Machine Learning 2007: Lecture 11

Instructor: Tim van Erven (Tim.van.Erven@cwi.nl)
Website: www.cwi.nl/˜erven/teaching/0708/ml/

November 28, 2007

Overview

- **Organisational Matters**
- Models
- Maximum Likelihood Parameter Estimation
- Probability Theory
- Bayesian Learning
  - The Bayesian Distribution
  - From Prior to Posterior
  - MAP Parameter Estimation
  - Bayesian Predictions
  - Discussion
  - Advanced Issues

Guest lecture:

- Next week, Peter Grünwald will give a special guest lecture about minimum description length (MDL) learning.

This Lecture versus Mitchell:

- Chapter 6 up to section 6.5.0 about Bayesian learning.
- I present things in a better order.
- Mitchell also covers the connection between MAP parameter estimation and least squares linear regression: It is good for you to study this, but I will not ask an exam question about it.

Overview

- Organisational Matters
- **Models**
- Maximum Likelihood Parameter Estimation
- Probability Theory
- Bayesian Learning
  - The Bayesian Distribution
  - From Prior to Posterior
  - MAP Parameter Estimation
  - Bayesian Predictions
  - Discussion
  - Advanced Issues

Prediction Example without Noise

Training data:

$D =$

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
0	1	0	1	0	1	0	1

Hypothesis Space:

$$\mathcal{H} = \{h_1, h_2, h_3\}$$

$$h_1\colon \quad y_n = 0$$

$$h_2\colon \quad y_n = \begin{cases} 0 & \text{if } n \text{ is odd} \\ 1 & \text{if } n \text{ is even} \end{cases}$$

$$h_3\colon \quad y_n = 1$$

Prediction Example without Noise

Training data:

$D =$

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
0	1	0	1	0	1	0	1

Hypothesis Space:

$$\mathcal{H} = \{h_1, h_2, h_3\}$$

$$h_1: \quad y_n = 0$$

$$h_2: \quad y_n = \begin{cases} 0 & \text{if } n \text{ is odd} \\ 1 & \text{if } n \text{ is even} \end{cases}$$

$$h_3: \quad y_n = 1$$

By simple list-then-eliminate:

- Only h_2 is consistent with the training data.
- Therefore we predict, in accordance with h_2, that $y_9 = 0$.

Turning Hypotheses into Distributions

Models:

- We may view each hypothesis as probability distribution that gives probability 1 to a certain outcome.
- A hypothesis space that contains such probabilistic hypotheses is called a (statistical) **model**.

The previous hypotheses as distributions:

$$\mathcal{M} = \{P_1, P_2, P_3\}$$

$$P_1: \quad P_1(y_n = 0) = 1$$

$$P_2: \quad P_2(y_n = 0) = \begin{cases} 1 & \text{if } n \text{ is odd} \\ 0 & \text{if } n \text{ is even} \end{cases}$$

$$P_3: \quad P_3(y_n = 1) = 1$$

Turning Hypotheses into Distributions

Models:

- We may view each hypothesis as probability distribution that gives probability 1 to a certain outcome.
- A hypothesis space that contains such probabilistic hypotheses is called a (statistical) **model**.

The previous hypotheses as distributions:

$$\mathcal{M} = \{P_1, P_2, P_3\}$$

$$P_1\colon \quad P_1(y_n = 0) = 1$$

$$P_2\colon \quad P_2(y_n = 0) = \begin{cases} 1 & \text{if } n \text{ is odd} \\ 0 & \text{if } n \text{ is even} \end{cases}$$

$$P_3\colon \quad P_3(y_n = 1) = 1$$

List-then-eliminate still works:

- A probabilistic hypothesis is consistent with the data if it gives positive probability to the data.

Prediction Example with Noise

Noise:

- Using probabilistic hypotheses is natural when there is noise in the data.
- Suppose we observe a measurement error with some (small) probability ϵ.

