Machine Learning 2007: Lecture 10

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

November 21, 2007

Overview

- **Rogier: Weka Demonstration**
- Organisational Matters
- Naive Bayes Continued
- Probability Theory
  - I.I.D. Distributions
  - Distributions on $\mathbb{R}$
- Models
- Maximum Likelihood Parameter Estimation
- Bayesian Learning (Part 1)

Overview

- Rogier: Weka Demonstration
- **Organisational Matters**
- Naive Bayes Continued
- Probability Theory
  - I.I.D. Distributions
  - Distributions on $\mathbb{R}$
- Models
- Maximum Likelihood Parameter Estimation
- Bayesian Learning (Part 1)

This Lecture versus Mitchell

This Lecture:

- Section 6.9 about naive Bayes.
- Chapter 6 up to section 6.5.0 about Bayesian learning.
- I present things in a better order.
- We will continue with Bayesian learning in the next lecture.

This Lecture versus Mitchell

This Lecture:

- Section 6.9 about naive Bayes.
- Chapter 6 up to section 6.5.0 about Bayesian learning.
- I present things in a better order.
- We will continue with Bayesian learning in the next lecture.

WARNING versus Mitchell:

- Although naive Bayes is in the chapter about Bayesian learning (explained in the next lecture), Mitchell does not explain how it can be viewed as a Bayesian method, which is not trivial!
- The way Mitchell presents naive Bayes, it does not look like a Bayesian method at all.

Overview

- Rogier: Weka Demonstration
- Organisational Matters
- **Naive Bayes Continued**
- Probability Theory
  - I.I.D. Distributions
  - Distributions on $\mathbb{R}$
- Models
- Maximum Likelihood Parameter Estimation
- Bayesian Learning (Part 1)

Naive Bayes

Classification:

- Suppose we want to classify d-dimensional feature vector $\mathbf{x}$.
- Then select the label y with highest conditional probability:

$$\begin{aligned}
\arg\max_y P(Y = y \mid X = \mathbf{x}) \\
&= \arg\max_y \frac{P(X = \mathbf{x} \mid Y = y) P(Y = y)}{P(X = \mathbf{x})} \\
&= \arg\max_y P(X = \mathbf{x} \mid Y = y) P(Y = y) \\
&= \arg\max_y \prod_{i=1}^{d} P(X_i = x_i \mid Y = y) P(Y = y)
\end{aligned}$$

- The last step assumes that the components of $\mathbf{x}$ are conditionally independent given the class label y.
- Probabilities are estimated from training data.

Naive Bayes Example

Fairy tale data set:

x_1 WearsBlack	x_2 SavesPrincess	x_3 HorseColour	y GoodOrEvil
No	Yes	Black	Good
Yes	No	Black	Evil
No	No	White	Good
Yes	Yes	Brown	Good

Classifying the new instance $\begin{pmatrix} \textbf{No} \\ \textbf{Yes} \\ \textbf{White} \end{pmatrix}$:

$$\prod_{i=1}^{3} P(X_i = x_i \mid Y = \text{Good}) P(Y = \text{Good}) = \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{3}{4}$$

$$> \prod_{i=1}^{3} P(X_i = x_i \mid Y = \text{Evil}) P(Y = \text{Evil}) = 0 \cdot 0 \cdot 0 \cdot \frac{1}{4}$$

Inductive Bias

Incorrect independence assumption:

- The assumption that components of $\mathbf{x}$ are conditionally independent given the class label is very strong. In fact it is often known to be false.

Inductive Bias

Incorrect independence assumption:

- The assumption that components of $\mathbf{x}$ are conditionally independent given the class label is very strong. In fact it is often known to be false.
- For example, naive Bayes is often used to classify e-mail as spam or not spam. Each component of $\mathbf{x}$ represents a word in the text of an e-mail.
- If one of the words ‘OEM’ and ‘software’ occurs in a spam message, then the other one is more likely to occur as well.
- Hence the components of $\mathbf{x}$ are clearly not independent.

Inductive Bias

Incorrect independence assumption:

- The assumption that components of $\mathbf{x}$ are conditionally independent given the class label is very strong. In fact it is often known to be false.
- For example, naive Bayes is often used to classify e-mail as spam or not spam. Each component of $\mathbf{x}$ represents a word in the text of an e-mail.
- If one of the words ‘OEM’ and ‘software’ occurs in a spam message, then the other one is more likely to occur as well.
- Hence the components of $\mathbf{x}$ are clearly not independent.

