Bayesian Inference and Hypothesis Testing

Christos Dimitrakakis

October 25, 2024

Conditional Probability and the Theorem of Bayes

Simple Bayesian hypothesis testing

Bayesian Inference

Bayes theorem

- Recall the definition of Conditional probability:

$$P(A|B) = P(A \cap B)/P(B)$$

i.e. the probability of A happening if B happens.

Bayes theorem

- Recall the definition of Conditional probability:

$$P(A|B) = P(A \cap B)/P(B)$$

  i.e. the probability of A happening if B happens.
- It is also true that:

$$P(B|A) = P(A \cap B)/P(A)$$

Bayes theorem

- Recall the definition of Conditional probability:

$$P(A|B) = P(A \cap B)/P(B)$$

  i.e. the probability of A happening if B happens.
- It is also true that:

$$P(B|A) = P(A \cap B)/P(A)$$

- Combining the two equations, reverse the conditioning:

$$P(A|B) = P(B|A)P(A)/P(B)$$

Bayes theorem

- Recall the definition of Conditional probability:

$$P(A|B) = P(A \cap B)/P(B)$$

  i.e. the probability of A happening if B happens.
- It is also true that:

$$P(B|A) = P(A \cap B)/P(A)$$

- Combining the two equations, reverse the conditioning:

$$P(A|B) = P(B|A)P(A)/P(B)$$

- So we can reverse the order of conditioning, i.e. relate to the probability of A given B to that of B given A.

The cards problem

1. Print out a number of cards, with either [A|A], [A|B] or [B|B] on their sides.
2. If you have an A, what is the probability of an A on the other side?
3. Have the students perform the experiment with:
   - 3.1 Draw a random card.
   - 3.2 Count the number of people with A.
   - 3.3 What is the probability that somebody with an A on one side will have an A on the other?
   - 3.4 Half of the people should have an A?

The cards problem

1. Print out a number of cards, with either [A|A], [A|B] or [B|B] on their sides.
2. If you have an A, what is the probability of an A on the other side?
3. Have the students perform the experiment with:
    3.1 Draw a random card.
    3.2 Count the number of people with A.
    3.3 What is the probability that somebody with an A on one side will have an A on the other?
    3.4 Half of the people should have an A?

The prior and posterior probabilities

A	A	2/6	A observed	2/3
A	B	1/6	A observed	1/3
B	A	1/6		
B	B	2/6		

Conditional Probability and the Theorem of Bayes

Simple Bayesian hypothesis testing

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.
  - H_1: They are guilty.

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.
  - H_1: They are guilty.

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.
  - H_1: They are guilty.

What is your belief that they have committed the crime?

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.
  - H_1: They are guilty.

What is your belief that they have committed the crime?

Prior elicitation

- All those that think the accused is guilty, raise their hand.

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.
  - H_1: They are guilty.

What is your belief that they have committed the crime?

Prior elicitation

- All those that think the accused is guilty, raise their hand.
- Divide by the number of people in class

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.
  - H_1: They are guilty.

What is your belief that they have committed the crime?

Prior elicitation

- All those that think the accused is guilty, raise their hand.
- Divide by the number of people in class
- Let us call this $P(H_1)$.

The murder problem

- A murder occurred in a house over Christmas. There were n people inside, plus the victim. Person X, the victim's son, is accused of a murder.
- There are two possibilities:
  - H_0: They are innocent.
  - H_1: They are guilty.

What is your belief that they have committed the crime?

Prior elicitation

- All those that think the accused is guilty, raise their hand.
- Divide by the number of people in class
- Let us call this $P(H_1)$.
- This is a purely subjective measure!

DNA test

- Let us now do a DNA test on the suspect

DNA test

- Let us now do a DNA test on the suspect

DNA test properties

- D: Test is positive

DNA test

- Let us now do a DNA test on the suspect

DNA test properties

- D: Test is positive
- $P(D|H_0) = 10\%$: False positive rate

DNA test

- Let us now do a DNA test on the suspect

DNA test properties

- D: Test is positive
- $P(D|H_0) = 10\%$: False positive rate
- $P(D|H_1) = 100\%$: True positive rate

DNA test

- Let us now do a DNA test on the suspect

DNA test properties

- D: Test is positive
- $P(D|H_0) = 10\%$: False positive rate
- $P(D|H_1) = 100\%$: True positive rate

DNA test

- Let us now do a DNA test on the suspect

DNA test properties

- D: Test is positive
- $P(D|H_0) = 10\%$: False positive rate
- $P(D|H_1) = 100\%$: True positive rate

Run the test

- The result is either positive or negative ($\neg D$).