This is easy to incorporate:

$$\mathcal{M} = \{P_1, P_2, P_3\}$$

$$P_1\colon \quad P_1(y_n = 0) = 1 - \epsilon$$

$$P_2\colon \quad P_2(y_n = 0) = \begin{cases} 1 - \epsilon & \text{if } n \text{ is odd} \\ \epsilon & \text{if } n \text{ is even} \end{cases}$$

$$P_3\colon \quad P_3(y_n = 1) = 1 - \epsilon$$

Prediction Example with Noise

Noise:

- Using probabilistic hypotheses is natural when there is noise in the data.
- Suppose we observe a measurement error with some (small) probability ϵ.

This is easy to incorporate:

$$\mathcal{M} = \{P_1, P_2, P_3\}$$

$$P_1: \quad P_1(y_n = 0) = 1 - \epsilon$$

$$P_2: \quad P_2(y_n = 0) = \begin{cases} 1 - \epsilon & \text{if } n \text{ is odd} \\ \epsilon & \text{if } n \text{ is even} \end{cases}$$

$$P_3: \quad P_3(y_n = 1) = 1 - \epsilon$$

List-then-eliminate does not work any more:

- For example, $P_1(D = 0, 1, 0, 1, 0, 1, 0, 1) = \epsilon^4 (1 - \epsilon)^4$.
- Typically many or all probabilistic hypotheses in our model will be consistent with the data.

Overview

- Organisational Matters
- Models
- **Maximum Likelihood Parameter Estimation**
- Probability Theory
- Bayesian Learning
  - The Bayesian Distribution
  - From Prior to Posterior
  - MAP Parameter Estimation
  - Bayesian Predictions
  - Discussion
  - Advanced Issues

Parameters

Parameters index the elements of a hypothesis space:

$$\mathcal{H} = \{h_1, h_2, h_3\} \qquad \iff \qquad \mathcal{H} = \{h_\theta \mid \theta \in \{1, 2, 3\}\}$$

Parameters

Parameters index the elements of a hypothesis space:

$$\mathcal{H} = \{h_1, h_2, h_3\} \qquad \Longleftrightarrow \qquad \mathcal{H} = \{h_\theta \mid \theta \in \{1, 2, 3\}\}$$

Usually in a convenient way:

Hypotheses are often expressed in terms of the parameters. In linear regression for example:

$$\mathcal{H} = \{h_{\mathbf{w}} \mid \mathbf{w} \in \mathbb{R}^2\} \quad \text{where } h_{\mathbf{w}} : y = w_0 + w_1 x.$$

Parameters

Parameters index the elements of a hypothesis space:

$$\mathcal{H} = \{h_1, h_2, h_3\} \qquad \Longleftrightarrow \qquad \mathcal{H} = \{h_\theta \mid \theta \in \{1, 2, 3\}\}$$

Usually in a convenient way:

Hypotheses are often expressed in terms of the parameters. In linear regression for example:

$$\mathcal{H} = \{h_\mathbf{w} \mid \mathbf{w} \in \mathbb{R}^2\} \quad \text{where } h_\mathbf{w} : y = w_0 + w_1 x.$$

Example where the hypothesis space is a model:

For example in prediction of binary outcomes:

$$\mathcal{M} = \left\{P_\theta \mid \theta \in \left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}\right\} \quad \text{where } P_\theta(y_n = 1) = \theta.$$

Maximum Likelihood Parameter Estimation

Training data and model:

$D =$

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
0	1	1	1	0	1	1	1

$$\mathcal{M} = \left\{ P_\theta \mid \theta \in \left\{ \frac{1}{4}, \frac{1}{2}, \frac{3}{4} \right\} \right\}$$ where $P_\theta(y_n = 1) = \theta$.

