Density estimation

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

- Well-defined theory for representing and manipulating uncertainty
- **Axioms of probability:**

  Let $A$ and $B$ be two events. Then:
  
  1. $0 \leq P(A) \leq 1$
  2. $P(\text{True}) = 1$ and $P(\text{False}) = 0$
  3. $P(A \lor B) = P(A) + P(B) - P(A \land B)$
Probability

• Let $A$ be an event, and $\neg A$ its complement.
  
  Then

\[
P(A) + P(\neg A) = 1
\]

\[
P(A \land \neg A) = 0
\]

\[
P(\text{False}) = 0
\]

\[
P(A \lor \neg A) = 1
\]

\[
P(\text{True}) = 1
\]
Joint probability

**Joint probability:**

- **Let A and B be two events.** The probability of an event A, B occurring jointly

  \[ P(A \land B) = P(A, B) \]

We can add more events, say, A, B, C

\[ P(A \land B \land C) = P(A, B, C) \]
Independence

Independence :

- Let A, B be two events. The events are independent if:

\[ P(A, B) = P(A)P(B) \]
Conditional probability:

- Let $A$, $B$ be two events. The conditional probability of $A$ given $B$ is defined as:

$$P(A | B) = \frac{P(A, B)}{P(B)}$$

Product rule:

- A rewrite of the conditional probability

$$P(A, B) = P(A | B)P(B)$$
Bayes theorem

\[ P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \]

Why?

\[ P(A \mid B) = \frac{P(A, B)}{P(B)} \]

\[ P(A, B) = P(B \mid A)P(A) \]

\[ P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \]
Random variable

A function that maps observed quantities to real valued outcomes

Binary random variables:

Mapped to 0,1

Example: Tail mapped to 0, Head mapped to 1

Note: Only one value for each outcome: either 0 or 1

\[ P(x = 0) \] probability of tail
\[ P(x = 1) \] probability of head

• Probability distribution:

\[ P(x) = \begin{bmatrix} 0.45 \\ 0.55 \end{bmatrix} \]

Assigns a probability to each possible outcome
Random variable

Discrete

– $x=0,1$ based on tail/head coin toss
– $x=1,2,3,4,5,6$ based on the roll of a dice
– $p(x)$ – assigns a probability to each possible outcomes

• Continuous

– $x$ height of a person
– $p(x)$ defined in terms of the probability density function

$$\int p(x)dx = 1$$
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 underlying 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 \( p(X) \))
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
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 \( p(D | \Theta, \xi) \)

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

\[
\log p(D | \Theta, \xi) = \sum_{i=1}^{n} \log P(D_i | \Theta, \xi)
\]
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
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** \( x_i \)

\[
P(x_i \mid \theta) = \theta^{x_i} (1 - \theta)^{1-x_i} \quad \text{Bernoulli distribution}
\]

- Combines the probability of a head and a tail
- So that \( x_i \) is going to pick its correct probability
- Gives \( \theta \) 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 coin flips $D = H H T H T H$
encoded as $D = 110101$

What is the probability of observing a data sequence $D$:

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

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

\[ \sum_{i=1}^{n} x_i \log \theta + (1 - x_i) \log(1 - \theta) = \log \theta \sum_{i=1}^{n} x_i + \log(1 - \theta) \sum_{i=1}^{n} (1 - x_i) \]

\[ N_1 \ - \text{number of heads seen} \quad N_2 \ - \text{number of tails seen} \]
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 T T H T H H H H T H H H H T H H
  
  - **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 T T H 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:**

$$(1 - \theta_{ML}) = \frac{N_2}{N} = \frac{N_2}{N_1 + N_2} = \frac{10}{25} = 0.4$$
Maximum a posteriori estimate

Maximum a posteriori estimate
– Selects the mode of the posterior distribution

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

Likelihood of data

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

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 \( \theta \)

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

\[ p(\theta \mid D, \xi) = \frac{P(D \mid \theta, \xi) \text{Beta}(\theta \mid \alpha_1, \alpha_2)}{P(D \mid \xi)} = \text{Beta}(\theta \mid \alpha_1 + N_1, \alpha_2 + N_2) \]
$p(\theta \mid \xi) = \text{Beta}(\theta \mid a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1 - \theta)^{b-1}$
Posterior distribution

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

Maximum a posteriori estimate
– Selects the mode of the posterior distribution

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

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

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

What is the MAP estimate?
MAP estimate example

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

What is the MAP estimate?

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

  – Heads: 15
  – Tails: 10

• Assume

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

  \[ \theta_{\text{MAP}} = \frac{19}{33} \]

  \[ p(\theta | \xi) = \text{Beta}(\theta | 5,20) \]

  \[ \theta_{\text{MAP}} = \frac{19}{48} \]