DNA test

- Let us now do a DNA test on the suspect

DNA test properties

- D: Test is positive
- $P(D|H_0) = 10\%$: False positive rate
- $P(D|H_1) = 100\%$: True positive rate

Run the test

- The result is either positive or negative ($\neg D$).
- What is your belief now that the suspect is guilty?

Everybody is a suspect

- Run a DNA test on everybody in the house.

Everybody is a suspect

- Run a DNA test on everybody in the house.
- What is different from before?

Everybody is a suspect

- Run a DNA test on everybody in the house.
- What is different from before?
- Who has a positive test?

Everybody is a suspect

- Run a DNA test on everybody in the house.
- What is different from before?
- Who has a positive test?
- What is your belief that the people with the positive test are guilty?

Explanation

- Prior: $P(H_i)$.

$$P(D) = P(D \cap H_0) + P(D \cap H_1) \tag{1}$$
$$= P(D|H_0)P(H_0) + P(D|H_1)P(H_1) \tag{2}$$

- Posterior: $P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D|H_0)P(H_0)+P(D|H_1)P(H_1)}$
- Assuming $P(D|H_1) = 1$, and setting $P(H_0) = q$, this gives

$$P(H_0|D) = \frac{0.1q}{0.1q + 1 - q} = \frac{q}{10 - 9q}$$

- The posterior can always be updated with more data!

Explanation

- Prior: $P(H_i)$.
- Likelihood $P(D|H_i)$.

$$
\begin{aligned}
P(D) &= P(D \cap H_0) + P(D \cap H_1) && (1) \\
&= P(D|H_0)P(H_0) + P(D|H_1)P(H_1) && (2)
\end{aligned}
$$

- Posterior: $P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D|H_0)P(H_0)+P(D|H_1)P(H_1)}$
- Assuming $P(D|H_1) = 1$, and setting $P(H_0) = q$, this gives

$$
P(H_0|D) = \frac{0.1q}{0.1q + 1 - q} = \frac{q}{10 - 9q}
$$

- The posterior can always be updated with more data!

Explanation

- Prior: $P(H_i)$.
- Likelihood $P(D|H_i)$.
- Posterior: $P(H_i|D) = P(D \cap H_i)/P(D) = P(D|H_i)P(H_i)/P(D)$

$$P(D) = P(D \cap H_0) + P(D \cap H_1) \tag{1}$$
$$= P(D|H_0)P(H_0) + P(D|H_1)P(H_1) \tag{2}$$

- Posterior: $P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D|H_0)P(H_0)+P(D|H_1)P(H_1)}$
- Assuming $P(D|H_1) = 1$, and setting $P(H_0) = q$, this gives

$$P(H_0|D) = \frac{0.1q}{0.1q + 1 - q} = \frac{q}{10 - 9q}$$

- The posterior can always be updated with more data!

Explanation

- Prior: $P(H_i)$.
- Likelihood $P(D|H_i)$.
- Posterior: $P(H_i|D) = P(D \cap H_i)/P(D) = P(D|H_i)P(H_i)/P(D)$
- Marginal probability:

$$\begin{aligned} P(D) &= P(D \cap H_0) + P(D \cap H_1) && (1) \\ &= P(D|H_0)P(H_0) + P(D|H_1)P(H_1) && (2) \end{aligned}$$

- Posterior: $P(H_0|D) = \frac{P(D|H_0)P(H_0)}{P(D|H_0)P(H_0)+P(D|H_1)P(H_1)}$
- Assuming $P(D|H_1) = 1$, and setting $P(H_0) = q$, this gives

$$P(H_0|D) = \frac{0.1q}{0.1q + 1 - q} = \frac{q}{10 - 9q}$$

- The posterior can always be updated with more data!

