Density estimation

Design cycle

Data cleaning and preprocessing

Data → Feature selection → Model selection → Learning → Evaluation
**Design cycle**

Data → Feature selection → Model selection → Learning → Evaluation

- **Feature selection**
  - Reduce the dimensionality of data, especially if the sample size is small

- **Model selection**
  - Select a class of models among which to search for the model (by human, semi-automatic)
Design cycle

Data

Feature selection

Model selection

Learning

Evaluation

Find the best model according to some optimization criterion
• efficiency matters

Assess the quality of the model
Evaluation of learning models

**Simple holdout method**
- Divide the data to the training and test data

- Typically 2/3 training and 1/3 testing

Other more complex methods
- Use multiple train/test sets
- Based on various random re-sampling schemes:
  - Random sub-sampling
  - Cross-validation
  - Bootstrap
**Evaluation**

- **Random sub-sampling**
  - Repeat a simple holdout method $k$ times

  ![Diagram](image)

  - Split randomly into 70% Train, 30% Test
  - Train
  - Test
  - Learning
  - Classify/Evaluate
  - Average Stats

**Cross-validation (k-fold)**

- Divide data into $k$ disjoint groups, test on $k$-th group/train on the rest
- Typically 10-fold cross-validation
- Leave one out cross-validation ($k = \text{size of the data } D$)

![Diagram](image)

- Split into $k$ groups of equal size
- Test = $i$th group, Train on the rest
- Train
- Test
- Learning
- Classify/Evaluate
- Average Stats
Evaluation

Bootstrap
- The training set of size $N = \text{size of the data } D$
- Sampling with the replacement

Data

Generate the training set of size $N$ with replacement, the rest goes to the test set

Train

Test

Learning

Classify/Evaluate

Average Stats

Design cycle

Feeding back the evaluation results may help to choose a better model
- but then be aware that you are picking a winner

Evaluation statistics for the winner model may not reflect its true performance

Fix: Add one more evaluation step
What if we want to compare the predictive performance on a classification or a regression problem for two different learning methods?

**Solution:** compare the error results on the test data set or the average statistics on the same training/testing data splits

**Answer:** the method with better (smaller) testing error gives a better generalization error.

But we need to use statistics to validate the choice.

---

**Outline**

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

Data: 
\[ D = \{D_1, D_2, \ldots, D_n\} \]
\[ D_i = x_i \quad \text{a vector of attribute values} \]

Attributes:
- modeled by random variables \( X = \{X_1, X_2, \ldots, X_d\} \) with:
  - Continuous values
  - Discrete values

E.g. blood pressure with numerical values
or chest pain with discrete values
[no-pain, mild, moderate, strong]

Underlying true probability distribution:
\[ p(X) \]

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

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

Types of density estimation:

**Parametric**
- the distribution is modeled using a set of parameters $\Theta$
  
  $p(X \mid \Theta)$
- **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

**Semi-parametric**

---

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 $\Theta$ : $\hat{p}(X \mid \Theta)$

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

**Objective**: find parameters $\Theta$ such that $p(X \mid \Theta)$ describes data $D$ the best
Parameter estimation.

- **Maximum likelihood (ML)**
  
  maximize \( p(D \mid \Theta, \xi) \)
  
  - yields: one set of parameters \( \Theta_{ML} \)
  
  - the target distribution is approximated as:
  
  \[ \hat{p}(X) = p(X \mid \Theta_{ML}) \]

- **Bayesian parameter estimation**
  
  - uses the posterior distribution over possible parameters
  
  \[ p(\Theta \mid D, \xi) = \frac{p(D \mid \Theta, \xi) p(\Theta \mid \xi)}{p(D \mid \xi)} \]
  
  - Yields: all possible settings of \( \Theta \) (and their “weights”)
  
  - The target distribution is approximated as:
  
  \[ \hat{p}(X) = p(X \mid D) = \int p(X \mid \Theta) p(\Theta \mid D, \xi) d\Theta \]

Parameter estimation.

**Other possible 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:**

  \[
  \begin{array}{cccccccccccc}
  H & H & T & T & H & H & T & T & T & H & T & H & H & H & H & H & H & H & T \\
  \end{array}
  \]

  – **Heads:** 15
  – **Tails:** 10

What would be your estimate of the probability of a head $\hat{\theta}$ ?

$\hat{\theta} = ?$
Parameter estimation. Example

• **Assume** the unknown and possibly biased coin
• Probability of the head is $\theta$
• **Data:**
  
  H H T T H H T H T T H T H T H T H H T H H H 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

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

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**

Can be rewritten using the Bernoulli distribution:

$$P(D \mid \theta) = \prod_{i=1}^{6} \theta^{x_i} (1 - \theta)^{(1-x_i)}$$
The goodness of fit to the data.

Learning: we do not know the value of the parameter $\theta$

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:

$$\text{Error } (D, \theta) = -P(D \mid \theta)$$

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)
\[
\begin{align*}
\ell(D, \theta) &= \log P(D \mid \theta, \xi) = \log \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{(1-x_i)} \\
&= \sum_{i=1}^{n} x_i \log \theta + (1-x_i) \log(1-\theta) = \log \theta \sum_{i=1}^{n} x_i + \log(1-\theta) \sum_{i=1}^{n} (1-x_i)
\end{align*}
\]

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

CS 2750 Machine Learning

Maximum likelihood (ML) estimate.

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

Set derivative to zero
\[ \frac{\partial \ell(D, \theta)}{\partial \theta} = \frac{N_1}{\theta} - \frac{N_2}{1-\theta} = 0 \]

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

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

CS 2750 Machine Learning
Maximum likelihood estimate. Example

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

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

\[
\theta_{\text{ML}} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} = \frac{15}{25} = 0.6
\]

\[
(1 - \theta_{\text{ML}}) = \frac{N_2}{N} = \frac{N_2}{N_1 + N_2} = \frac{10}{25} = 0.4
\]
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

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

\( \Gamma(x) \) - 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 trials - conjugate choices

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

Posterior distribution is again a Beta distribution

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

\[
p(\theta | D, \xi) = \frac{P(D | \theta, \xi) \text{Beta}(\theta | \alpha_1, \alpha_2)}{P(D | \xi)} = \text{Beta}(\theta | \alpha_1 + N_1, \alpha_2 + N_2)
\]
Maximum 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 T H 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:
  H H T T H H T H T T T H H T H H H H H T H H H H T
  – Heads: 15
  – Tails: 10
• Assume $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$

What is the MAP estimate?

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

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

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

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

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