But it works anyway:

According to [Domingos and Pazzani, 1996]:

- Even if $P(y \mid \mathbf{x})$ is not estimated correctly;
- Often $\arg\max_y P(y \mid \mathbf{x})$ is still correct.

Overview

- Rogier: Weka Demonstration
- Organisational Matters
- Naive Bayes Continued
- Probability Theory
  - **I.I.D. Distributions**
  - Distributions on $\mathbb{R}$
- Models
- Maximum Likelihood Parameter Estimation
- Bayesian Learning (Part 1)

I.I.D. Distributions

Definition:

- Suppose we have data $D = y_1, \ldots, y_n$.
- Suppose each outcome y_i is distributed according to the same distribution P that does not depend on the previous outcomes $y_1, \ldots, y_{i-1}$.
- Then we say that the outcomes $y_1, \ldots, y_n$ are **independent and identically distributed** (i.i.d.).
- We have that $P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} P(Y_i = y_i)$.

I.I.D. Distributions

Definition:

- Suppose we have data $D = y_1, \ldots, y_n$.
- Suppose each outcome y_i is distributed according to the same distribution P that does not depend on the previous outcomes $y_1, \ldots, y_{i-1}$.
- Then we say that the outcomes $y_1, \ldots, y_n$ are **independent and identically distributed** (i.i.d.).
- We have that $P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} P(Y_i = y_i)$.

Example:

- Suppose we draw six cards $y_1, \ldots, y_6$ from a deck **with replacement**.
- Then for each draw y_i the probability of drawing, say, a queen of hearts, is the same and does not depend on our previous draws: The draws are i.i.d.
- Without replacement, the draws would not be i.i.d!

Overview

- Rogier: Weka Demonstration
- Organisational Matters
- Naive Bayes Continued
- Probability Theory
  - I.I.D. Distributions
  - **Distributions on $\mathbb{R}$**
- Models
- Maximum Likelihood Parameter Estimation
- Bayesian Learning (Part 1)

Distributions on $\mathbb{R}$

Finite sample space:

- Suppose $\Omega = \{\omega_1, \ldots, \omega_m\}$
- Then the probability of an event $A \subseteq \Omega$ is

$$P(A) = \sum_{\omega_i \in A} p(\omega_i),$$

- where the **mass function** p satisfies:
  1. $0 \leq p(\omega) \leq 1$ (for all $\omega \in \Omega$)
  2. $p(\omega_1) + \ldots + p(\omega_m) = 1$
- Note that, for all $\omega \in \Omega$, $P(\{\omega\}) = p(\omega)$.

The sample space $\mathbb{R}$:

- Suppose $\Omega = \mathbb{R}$.
- Then the probability of an event $A \subseteq \Omega$ is

$$P(A) = \int_{x \in A} p(x)\, dx,$$

- where the **density function** p satisfies:
  1. $0 \leq p(x)$ (for all $x \in \Omega$)
  2. $\int_{x \in \Omega} p(x)\, dx = 1$
- Note that, for all $x \in \Omega$, $P(\{x\}) = 0 \neq p(x)$!

Example: The Uniform Distribution

Finite sample space:

- Suppose $\Omega = \{\omega_1, \ldots, \omega_m\}$
- Then the **uniform distribution** on Ω gives the same probability to all outcomes.
- Its mass function is given by

$$p(\omega) = 1/m.$$

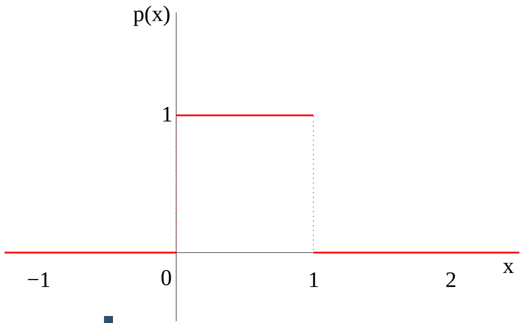
Examples:

- $P(\{\omega_1, \ldots, \omega_{m/2}\}) = \frac{1}{2}$
- $P(\{\omega_i\}) = 1/m = p(\omega_i)$

The interval $[0, 1]$:

- Suppose $\Omega = \mathbb{R}$.
- Then the uniform distribution on $[0, 1]$ is defined by the density function