Python example

```
def get_posterior(prior, data, likelihood):
    marginal = prior * likelihood[data][0] + (1 - prior) * likeliho
    posterior = prior * likelihood[data][0] / marginal
    return posterior

import numpy as np
prior = 0.9 # Pr(H1)
likelihood = np.zeros([2, 2])
likelihood[0][0] = 0.9 # Pr(F|H0)
likelihood[1][0] = 0.1 # Pr(T|H0)
likelihood[0][1] = 0 # Pr(F|H1)
likelihood[1][1] = 1 # Pr(T|H1)
data = 1
return get_posterior(prior, data, likelihood)
```

Types of hypothesis testing problems

Simple Hypothesis Test

Example: DNA evidence, Covid tests

- Two hypothesese H_0, H_1
- $P(D|H_i)$ is defined for all i

Multiple Hypotheses Test

Example: Model selection

- H_i: One of many mutually exclusive models
- $P(D|H_i)$ is defined for all i

Null Hypothesis Test

Example: Are men's and women's heights the same?

- H_0: The 'null' hypothesis
- $P(D|H_0)$ is defined
- The alternative is undefined

Pitfalls

Problem definition

- Defining the models $P(D|H_i)$ incorrectly.

Pitfalls

Problem definition

- Defining the models $P(D|H_i)$ incorrectly.
- Using an "unreasonable" prior $P(H_i)$

Pitfalls

Problem definition

- Defining the models $P(D|H_i)$ incorrectly.
- Using an "unreasonable" prior $P(H_i)$

Pitfalls

Problem definition

- Defining the models $P(D|H_i)$ incorrectly.
- Using an "unreasonable" prior $P(H_i)$

The garden of many paths

- Having a huge hypothesis space

Pitfalls

Problem definition

- Defining the models $P(D|H_i)$ incorrectly.
- Using an "unreasonable" prior $P(H_i)$

The garden of many paths

- Having a huge hypothesis space
- Selecting the relevant hypothesis after seeing the data

Probabilistic models

- Model family $\{P_\theta | \theta \in \Theta\}$

Probabilistic models

- Model family $\{P_\theta | \theta \in \Theta\}$
- Data $x \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.

Probabilistic models

- Model family $\{P_\theta | \theta \in \Theta\}$
- Data $x \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.
- How can we estimate the correct θ?

Probabilistic models

- Model family $\{P_\theta | \theta \in \Theta\}$
- Data $x \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.
- How can we estimate the correct θ?
- How can we predict a new data point?

Probabilistic models

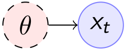
- Model family $\{P_\theta | \theta \in \Theta\}$
- Data $x \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.
- How can we estimate the correct θ?
- How can we predict a new data point?

Probabilistic models

- Model family $\{P_\theta | \theta \in \Theta\}$
- Data $x \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.
- How can we estimate the correct θ?
- How can we predict a new data point?

Example (Bernoulli model)

- $x \in \{0, 1\}$, $\theta \in [0, 1]$

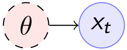


Probabilistic models

- Model family $\{P_\theta | \theta \in \Theta\}$
- Data $x \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.
- How can we estimate the correct θ?
- How can we predict a new data point?

Example (Bernoulli model)

- $x \in \{0, 1\}$, $\theta \in [0, 1]$
- $x \mid \theta \sim \mathrm{Bernoulli}(\theta)$



Probabilistic models

- Model family $\{P_\theta | \theta \in \Theta\}$
- Data $x \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.
- How can we estimate the correct θ?
- How can we predict a new data point?

Example (Bernoulli model)

- $x \in \{0, 1\}$, $\theta \in [0, 1]$
- $x \mid \theta \sim \mathrm{Bernoulli}(\theta)$
- $\mathrm{P}_\theta(\mathrm{x} = 1) = \theta$

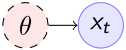
$$\theta \longrightarrow x_t$$

Probabilistic models

- Model family $\{P_\theta | \theta \in \Theta\}$
- Data $x \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.
- How can we estimate the correct θ?
- How can we predict a new data point?

Example (Bernoulli model)

- $x \in \{0, 1\}$, $\theta \in [0, 1]$
- $x \mid \theta \sim \mathrm{Bernoulli}(\theta)$
- $\mathrm{P}_\theta(\mathrm{x} = 1) = \theta$
- $\mathrm{P}_\theta(\mathrm{x} = 0) = 1 - \theta$.

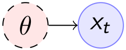


Probabilistic models

- Model family $\{P_\theta | \theta \in \Theta\}$
- Data $x \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.
- How can we estimate the correct θ?
- How can we predict a new data point?

Example (Bernoulli model)

- $x \in \{0, 1\}$, $\theta \in [0, 1]$
- $x \mid \theta \sim \mathrm{Bernoulli}(\theta)$
- $P_\theta(x = 1) = \theta$
- $P_\theta(x = 0) = 1 - \theta$.