Likelihood:

θ	$1/4$	$1/2$	$3/4$
$P_\theta(D)$	$(1/4)^6(3/4)^2$ $= 9/65536$	$(1/2)^8$ $= 256/65536$	$(3/4)^6(1/4)^2$ $= 729/65536$

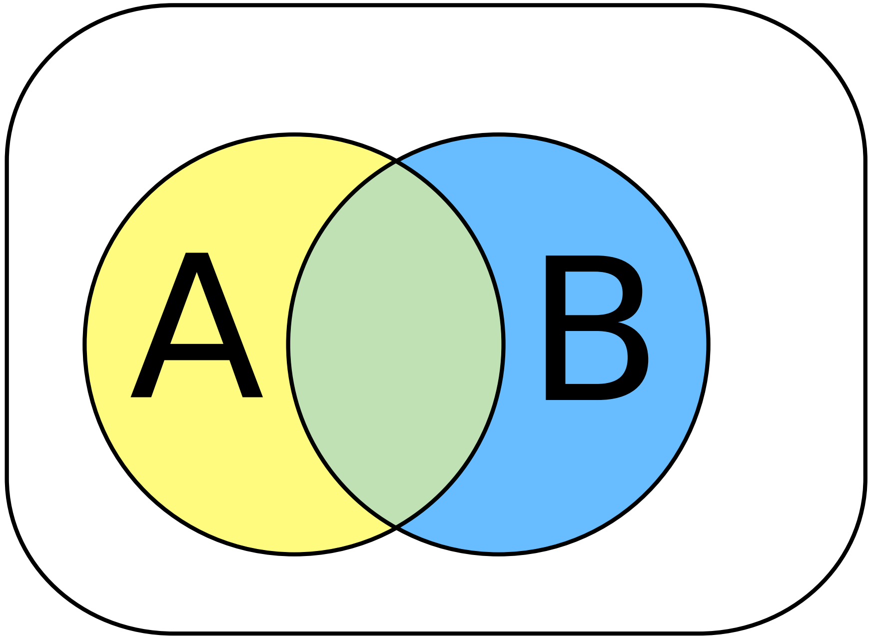
Maximum Likelihood Parameter Estimation:

$$\hat{\theta} = \arg\max_\theta P_\theta(D) = 3/4$$

Overview

- Organisational Matters
- Models
- Maximum Likelihood Parameter Estimation
- **Probability Theory**
- Bayesian Learning
  - The Bayesian Distribution
  - From Prior to Posterior
  - MAP Parameter Estimation
  - Bayesian Predictions
  - Discussion
  - Advanced Issues

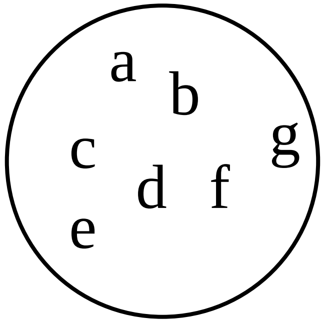
Relating Unions and Intersections



For any two events A and B:

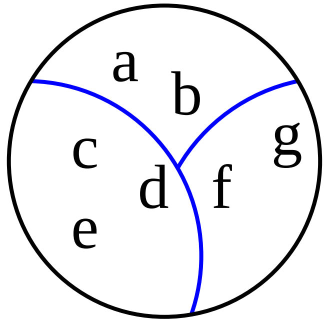
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The Law of Total Probability



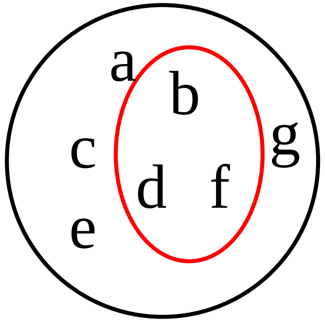
- Suppose $\Omega = \{a, b, c, d, e, f, g\}$.

The Law of Total Probability



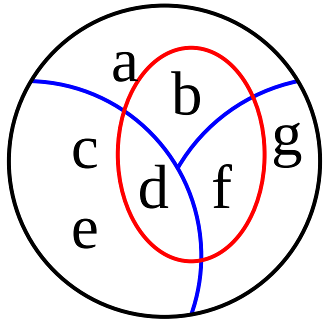
- Suppose $\Omega = \{a, b, c, d, e, f, g\}$.
- A **partition** of Ω cuts it into parts:
  - Let the parts be $A_1 = \{a, b\}$, $A_2 = \{c, d, e\}$ and $A_3 = \{f, g\}$
  - The parts do not overlap, and together cover Ω.

