Machine learning: Density estimation

Data:
- $D = \{D_1, D_2, \ldots, D_n\}$
- $D_i = x_i$, 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

true distribution \( p(X) \) \n
\[ D = \{ D_1, D_2, \ldots, D_n \} \]

estimate \( \hat{p}(X) \)

**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

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. Basic criteria.

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

  \(\xi\) - represents prior (background) knowledge

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

  Selects the mode of the posterior

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

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:**
  
  - 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:**
  
  - Heads: 15
  - Tails: 10

What would be your choice of the probability of a head?

**Solution:** use frequencies of outcomes 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 $\theta$ (0.6)
- probability of a tail $(1 - \theta)$ (0.4)

**Assume:** we know the probability $\theta$

**Probability of an outcome of a coin flip $x_i$**

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

- Combines the probability of a head and a tail
- So that $x_i$ is going to pick its correct probability
- Gives $\theta$ or $0.6$ for $x_i = 1$
- Gives $(1 - \theta)$ or $0.4$ 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 $\theta$ (0.6)
- probability of a tail $(1 - \theta)$ (0.4)

**Assume:** a sequence of independent coin flips

$D = \text{H H T H T H}$ (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 $\theta$ (0.6)
probability of a tail $(1-\theta)$ (0.4)

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 $D$:

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

$$P(D | \theta) = 0.6 \times 0.6 \times 0.4 \times 0.6 \times 0.4 \times 0.6 = 0.6^4 \times 0.4^2$$

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 $D$:

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

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?

**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:

\[
Error(D, \theta) = -P(D \mid \theta)
\]
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) 
\[ l(D, \theta) = \log P(D \mid \theta, \xi) = \log \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} = \sum_{i=1}^{n} x_i \log \theta + (1-x_i) \log(1-\theta) = \log \theta \sum_{i=1}^{n} x_i + \log(1-\theta) \sum_{i=1}^{n} (1-x_i) \]

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

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:**
  - **Heads:** 15
  - **Tails:** 10

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

\[
\begin{align*}
\text{Head:} & \quad \theta_{ML} = \frac{N_1}{N} = \frac{N_1}{N_1 + N_2} = \frac{15}{25} = 0.6 \\
\text{Tail:} & \quad (1 - \theta_{ML}) = \frac{N_2}{N} = \frac{N_2}{N_1 + N_2} = \frac{10}{25} = 0.4
\end{align*}
\]
Learning of BBN parameters. Example.

Example:

\[
\begin{array}{c}
P(\text{Pneumonia}) \\
T & F \\
? & ? \\
\end{array}
\]

\[
\begin{array}{c}
P(\text{HWBC|Pneum}) \\
Pn & T & F \\
T & ? & ? \\
F & ? & ? \\
\end{array}
\]

\[
\begin{array}{c}
P(\text{Paleness|Pneum}) \\
P(\text{Fever|Pneum}) \\
P(\text{Cough|Pneum}) \\
? & ? & ? \\
\end{array}
\]

Data D (different patient cases):

<table>
<thead>
<tr>
<th>Pal</th>
<th>Fev</th>
<th>Cou</th>
<th>HWB</th>
<th>Pneu</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

CS 1571 Intro to AI
Estimates of parameters of BBN

- Much like multiple coin tosses
- A “smaller” learning problem corresponds to the learning of exactly one conditional distribution
- Example: \( P(\text{Fever} \mid \text{Pneumonia} = T) \)
- Problem: How to pick the data to learn?

Learning of BBN parameters. Example.

Data D (different patient cases):

<table>
<thead>
<tr>
<th>Pal</th>
<th>Fev</th>
<th>Cou</th>
<th>HWB</th>
<th>Pneu</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

How to estimate: \( P(\text{Fever} \mid \text{Pneumonia} = T) = ? \)
Learning of BBN parameters. Example.

Learn: \( P(\text{Fever} | \text{Pneumonia} = T) \)

Step 1: Select data points with Pneumonia=T

<table>
<thead>
<tr>
<th>Pal</th>
<th>Fev</th>
<th>Cou</th>
<th>HWB</th>
<th>Pneu</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

Learning of BBN parameters. Example.

Learn: \( P(\text{Fever} | \text{Pneumonia} = T) \)

Step 1: Ignore the rest

<table>
<thead>
<tr>
<th>Pal</th>
<th>Fev</th>
<th>Cou</th>
<th>HWB</th>
<th>Pneu</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>
Learning of BBN parameters. Example.

Learn:  \( P(Fever \mid Pneumonia = T) \)

Step 2: Select values of the random variable defining the distribution of Fever

<table>
<thead>
<tr>
<th>Pal</th>
<th>Fev</th>
<th>Cou</th>
<th>HWB</th>
<th>Pneu</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

Learning of BBN parameters. Example.

Learn:  \( P(Fever \mid Pneumonia = T) \)

Step 2: Ignore the rest

Fev
F
F
T
T
T
Learning of BBN parameters. Example.

Learn: \( P(\text{Fever} | \text{Pneumonia} = T) \)

Step 3: Learning the ML estimate

\[
\begin{array}{c|c}
\text{Fever} & \text{F} \\
\text{F} & \text{T} \\
\text{T} & \text{T} \\
\text{T} & \text{T} \\
\end{array}
\]

\[
P(\text{Fever} | \text{Pneumonia} = T) = \begin{cases} 0.6 & \text{if} \text{Fever} = \text{T} \\ 0.4 & \text{if} \text{Fever} = \text{F} \end{cases}
\]

Maximum a posteriori estimate

Maximum a posteriori estimate

- Selects the mode of the posterior distribution

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

Likelihood of data

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

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

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

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

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

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

Beta distribution

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

CS 2710 Foundations of AI
\[
p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) \text{Beta}(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = \text{Beta}(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2)
\]

**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:**
\[
\theta_{\text{MAP}} = \frac{\alpha_1 + N_1 - 1}{\alpha_1 + \alpha_2 + N_1 + N_2 - 2}
\]
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 H H H T H H T
  
  - **Heads:** 15
  - **Tails:** 10
- Assume $p(\theta \mid \xi) = Beta(\theta \mid 5,5)$

What is the MAP estimate?

$$\theta_{MAP} = \frac{N_1 + \alpha_1 - 1}{N_1 + N_2 + \alpha_1 + \alpha_2 - 2} = \frac{19}{33}$$
**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 T H T H T H H H T H H H T T
  
  - Heads: 15
  - Tails: 10

- Assume
  
  \[
  p(\theta \mid \xi) = \text{Beta}(\theta \mid 5,5) \quad \theta_{\text{MAP}} = \frac{19}{33}
  \]
  
  \[
  p(\theta \mid \xi) = \text{Beta}(\theta \mid 5,20) \quad \theta_{\text{MAP}} = \frac{19}{48}
  \]

---

**Learning of BBN parameters**

**Learn:** \( P(\text{Fever} \mid \text{Pneumonia} = T) \)

Assume the prior

\[
\theta_{\text{Fever} \mid \text{Pneumonia} = T} \sim \text{Beta}(3,4)
\]

Fever:

\[\text{Fever} \quad \text{F} \quad \text{T} \quad \text{T} \quad \text{T}\]

Posterior:

\[
\theta_{\text{Fever} \mid \text{Pneumonia} = T} \sim \text{Beta}(6,6)
\]

\[
\theta_{\text{MAP}}^{\text{Fever} \mid \text{Pneumonia} = T} = \frac{6 - 1}{6 + 6 - 2} = 0.5
\]

MAP estimates

<table>
<thead>
<tr>
<th>T</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>