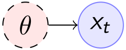


Probabilistic models

- Model family $\{P_\theta | \theta \in \Theta\}$
- Data $x \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.
- How can we estimate the correct θ?
- How can we predict a new data point?

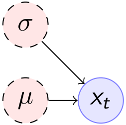
Example (Bernoulli model)

- $x \in \{0, 1\}$, $\theta \in [0, 1]$
- $x \mid \theta \sim \mathrm{Bernoulli}(\theta)$
- $\mathrm{P}_\theta(\mathrm{x} = 1) = \theta$
- $\mathrm{P}_\theta(\mathrm{x} = 0) = 1 - \theta$.



Example (Gaussian model)

- $x \in \mathbb{R}$, $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+$

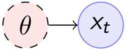


Probabilistic models

- Model family $\{P_\theta | \theta \in \Theta\}$
- Data $x \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.
- How can we estimate the correct θ?
- How can we predict a new data point?

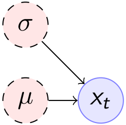
Example (Bernoulli model)

- $x \in \{0, 1\}$, $\theta \in [0, 1]$
- $x \mid \theta \sim \mathrm{Bernoulli}(\theta)$
- $\mathrm{P}_\theta(\mathrm{x} = 1) = \theta$
- $\mathrm{P}_\theta(\mathrm{x} = 0) = 1 - \theta$.



Example (Gaussian model)

- $x \in \mathbb{R}$, $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+$
- $x \mid \mu, \sigma \sim \mathrm{Normal}(\mu, \sigma)$

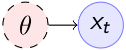


Probabilistic models

- Model family $\{P_\theta | \theta \in \Theta\}$
- Data $x \sim P_{\theta^*}$ for some $\theta^* \in \Theta$.
- How can we estimate the correct θ?
- How can we predict a new data point?

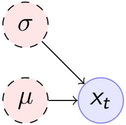
Example (Bernoulli model)

- $x \in \{0, 1\}$, $\theta \in [0, 1]$
- $x \mid \theta \sim \mathrm{Bernoulli}(\theta)$
- $\mathrm{P}_\theta(\mathrm{x} = 1) = \theta$
- $\mathrm{P}_\theta(\mathrm{x} = 0) = 1 - \theta$.



Example (Gaussian model)

- $x \in \mathbb{R}$, $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+$
- $x \mid \mu, \sigma \sim \mathrm{Normal}(\mu, \sigma)$
- $p_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$



Maximum likelihood (ML) inference

- Family $\{P_\theta | \theta \in \Theta\}$
- Data x with likelihood $P_\theta(x)$ for each parameter value θ.
- $\theta_{\mathrm{ML}}(x) = \arg\max_\theta P_\theta(x)$

Example (Bernoulli model)

- $x_t \in \{0, 1\}$, for $t \in [T]$, $\theta \in [0, 1]$

Maximum likelihood (ML) inference

- Family $\{P_\theta | \theta \in \Theta\}$
- Data x with likelihood $P_\theta(x)$ for each parameter value θ.
- $\theta_{\mathrm{ML}}(x) = \arg\max_\theta P_\theta(x)$

Example (Bernoulli model)

- $x_t \in \{0, 1\}$, for $t \in [T]$, $\theta \in [0, 1]$
- $x_t \mid \theta \sim \mathrm{Bernoulli}(\theta)$

Maximum likelihood (ML) inference

- Family $\{P_\theta | \theta \in \Theta\}$
- Data x with likelihood $P_\theta(x)$ for each parameter value θ.
- $\theta_{\mathrm{ML}}(x) = \arg\max_\theta P_\theta(x)$

Example (Bernoulli model)

- $x_t \in \{0, 1\}$, for $t \in [T]$, $\theta \in [0, 1]$
- $x_t \mid \theta \sim \mathrm{Bernoulli}(\theta)$
- $P_\theta(x_1, \ldots, x_T) = \prod_{t=1}^{T} P_\theta(x_t)$

Maximum likelihood (ML) inference

- Family $\{P_\theta | \theta \in \Theta\}$
- Data x with likelihood $P_\theta(x)$ for each parameter value θ.
- $\theta_{\mathrm{ML}}(x) = \arg\max_\theta P_\theta(x)$

Example (Bernoulli model)

- $x_t \in \{0, 1\}$, for $t \in [T]$, $\theta \in [0, 1]$
- $x_t \mid \theta \sim \mathrm{Bernoulli}(\theta)$
- $P_\theta(x_1, \ldots, x_T) = \prod_{t=1}^T P_\theta(x_t)$
- What maximises the likelihood?