The Law of Total Probability



- Suppose $\Omega = \{a, b, c, d, e, f, g\}$.
- A **partition** of Ω cuts it into parts:
  - Let the parts be $A_1 = \{a, b\}$, $A_2 = \{c, d, e\}$ and $A_3 = \{f, g\}$
  - The parts do not overlap, and together cover Ω.
- $B = \{b, d, f\}$

The Law of Total Probability



- Suppose $\Omega = \{a, b, c, d, e, f, g\}$.
- A **partition** of Ω cuts it into parts:
  - Let the parts be $A_1 = \{a, b\}$, $A_2 = \{c, d, e\}$ and $A_3 = \{f, g\}$
  - The parts do not overlap, and together cover Ω.
- $B = \{b, d, f\}$

Law of Total Probability:

$$P(B) = \sum_{i=1}^{3} P(B \cap A_i) = \sum_{i=1}^{3} P(B \mid A_i) P(A_i)$$

Marginal Probability

- Suppose we throw a **blue** and a **red** die.
- Let X and Y be random variables, where
  X: outcome blue die; Y: outcome red die
- If we only know $P(X, Y)$, how do we compute $P(X)$?

Marginal Probability

- Suppose we throw a **blue** and a **red** die.
- Let X and Y be random variables, where
  X: outcome blue die; Y: outcome red die
- If we only know $P(X, Y)$, how do we compute $P(X)$?

Marginal Probability of X:

$X \setminus Y$	1	2	3	4	5	6	
1							$1/6$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	**1/6**
3							$1/6$
4			$P(X, Y)$				$1/6$
5							$1/6$
6							$1/6$
	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	1

$$P(X = 2) = \sum_{y=1}^{6} P(X = 2, Y = y) = 1/6$$

Overview

- Organisational Matters
- Models
- Maximum Likelihood Parameter Estimation
- Probability Theory
- **Bayesian Learning**
  - The Bayesian Distribution
  - From Prior to Posterior
  - MAP Parameter Estimation
  - Bayesian Predictions
  - Discussion
  - Advanced Issues

Bayesian Learning

Very popular:

- Bayesian learning can be used with any model, and even if we have multiple models.
- It is widely used in machine learning.

Nice properties:

- It avoids overfitting.
- Makes preference bias clearly visible.

Bayesian Learning

Very popular:

- Bayesian learning can be used with any model, and even if we have multiple models.
- It is widely used in machine learning.

Nice properties:

- It avoids overfitting.
- Makes preference bias clearly visible.

Main idea:

- Given some model with parameter θ, construct a **single** distribution P_{Bayes} on both data D and the parameter θ.
- Now we can compute the probability of
  - parameters given the training data: $P_{\text{Bayes}}(\theta = 3/4 \mid D)$;
  - the next outcome given the training data: $P_{\text{Bayes}}(y_{n+1} = 1 \mid D)$.

Overview

- Organisational Matters
- Models
- Maximum Likelihood Parameter Estimation
- Probability Theory
- Bayesian Learning
  - **The Bayesian Distribution**
  - From Prior to Posterior
  - MAP Parameter Estimation
  - Bayesian Predictions
  - Discussion
  - Advanced Issues

The Bayesian Distribution

Prior Distribution:

- A model contains **many** distributions. For example, $\mathcal{M} = \{P_\theta \mid \theta \in \{1, \ldots, 10\}\}$.
- We put a **prior distribution** π on the parameter θ.

The Bayesian Distribution

Prior Distribution:

- A model contains **many** distributions. For example, $\mathcal{M} = \{P_\theta \mid \theta \in \{1, \ldots, 10\}\}$.
- We put a **prior distribution** π on the parameter θ.
- $\pi(\theta)$ reflects our *a priori* [1] degree of belief that θ is the right parameter.

The Bayesian Distribution

Prior Distribution:

- A model contains **many** distributions. For example, $\mathcal{M} = \{P_\theta \mid \theta \in \{1, \ldots, 10\}\}$.
- We put a **prior distribution** π on the parameter θ.
- $\pi(\theta)$ reflects our *a priori* [1] degree of belief that θ is the right parameter.

