]

- **Derivatives:**

\[
\frac{\partial A(\eta)}{\partial \eta} = \frac{\partial}{\partial \eta} \log(1 + e^\eta) = \frac{e^\eta}{1 + e^\eta} = \frac{1}{1 + e^{-\eta}} = \pi
\]

\[
\frac{\partial A(\eta)}{\partial \eta^2} = \frac{\partial}{\partial \eta} \frac{1}{1 + e^{-\eta}} = \pi(1 - \pi)
\]
Multivariate normal distribution

- **Multivariate normal**: \( \mathbf{x} \sim N(\mathbf{\mu}, \mathbf{\Sigma}) \)
- **Parameters**: 
  - \( \mathbf{\mu} \): mean
  - \( \mathbf{\Sigma} \): covariance matrix
- **Density function**:

  \[
  p(\mathbf{x} | \mathbf{\mu}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \mathbf{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{\mu})\right]
  \]

- **Example**: 

![Multivariate normal distribution graph](image)
Parameter estimates

- **Loglikelihood**
  \[ l(D, \mu, \Sigma) = \log \prod_{i=1}^{n} p(x_i | \mu, \Sigma) \]

- **ML estimates of the mean and covariances:**
  \[ \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \]
  \[ \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T \]
  - Covariance estimate is biased
  \[ E_n(\hat{\Sigma}) = E_n \left( \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T \right) = \frac{n-1}{n} \Sigma \neq \Sigma \]

- **Unbiased estimate:**
  \[ \hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T \]

---

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 \( \Theta \)
- **Data** \( D = \{ D_1, D_2, \ldots, D_n \} \)

**Objective:** find parameters \( \hat{\Theta} \) that fit the data the best

What is the best set of parameters? There are various criteria one can apply here …
Parameter estimation.

- **Maximum likelihood (ML)**
  
  
  \[
  \text{maximize } p(D \mid \Theta, \xi) \\
  \xi \text{ - represents prior (background) knowledge}
  \]

- **Maximum a posteriori probability (MAP)**
  
  \[
  \text{maximize } p(\Theta \mid D, \xi) \\
  \text{Selects the mode of the posterior}
  \]

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

- **Bayesian framework**
  
  - use a posterior density
  - no optimization

Posterior of a multivariate normal

- **Assume that we use only a prior on the mean:** \( \mu \)
- **A prior** 
  
  \[
  \mu \approx N(\mu_p, \Sigma_p)
  \]

- **Then the posterior is:**
  
  - **Normally**
    
    \[
    p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]
    \]

- **ML estimates of the mean and covariances:**
  
  \[
  \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T
  \]
  
  - Covariance estimate is biased
    
    \[
    E_n (\Sigma) = E_n \left( \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T \right) = \frac{n-1}{n} \Sigma \neq \Sigma
    \]
  - **Unbiased estimate:**
    
    \[
    \Sigma = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T
    \]
Parameter estimates

- **Loglikelihood**
  \[ l(D, \mu, \Sigma) = \log \prod_{i=1}^{n} p(x_i | \mu, \Sigma) \]

- **ML estimates of the mean and covariances:**
  \[
  \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \\
  \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T
  \]

- **Unbiased estimate:**
  \[
  \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T
  \]

Unsupervised learning

- **Data:** \( D = \{D_1, D_2, ..., D_n\} \)
  \( D_i = x_i \) a vector of attribute values
  - e.g. the description of a patient
  - no specific target attribute we want to predict (no output \( y \))

- **Objective:**
  - learn (describe) relations between attributes, examples

**Types of problems:**

- **Clustering**
  Group together “similar” examples

- **Density estimation**
  - Model probabilistically the population of examples
### Beta distribution

- $\alpha_1 = \alpha_2 = 0.5$
- $\alpha_1 = 2.5, \alpha_2 = 5$
- $\alpha_1 = \alpha_2 = 2.5$

### Exponential family

- **Exponential family of distributions**

$$f(x, \theta, \phi) = \exp \left\{ \frac{(\theta x - b(\theta))}{a(\phi)} + c(x, \phi) \right\}$$

- **Parameters:**
  - $\theta$ - location parameters
  - $\phi$ - scaling parameters

- **Example:**
  - $$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]$$
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