Maximum likelihood (ML) inference

- Family $\{P_\theta | \theta \in \Theta\}$
- Data x with likelihood $P_\theta(x)$ for each parameter value θ.
- $\theta_{\mathrm{ML}}(x) = \arg\max_\theta P_\theta(x)$

Example (Bernoulli model)

- $x_t \in \{0, 1\}$, for $t \in [T]$, $\theta \in [0, 1]$
- $x_t \mid \theta \sim \mathrm{Bernoulli}(\theta)$
- $P_\theta(x_1, \ldots, x_T) = \prod_{t=1}^T P_\theta(x_t)$
- What maximises the likelihood?
- Define $s_T = \sum_{t=1}^T x_t$.

Maximum likelihood (ML) inference

- Family $\{P_\theta | \theta \in \Theta\}$
- Data x with likelihood $P_\theta(x)$ for each parameter value θ.
- $\theta_{\mathrm{ML}}(x) = \arg\max_\theta P_\theta(x)$

Example (Bernoulli model)

- $x_t \in \{0, 1\}$, for $t \in [T]$, $\theta \in [0, 1]$
- $x_t \mid \theta \sim \mathrm{Bernoulli}(\theta)$
- $P_\theta(x_1, \ldots, x_T) = \prod_{t=1}^{T} P_\theta(x_t)$
- What maximises the likelihood?
- Define $s_T = \sum_{t=1}^{T} x_t$.
- Show that $\theta_{\mathrm{ML}}(x) = s_T / T$.

Maximum likelihood (ML) inference

- Family $\{P_\theta | \theta \in \Theta\}$
- Data x with likelihood $P_\theta(x)$ for each parameter value θ.
- $\theta_{\mathrm{ML}}(x) = \arg\max_\theta P_\theta(x)$

Example (Bernoulli model)

- $x_t \in \{0, 1\}$, for $t \in [T]$, $\theta \in [0, 1]$
- $x_t \mid \theta \sim \mathrm{Bernoulli}(\theta)$
- $P_\theta(x_1, \ldots, x_T) = \prod_{t=1}^T P_\theta(x_t)$
- What maximises the likelihood?
- Define $s_T = \sum_{t=1}^T x_t$.
- Show that $\theta_{\mathrm{ML}}(x) = s_T / T$.
- What is the problem with this estimate?

Maximum a posteriori (MAP) inference

- Family $\{P_\theta | \theta \in \Theta\}$
- Data x with likelihood $P_\theta(x)$ for each parameter value θ.
- Prior $\beta(\theta)$.
- $\theta_{\mathrm{MAP}}(x) = \arg\max_\theta P_\theta(x)\beta(\theta)$
- Experiment with the prior for the Bernoulli model.

Bayesian Inference

- Model family $\{P_\theta | \theta \in \Theta\}$

Bayesian Inference

- Model family $\{P_\theta | \theta \in \Theta\}$
- Each model θ assigns probabilities $P_\theta(x)$ to possible $x \in X$.

Bayesian Inference

- Model family $\{P_\theta | \theta \in \Theta\}$
- Each model θ assigns probabilities $P_\theta(x)$ to possible $x \in X$.
- We also have a (subjective) prior distribution β over the parameters.

Bayesian Inference

- Model family $\{P_\theta | \theta \in \Theta\}$
- Each model θ assigns probabilities $P_\theta(x)$ to possible $x \in X$.
- We also have a (subjective) prior distribution β over the parameters.
- Given x, we calculate the posterior distribution

Bayesian Inference

- Model family $\{P_\theta | \theta \in \Theta\}$
- Each model θ assigns probabilities $P_\theta(x)$ to possible $x \in X$.
- We also have a (subjective) prior distribution β over the parameters.
- Given x, we calculate the posterior distribution