Definition of P_{Bayes}:

- The **single** distribution P_{Bayes} on both parameters and data is defined by:

$$P_{\text{Bayes}}(\theta) = \pi(\theta) \quad \text{and} \quad P_{\text{Bayes}}(D \mid \theta) = P_\theta(D)$$

- This implies that $P_{\text{Bayes}}(D, \theta) = P_\theta(D)\pi(\theta)$

[1] "A priori" means before seeing the data.

Example

Model, prior and training data:

- Model: $\mathcal{M} = \{P_\theta \mid \theta \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}\}$ where $P_\theta(y_n = 1) = \theta$.
- Prior: $\pi\left(\frac{1}{4}\right) = \pi\left(\frac{1}{2}\right) = \pi\left(\frac{3}{4}\right) = \frac{1}{3}$
- Data: $D =$

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
0	1	1	1	0	1	1	1

Joint Probabilities:

$P_{\text{Bayes}}(D, \theta) = P_\theta(D)\pi(\theta)$:

θ	$P_{\text{Bayes}}(D, \theta)$
$1/4$	$1/3 \cdot (1/4)^6(3/4)^2 = 9/196608$
$1/2$	$1/3 \cdot (1/2)^8 = 256/196608$
$3/4$	$1/3 \cdot (3/4)^6(1/4)^2 = 729/196608$

The Marginal Probability of the Data

The marginal probability of the data:

$$P_{\text{Bayes}}(D) = \sum_{\theta} P_{\text{Bayes}}(D, \theta) = \sum_{\theta} P_{\theta}(D)\pi(\theta)$$

Example:

θ	$P_{\text{Bayes}}(D, \theta)$
$1/4$	$9/196608$
$1/2$	$256/196608$
$3/4$	$729/196608$

$$\Longrightarrow \quad P_{\text{Bayes}}(D) = \frac{9 + 256 + 729}{196608} = \frac{994}{196608}$$

Remarks:

- The marginal probability $P_{\text{Bayes}}(D)$ is a weighted average of $P_{\theta}(D)$, where each θ has the weight $\pi(\theta)$.
- This weight $\pi(\theta)$ does not depend on the data.

Overview

- Organisational Matters
- Models
- Maximum Likelihood Parameter Estimation
- Probability Theory
- Bayesian Learning
  - The Bayesian Distribution
  - **From Prior to Posterior**
  - MAP Parameter Estimation
  - Bayesian Predictions
  - Discussion
  - Advanced Issues

From Prior to Posterior Distribution

Updating beliefs:

- The prior $\pi(\theta)$ gives the probability of θ **before** we observe any data.
- The **posterior distribution** $P_{\text{Bayes}}(\theta \mid D)$ gives the probability of θ **after** observing data D.

From Prior to Posterior Distribution

Updating beliefs:

- The prior $\pi(\theta)$ gives the probability of θ **before** we observe any data.
- The **posterior distribution** $P_{\text{Bayes}}(\theta \mid D)$ gives the probability of θ **after** observing data D.
- This is the Bayesian way to update beliefs about parameters based on data D.

From Prior to Posterior Distribution

Updating beliefs:

- The prior $\pi(\theta)$ gives the probability of θ **before** we observe any data.
- The **posterior distribution** $P_{\text{Bayes}}(\theta \mid D)$ gives the probability of θ **after** observing data D.
- This is the Bayesian way to update beliefs about parameters based on data D.

Notation:

- The prior and the posterior both represent beliefs about θ.
- It is therefore common to write $\pi(\theta \mid D)$ for $P_{\text{Bayes}}(\theta \mid D)$.