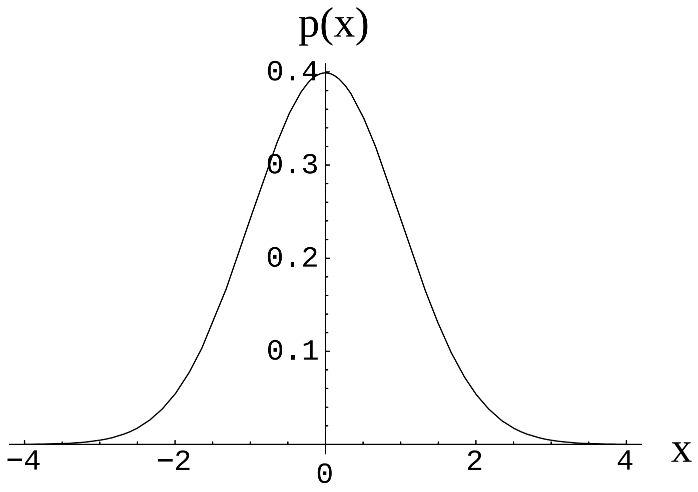
$$p(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$



Examples:

- $P([0, \frac{1}{2}]) = \frac{1}{2}$
- $P(\{0.1\}) = 0 \neq 1 = p(0.1)$

Example: The Normal Distribution



$$p_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Remarks:

- Its **mean** μ controls where it is centered.
- Its **variance** σ^2 controls how spread out it is (larger variance makes it flatter and wider).
- The normal distribution is also called the Gaussian distribution.

Overview

- Rogier: Weka Demonstration
- Organisational Matters
- Naive Bayes Continued
- Probability Theory
  - I.I.D. Distributions
  - Distributions on $\mathbb{R}$
- **Models**
- Maximum Likelihood Parameter Estimation
- Bayesian Learning (Part 1)

Models

Definition:

A **(statistical) model** is a hypothesis space that contains only probability distributions.

Models

Definition:

A **(statistical) model** is a hypothesis space that contains only probability distributions.

Example: the Bernoulli model for prediction

- For binary outcomes $y \in \{0, 1\}$ define the **Bernoulli distribution** with probability of success θ by

$$p_\theta(y) = \theta^y (1-\theta)^{1-y} = \begin{cases} \theta & \text{if } y = 1, \\ 1-\theta & \text{if } y = 0. \end{cases}$$

- Then the **Bernoulli model** (with parameter θ) is the set of all possible Bernoulli distributions[1]:

$$\mathcal{M}_{\text{Bernoulli}} = \{p_\theta \mid \theta \in [0, 1]\}$$

[1] For the remainder of the lectures I will be a bit sloppy about the distinction between distributions and density functions to avoid distracting technicalities.

Models in Classification or Regression

The label depends on the input:

- In classification or regression we get an input $\mathbf{x}$ and we need to produce an output y.
- Thus our estimate of y will depend on the input $\mathbf{x}$ that we get.
- For example (for 1-dimensional x): $y = 3 + 2x + x^2$.

Models in Classification or Regression

The label depends on the input:

- In classification or regression we get an input $\mathbf{x}$ and we need to produce an output y.
- Thus our estimate of y will depend on the input $\mathbf{x}$ that we get.
- For example (for 1-dimensional x): $y = 3 + 2x + x^2$.

The same holds with models:

For example, for binary $y \in \{0, 1\}$ and 1-dimensional x define the model $\mathcal{M} = \{p_{\theta,x} \mid \theta \in [0, 1]\}$ (with parameter θ), where

$$p_{\theta,x}(y) = \begin{cases} \theta^y(1-\theta)^{1-y} & \text{if } x < 0, \\ 1 - \theta^y(1-\theta)^{1-y} & \text{if } x \geq 0. \end{cases}$$

Models in Classification or Regression

The label depends on the input:

- In classification or regression we get an input $\mathbf{x}$ and we need to produce an output y.
- Thus our estimate of y will depend on the input $\mathbf{x}$ that we get.
- For example (for 1-dimensional x): $y = 3 + 2x + x^2$.

The same holds with models:

For example, for binary $y \in \{0, 1\}$ and 1-dimensional x define the model $\mathcal{M} = \{p_{\theta,x} \mid \theta \in [0, 1]\}$ (with parameter θ), where

$$p_{\theta,x}(y) = \begin{cases} \theta^y(1-\theta)^{1-y} & \text{if } x < 0, \\ 1 - \theta^y(1-\theta)^{1-y} & \text{if } x \geq 0. \end{cases}$$

- We are usually interested in distributions on y; $\mathbf{x}$ is considered as given.
- Naive Bayes is an exception.

From Hypothesis Space to Model

Deterministic hypotheses + noise. . .