Bayesian Inference

- Model family $\{P_\theta | \theta \in \Theta\}$
- Each model θ assigns probabilities $P_\theta(x)$ to possible $x \in X$.
- We also have a (subjective) prior distribution β over the parameters.
- Given x, we calculate the posterior distribution

$$\beta(\theta|x) = \frac{P_\theta(x)\beta(\theta)}{\sum_{\theta' \in \Theta} P_{\theta'}(x)\beta(\theta')}, \qquad \text{(finite } \Theta, \beta \text{ is a probability)}$$

$$\beta(\theta|x) = \frac{P_\theta(x)\beta(\theta)}{\int_\Theta P_{\theta'}(x)\beta(\theta')d\theta'}, \qquad \text{(continuous } \Theta, \beta \text{ is a density)}$$

$$\beta(B|x) = \frac{\int_B P_{\theta'}(x)d\beta(\theta)}{\int_\Theta P_{\theta'}(x)d\beta(\theta)}, \qquad B \subset \Theta \qquad \text{(arbitrary } \Theta, \beta \text{ is a measure)}$$

Bayesian Inference

- Model family $\{P_\theta | \theta \in \Theta\}$
- Each model θ assigns probabilities $P_\theta(x)$ to possible $x \in X$.
- We also have a (subjective) prior distribution β over the parameters.
- Given x, we calculate the posterior distribution

$$\beta(\theta|x) = \frac{P_\theta(x)\beta(\theta)}{\sum_{\theta' \in \Theta} P_{\theta'}(x)\beta(\theta')}, \qquad \text{(finite } \Theta, \beta \text{ is a probability)}$$

$$\beta(\theta|x) = \frac{P_\theta(x)\beta(\theta)}{\int_\Theta P_{\theta'}(x)\beta(\theta')d\theta'}, \qquad \text{(continuous } \Theta, \beta \text{ is a density)}$$

$$\beta(B|x) = \frac{\int_B P_{\theta'}(x)d\beta(\theta)}{\int_\Theta P_{\theta'}(x)d\beta(\theta)}, \qquad B \subset \Theta \qquad \text{(arbitrary } \Theta, \beta \text{ is a measure)}$$

Alternative notation for different probability spaces

- The prior $\beta(\theta) = \mathbb{P}(\theta)$ and posterior $\beta(\theta \mid x) = \mathbb{P}(\theta \mid x)$ belief.
- The likelihood $P_\theta(x) = \mathbb{P}(x \mid \theta)$
- The marginal $\mathbb{P}_\beta(x) = \sum_\theta P_\theta(x)\beta(\theta)$.

Probabilistic machine learning

- Model family $\{P_\theta | \theta \in \Theta\}$

Probabilistic machine learning

- Model family $\{P_\theta | \theta \in \Theta\}$
- Prior β on Θ

Probabilistic machine learning

- Model family $\{P_\theta | \theta \in \Theta\}$
- Prior β on Θ
- Observations $x = x_1, \ldots, x_t$.

Probabilistic machine learning

- Model family $\{P_\theta | \theta \in \Theta\}$
- Prior β on Θ
- Observations $x = x_1, \ldots, x_t$.

Probabilistic machine learning

- Model family $\{P_\theta | \theta \in \Theta\}$
- Prior β on Θ
- Observations $x = x_1, \ldots, x_t$.

Maximum likelihood approach

- Model selection: $\theta^*_{ML}(x) = \arg\max_\theta P_\theta(x)$.
- Model prediction: $P_{\theta^*_{ML}(x)}(x_{t+1})$

Probabilistic machine learning

- Model family $\{P_\theta | \theta \in \Theta\}$
- Prior β on Θ
- Observations $x = x_1, \ldots, x_t$.

Maximum likelihood approach

- Model selection: $\theta^*_{ML}(x) = \arg\max_\theta P_\theta(x)$.
- Model prediction: $P_{\theta^*_{ML}(x)}(x_{t+1})$

Maximum a posteriori approach

- Model selection: $\theta^*_{MAP}(x) = \arg\max_\theta P_\theta(x)\beta(\theta)$.
- Model prediction: $P_{\theta^*_{MAP}(x)}(x_{t+1})$

Probabilistic machine learning

- Model family $\{P_\theta | \theta \in \Theta\}$
- Prior β on Θ
- Observations $x = x_1, \ldots, x_t$.