Example

Previous example continued:

θ	$P_{\text{Bayes}}(D, \theta)$
$1/4$	$9/196608$
$1/2$	$256/196608$
$3/4$	$729/196608$

$$\Longrightarrow \quad P_{\text{Bayes}}(D) = \frac{994}{196608}$$

Posterior probability:

$$\pi(\theta \mid D) = \frac{P_{\text{Bayes}}(D, \theta)}{P_{\text{Bayes}}(D)} \quad \Longrightarrow$$

θ	$\pi(\theta \mid D)$
$1/4$	$\frac{9/196608}{994/196608} = 9/994$
$1/2$	$\frac{256/196608}{994/196608} = 256/994$
$3/4$	$\frac{729/196608}{994/196608} = 729/994$

- We started with equal prior probabilities.
- After observing the data, $\theta = 3/4$ is considered much more likely than the other θ.

Overview

- Organisational Matters
- Models
- Maximum Likelihood Parameter Estimation
- Probability Theory
- Bayesian Learning
  - The Bayesian Distribution
  - From Prior to Posterior
  - **MAP Parameter Estimation**
  - Bayesian Predictions
  - Discussion
  - Advanced Issues

MAP Parameter Estimation

Definition:

The maximum a posteriori (MAP) parameter estimate is the parameter with largest posterior (= a posteriori) probability:

$$\theta_{\text{MAP}} = \arg\max_\theta \pi(\theta \mid D)$$

Example continued:

θ	$\pi(\theta \mid D)$
$1/4$	$9/994$
$1/2$	$256/994$
$3/4$	$729/994$

$$\Longrightarrow \quad \theta_{\text{MAP}} = 3/4$$

Overview

- Organisational Matters
- Models
- Maximum Likelihood Parameter Estimation
- Probability Theory
- Bayesian Learning
  - The Bayesian Distribution
  - From Prior to Posterior
  - MAP Parameter Estimation
  - **Bayesian Predictions**
  - Discussion
  - Advanced Issues

The Predictive Distribution

Definition:

- Suppose $D = y_1, \ldots, y_n$.
- Then the Bayesian predictive distribution is $P_{\text{Bayes}}(y_{n+1} \mid D)$.

Understanding the predictive distribution:

It can be shown that:

$$P_{\text{Bayes}}(y_{n+1} \mid D) = \sum_{\theta} P_\theta(y_{n+1}) \pi(\theta \mid D)$$

- The predictive probability $P_{\text{Bayes}}(y_{n+1} \mid D)$ is a weighted average of $P_\theta(y_{n+1})$, where each θ has the weight $\pi(\theta \mid D)$.

Example Continued

Previous example continued:

- Recall that in this example $P_\theta(y_{n+1} = 1) = \theta$.

θ	$\pi(\theta \mid D)$
0.25	$9/994$
0.5	$256/994$
0.75	$729/994$

Predictive probability:

$$
\begin{aligned}
P_{\text{Bayes}}(y_{n+1} = 1 \mid D) &= \sum_{\theta=1}^{3} P_\theta(y_{n+1} = 1)\pi(\theta \mid D) \\
&= \frac{1}{4} \cdot \frac{9}{994} + \frac{1}{2} \cdot \frac{256}{994} + \frac{3}{4} \cdot \frac{729}{994} \\
&\approx 0.68
\end{aligned}
$$

- Notice that 0.68 is pretty close to 0.75.

Overview

MAP versus Predictive Distribution

- Prediction with map: $P_{\theta_{\text{MAP}}}(y_{n+1})$, where $\theta_{\text{MAP}} = \arg\max_\theta \pi(\theta \mid D)$
- Predictive distribution: $\sum_\theta P_\theta(y_{n+1})\pi(\theta \mid D)$

New example:

Two hypotheses that predict a 1 with high probability, one MAP hypothesis that predicts a 0 with high probability:

$P_\theta(y_{n+1} = 1)$	$1/10$	$8/10$	$9/10$
$\pi(\theta \mid D)$	$4/10$	$3/10$	$3/10$

$$P_{\text{Bayes}}(y_{n+1} = 1 \mid D) = \frac{4 \cdot 1}{100} + \frac{3 \cdot 8}{100} + \frac{3 \cdot 9}{100} = \frac{55}{100}$$

- Together the hypotheses that predict 1 have higher posterior probability than the MAP hypothesis that predicts 0.
- If we use the MAP, then we ignore their predictions!