- Suppose $\mathcal{H} = \{h_{\mathbf{w}} \mid \mathbf{w} \in \mathbb{R}^3\}$ is the set of all 2nd degree polynomials: $h_{\mathbf{w}}(x) = w_0 + w_1 x + w_2 x^2$.
- Suppose we assume normally distributed noise ϵ with mean $\mu = 0$ and variance $\sigma = 1$.
- Then $y = h_{\mathbf{w}}(x) + \epsilon$.

From Hypothesis Space to Model

Deterministic hypotheses + noise...

- Suppose $\mathcal{H} = \{h_{\mathbf{w}} \mid \mathbf{w} \in \mathbb{R}^3\}$ is the set of all 2nd degree polynomials: $h_{\mathbf{w}}(x) = w_0 + w_1 x + w_2 x^2$.
- Suppose we assume normally distributed noise ϵ with mean $\mu = 0$ and variance $\sigma = 1$.
- Then $y = h_{\mathbf{w}}(x) + \epsilon$.

...gives distributions:

- Adding $h_{\mathbf{w}}(x)$ to a normal distribution only changes its mean: $\mu = 0 + h_{\mathbf{w}}(x)$.
- Hence the density of y is $\frac{1}{\sqrt{2\pi}} e^{-\frac{(y - h_{\mathbf{w}}(x))^2}{2}}$.
- So we get the model $\mathcal{M} = \{p_{\mathbf{w},x} \mid \mathbf{w} \in \mathbb{R}^3\}$ (with parameters $\mathbf{w}$), where

$$p_{\mathbf{w},x}(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y - h_{\mathbf{w}}(x))^2}{2}}$$

Overview

- Rogier: Weka Demonstration
- Organisational Matters
- Naive Bayes Continued
- Probability Theory
  - I.I.D. Distributions
  - Distributions on $\mathbb{R}$
- Models
- **Maximum Likelihood Parameter Estimation**
- Bayesian Learning (Part 1)

Maximum Likelihood Parameter Estimation

Parameter Estimation:

- **Model** $\mathcal{M} = \{p_\theta \mid \theta \in \Theta\}$ with parameter θ. (Θ is the set of possible parameter values.)
- **Data** $D = d_1, \ldots, d_n$, which is distributed according to an unknown distribution $p_{\theta^*} \in \mathcal{M}$.
- We want to **estimate the parameter** θ^* from the data D.

Maximum Likelihood Parameter Estimation

Parameter Estimation:

- **Model** $\mathcal{M} = \{p_\theta \mid \theta \in \Theta\}$ with parameter θ. (Θ is the set of possible parameter values.)
- **Data** $D = d_1, \ldots, d_n$, which is distributed according to an unknown distribution $p_{\theta^*} \in \mathcal{M}$.
- We want to **estimate the parameter** θ^* from the data D.

Maximum Likelihood:

Maximum likelihood parameter estimation selects the parameter $\hat{\theta}$ that maximizes the density[2] of the data:

$$\hat{\theta} = \arg\max_\theta p_\theta(D)$$

[2] If D takes values in a finite sample space, then the probability mass is used instead of the density.

Maximum Likelihood in the Bernoulli Model

Bernoulli distribution for n outcomes:

- Given binary data $D = y_1, \ldots, y_n$, we want to predict y_{n+1}.
- We assume that the outcomes in D are i.i.d. according to a Bernoulli distribution.
- If n_0 and n_1 respectively denote the number of zeroes and ones in D, then

$$p_\theta(D) = \theta^{n_1}(1-\theta)^{n_0}$$

Maximum Likelihood:[3]

$$\hat{\theta} = \arg\max_\theta \theta^{n_1}(1-\theta)^{n_0} = \arg\max_\theta\, n_1 \ln\theta + n_0 \ln(1-\theta)$$

Solving $\frac{d}{d\theta} n_1 \ln\theta + n_0 \ln(1-\theta) = 0$, gives: $\hat{\theta} = \frac{n_1}{n_1+n_0} = \frac{n_1}{n}$.

[3] Ignoring minor technical issues for $\theta = 0$ or $\theta = 1$.