Maximum likelihood approach

- Model selection: $\theta^*_{ML}(x) = \arg\max_\theta P_\theta(x)$.
- Model prediction: $P_{\theta^*_{ML}(x)}(x_{t+1})$

Maximum a posteriori approach

- Model selection: $\theta^*_{MAP}(x) = \arg\max_\theta P_\theta(x)\beta(\theta)$.
- Model prediction: $P_{\theta^*_{MAP}(x)}(x_{t+1})$

Bayesian approach

- Posterior calculation: $\beta(\theta|x) = P_\theta(x)\beta(\theta)/\,\mathbb{P}_\beta(x)$
- Model prediction: $\mathbb{P}_\beta(x_{t+1}|x) = \sum_\theta P_\theta(x_{t+1})\beta(\theta|x)$

Differences between approaches

Maximum likelihood approach

- Ignores model complexity
- Is an optimisation problem

Maximum a posteriori approach

- Regularises model selection using the prior
- Can be seen as solving the optimisation problem

$$\max_{\theta} \ln P_{\theta}(x) + \ln \beta(\theta),$$

where the prior term $\ln \beta(\theta)$ acts as a regulariser.

Bayesian approach

- Does not select a single model
- Averages over all models according to their fit and the prior
- Does not result in an optimisation problem.

The n-meteorologists problem

- Consider n meteorological stations $\{\mu\}$ predicting rainfall.
- $x_t \in \{0, 1\}$ with $x_t = 1$ if it rains on day t.
- We have a prior distribution $\beta(\mu)$ for each station.
- At time t, station μ makes as a prediction $P_\mu(x_{t+1}|x_1, \dots, x_t)$
- We observe x_{t+1} and calculate the posterior $\beta(\mu|x_1, \dots, x_t, x_{t+1})$.

The marginal distribution

To take into account all stations, we can marginalise:

$$\mathbb{P}_\beta(x_{t+1} \mid x_1, \dots x_t) = \sum_\mu P_\mu(x_{t+1}|x_t)\beta(\mu)$$

The posterior

- Show that

$$\beta(\mu \mid x_1, \dots, x_{t+1}) = \frac{P_\mu(x_t \mid x_1, \dots, x_t)\beta(\mu|x_1, \dots, x_t)}{\sum_{\mu'} P_{\mu'}(x_t \mid x_1, \dots, x_t)\beta(\mu'|x_1, \dots, x_t)}$$

- How would you implement an ML or a MAP solution to this problem?

Sufficient statistics

A statistic f

This is any function $f : X \to S$ where

- X is the data space
- S is an arbitrary space

Example statistics for $X = \mathbb{R}^*$ (the set of all real-valued sequences)

- The sample mean of a sequence $1/T \sum_{t=1}^{T} x_t$
- The total number of samples T

Sufficient statistic

f is sufficient for a family $\{P_\theta : \theta \in \Theta\}$ when

$$f(x) = f(x') \Rightarrow P_\theta(x) = P_\theta(x') \forall \theta \in \Theta.$$

If there exists a finite-dimensional sufficient statistic, Bayesian and ML learning can be done in closed form within the family.

Conjugate priors

Consider a parametrised family of priors $\mathcal{B}$ on Θ and a distribution family $\{P_\theta\}$
The pair is conjugate if, for any prior $\beta \in \mathcal{B}$, and any observation x, there exists $\beta' \in \mathcal{B}$ such that $\beta'(\theta) = \beta(\theta|x)$

Standard Parametric conjugate families

Prior	Likelihood	Parameters θ	Observations x
Beta	Bernoulli	$[0, 1]$	$\{0, 1\}^T$
Multinomial	Dirichlet	$\mathbb{\Delta}^n$	$\{1, \ldots, n\}^T$
Gamma	Normal	$\mathbb{R}, \mathbb{R}$	$\mathbb{R}^T$
Wishart	Normal	$\mathbb{R}^n, \mathbb{R}^{n \times n}$	$\mathbb{R}^{n \times T}$

The Simplex $\mathbb{\Delta}^n = \{\boldsymbol{\theta} \in [0, 1]^n : \|\boldsymbol{\theta}\|_1\}$ is the set of all n-dimensional probability vectors.

Extensions

- Discrete Bayesian Networks.
- Linear-Gaussian Models (i.e. Bayesian linear regression)
- Gaussian Processes.

Beta-Bernoulli

$$\boldsymbol{\alpha} \longrightarrow \theta \longrightarrow x_t$$

Definition of the Bernoulli distribution

If $x_t \mid \theta \sim \mathrm{Bernoulli}(\theta)$. $\theta \in [0,1]$, $x_t \in \{0,1\}$ and:

$$P_\theta(x_t = 1) = \theta$$

Definition of the Beta density

If $\theta \sim \mathrm{Beta}(\alpha_1, \alpha_0)$, $\alpha_0, \alpha_1 > 0$ and

$$p(\theta | \alpha_1, \alpha_0) \propto \theta^{\alpha_1 - 1}(1 - \theta)^{\alpha_0 - 1}$$

Beta-Bernoulli conjugate pair

- $\theta \sim \mathrm{Beta}(\alpha_1, \alpha_0)$.
- $x_t \mid \theta \sim \mathrm{Bernoulli}(\theta)$.