The Prior Determines the Preference Bias

Marginal probability of the data:

$$P_{\text{Bayes}}(D) = \sum_{\theta} P_{\text{Bayes}}(D, \theta) = \sum_{\theta} P_{\theta}(D)\pi(\theta)$$

Posterior distribution:

$$\pi(\theta \mid D) = \frac{P_{\text{Bayes}}(D, \theta)}{P_{\text{Bayes}}(D)} = \frac{P_{\theta}(D)\pi(\theta)}{P_{\text{Bayes}}(D)}$$

Dependence on the prior:

- The most important probabilities in Bayesian inference.
- Both use $P_{\theta}(D)$ and $\pi(\theta)$.
- $P_{\theta}(D)$ depends on the data, but $\pi(\theta)$ does not!
- $\pi(\theta)$ determines the relative importance of each parameter θ.

The Prior Determines the Preference Bias

Marginal probability of the data:

$$P_{\text{Bayes}}(D) = \sum_{\theta} P_{\text{Bayes}}(D, \theta) = \sum_{\theta} P_{\theta}(D)\pi(\theta)$$

Posterior distribution:

$$\pi(\theta \mid D) = \frac{P_{\text{Bayes}}(D, \theta)}{P_{\text{Bayes}}(D)} = \frac{P_{\theta}(D)\pi(\theta)}{P_{\text{Bayes}}(D)}$$

Dependence on the prior:

- The most important probabilities in Bayesian inference.
- Both use $P_{\theta}(D)$ and $\pi(\theta)$.
- $P_{\theta}(D)$ depends on the data, but $\pi(\theta)$ does not!
- $\pi(\theta)$ determines the relative importance of each parameter θ.
- However, if we get a lot of data, then the effect of $P_{\theta}(D)$ becomes much more important than the effect of the prior.

Overview

- Organisational Matters
- Models
- Maximum Likelihood Parameter Estimation
- Probability Theory
- Bayesian Learning
  - The Bayesian Distribution
  - From Prior to Posterior
  - MAP Parameter Estimation
  - Bayesian Predictions
  - Discussion
  - **Advanced Issues**

Different Interpretations of Probability

- Suppose P is a distribution on $\Omega = \{a, b, c, d, e, f, g\}$ and $A = \{c, d, f\}$ is an event.

Frequentist: If we perform this same experiment n times, then the **relative frequency** of observing an outcome in A goes to $P(A)$ as $n \to \infty$.

Subjective Bayesian:[2] Before observing the outcome of the experiment, $P(A)$ is our **degree of belief** that we will get an outcome in A.

[2] There are other Bayesian interpretations of probability as well.

Different Interpretations of Probability

- Suppose P is a distribution on $\Omega = \{a, b, c, d, e, f, g\}$ and $A = \{c, d, f\}$ is an event.

Frequentist: If we perform this same experiment n times, then the **relative frequency** of observing an outcome in A goes to $P(A)$ as $n \to \infty$.

- Considers infinite number of repetitions of the experiment.
- Requires that it is possible (in principle) to observe the outcome of the experiment.
- Objective, the same for everyone.

Subjective Bayesian:[2] Before observing the outcome of the experiment, $P(A)$ is our **degree of belief** that we will get an outcome in A.

- Considers only one repetition of the experiment.
- Does not require that we can observe the outcome of the experiment.
- Subjective: My probability may be different from your probability.

[2] There are other Bayesian interpretations of probability as well.

Overview

- Organisational Matters
- Models
- Maximum Likelihood Parameter Estimation
- Probability Theory
- Bayesian Learning
  - The Bayesian Distribution
  - From Prior to Posterior
  - MAP Parameter Estimation
  - Bayesian Predictions
  - Discussion
  - Advanced Issues

References

- A.N. Shiryaev, “Probability”, Second Edition, 1996
- P. Grünwald, “The Minimum Description Length Principle”, 2007
- T.M. Mitchell, “Machine Learning”, McGraw-Hill, 1997