Then, for any $x = x_1, \ldots, x_T$, the posterior distribution is

- $\theta \mid x \sim \mathrm{Beta}(\alpha_1 + \sum_t x_t, \alpha_0 + T - \sum_t x_t)$.

Dirichlet-Multinomial

$$\boldsymbol{\alpha} \longrightarrow \boldsymbol{\theta} \longrightarrow x_t$$

Definition of the Multinomial distribution

If $x_t \mid \boldsymbol{\theta} \sim \mathrm{Mult}(\boldsymbol{\theta})$, with $\theta \in \mathbb{A}^n$ and $x_t \in \{1, \ldots, n\}$ and:

$$P_{\boldsymbol{\theta}}(x_t = i) = \theta_i$$

Definition of the Dirichlet density

If $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, with $\boldsymbol{\alpha} \in \mathbb{R}^n_+$ then

$$p(\theta | \boldsymbol{\alpha}) \propto \prod_i \theta_i^{\alpha_i - 1}$$

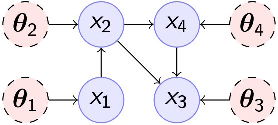
Dirichlet-Multinomial conjugate pair

- $\theta \sim \mathrm{Dir}(\boldsymbol{\alpha})$.
- $x_t \mid \theta \sim \mathrm{Bernoulli}(\boldsymbol{\theta})$.

Then, for any $x = x_1, \ldots, x_T$, the posterior distribution is

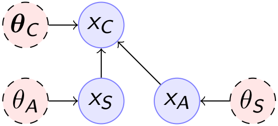
- $\theta \mid x \sim \mathrm{Dir}(\boldsymbol{\alpha} + \boldsymbol{s_T})$, where $s_{T,i} = \sum_{t=1}^T \mathbb{I}\{x_t = i\}$.

Discrete Bayesian Networks



- A directed acyclic graph (DAG) defined on variables $x_1, \ldots, x_n$ with each x_n taking a finite number of values,
- Let S_i be the indices corresponding to parent variables of x_i.
- $x_i \mid \boldsymbol{\theta}_i, x_{S_i} = k \sim \mathrm{Mult}(\boldsymbol{\theta}_{i,k})$.

Example: Lung cancer, smoking and asbestos

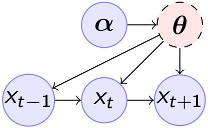


$$P_{\theta_A}(x_A = 1) = \theta_A \tag{3}$$

$$P_{\theta_S}(x_S = 1) = \theta_S \tag{4}$$

$$P_{\boldsymbol{\theta}_C}(x_C = 1 \mid X_A = j, X_S = k) = \theta_{C,j,k} \tag{5}$$

Markov model



A Markov model obeys

$$\mathbb{P}_{\boldsymbol{\theta}}(x_{k+1}|x_k, \ldots, x_1) = \mathbb{P}_{\boldsymbol{\theta}}(x_{k+1}|x_k)$$

i.e. the graphical model is a chain. We are usually interested in homogeneous models, where

$$\mathbb{P}_{\boldsymbol{\theta}}(x_{k+1} = i \mid x_k = j) = \theta_{i,j} \qquad \forall k$$

Inference for finite Markov models

- If $x_t \in [n]$ then $x_{t+1} \mid \boldsymbol{\theta}, x_t = i \sim \mathrm{Mult}(\boldsymbol{\theta}_i)$, $\boldsymbol{\theta}_i \in \mathbb{A}^n$
- Prior $\boldsymbol{\theta}_i \mid \boldsymbol{\alpha} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ for all $i \in [n]$.
- Posterior $\boldsymbol{\theta}_i \mid x_1, \ldots, x_t, \boldsymbol{\alpha} \sim \mathrm{Dir}(\boldsymbol{\alpha}^{(t)})$ with

$$\alpha^t_{i,j} = \alpha_{i,j} + \sum_{k=1}^{t} \mathbb{I}\left\{x_k = i \wedge x_{k+1} = j\right\}, \qquad \boldsymbol{\alpha}^0 = \boldsymbol{\alpha}.$$