Least Mean Squares as Maximum Likelihood

Model with normally distributed noise:

- Suppose we get i.i.d. data $D = (y_1, x_1)^\top, \ldots, (y_n, x_n)^\top$.
- We use the model of **second degree polynomials** with normally distributed noise, with $\mu = 0$ and $\sigma = 1$.
- Then, writing x^n for $x_1, \ldots, x_n$,

$$p_{\mathbf{w},x^n}(y_1, \ldots, y_n) = \prod_{i=1}^{n} p_{\mathbf{w},x_i}(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - h_{\mathbf{w}}(x_i))^2}{2}}$$

Least Mean Squares as Maximum Likelihood

Model with normally distributed noise:

- Suppose we get i.i.d. data $D = (y_1, x_1)^\top, \ldots, (y_n, x_n)^\top$.
- We use the model of **second degree polynomials** with normally distributed noise, with $\mu = 0$ and $\sigma = 1$.
- Then, writing x^n for $x_1, \ldots, x_n$,

$$p_{\mathbf{w},x^n}(y_1, \ldots, y_n) = \prod_{i=1}^{n} p_{\mathbf{w},x_i}(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - h_{\mathbf{w}}(x_i))^2}{2}}$$

Maximum likelihood gives least mean squares:

$$\arg\max_{\mathbf{w}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - h_{\mathbf{w}}(x_i))^2}{2}} = \arg\max_{\mathbf{w}} \ln \prod_{i=1}^{n} e^{-\frac{(y_i - h_{\mathbf{w}}(x_i))^2}{2}}$$

$$= \arg\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - h_{\mathbf{w}}(x_i))^2$$

Least Mean Squares as Maximum Likelihood

Model with normally distributed noise:

- Suppose we get i.i.d. data $D = (y_1, x_1)^\top, \ldots, (y_n, x_n)^\top$.
- We use the model of **second degree polynomials** with normally distributed noise, with $\mu = 0$ and $\sigma = 1$.
- Then, writing x^n for $x_1, \ldots, x_n$,

$$p_{\mathbf{w},x^n}(y_1, \ldots, y_n) = \prod_{i=1}^{n} p_{\mathbf{w},x_i}(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - h_{\mathbf{w}}(x_i))^2}{2}}$$

Maximum likelihood gives least mean squares:

$$\arg\max_{\mathbf{w}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - h_{\mathbf{w}}(x_i))^2}{2}} = \arg\max_{\mathbf{w}} \ln \prod_{i=1}^{n} e^{-\frac{(y_i - h_{\mathbf{w}}(x_i))^2}{2}}$$

$$= \arg\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - h_{\mathbf{w}}(x_i))^2$$

Remark: Maximum likelihood will overfit if we apply it to a very large hypothesis space/model. (E.g. high degree polynomials.)

Overview

- Rogier: Weka Demonstration
- Organisational Matters
- Naive Bayes Continued
- Probability Theory
  - I.I.D. Distributions
  - Distributions on $\mathbb{R}$
- Models
- Maximum Likelihood Parameter Estimation
- **Bayesian Learning (Part 1)**

Bayesian Learning

Very important:

- Bayesian learning is a general framework for doing machine learning that can be used with any model.
- It avoids overfitting.
- It is widely used in machine learning.

Bayesian Learning

Very important:

- Bayesian learning is a general framework for doing machine learning that can be used with any model.
- It avoids overfitting.
- It is widely used in machine learning.

Motivation:

- A model $\mathcal{M} = \{P_\theta \mid \theta \in \Theta\}$ contains **many** distributions P_θ for the data $D \in \Omega$.
- Suppose we want to calculate the probability $P(\theta \mid D)$.
- Then this is not defined: What is P? What is its sample space?

The Bayesian Distribution

The Idea:

- We start with a model $\mathcal{M} = \{P_\theta \mid \theta \in \Theta\}$, which contains many distributions.
- Then we put a **prior distribution** π on the parameter θ.
- We get a single distribution P_{Bayes} on both parameters and data!

The Bayesian Distribution

The Idea:

- We start with a model $\mathcal{M} = \{P_\theta \mid \theta \in \Theta\}$, which contains many distributions.
- Then we put a **prior distribution** π on the parameter θ.
- We get a single distribution P_{Bayes} on both parameters and data!

The details:

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

The Bayesian Distribution

The Idea:

- We start with a model $\mathcal{M} = \{P_\theta \mid \theta \in \Theta\}$, which contains many distributions.
- Then we put a **prior distribution** π on the parameter θ.
- We get a single distribution P_{Bayes} on both parameters and data!

The details:

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

- P_{Bayes} is a **single** distribution on $\Omega' = \Omega \times \Theta$, which contains both the data and θ.
- Therefore $P_{\text{Bayes}}(\theta \mid D)$ is well-defined.

Example

- Suppose our data consists of one binary outcome y.
- Consider the model $\mathcal{M} = \{P_\theta \mid \theta \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}\}$, where $P_\theta(y) = \theta^y(1-\theta)^{1-y}$ is a Bernoulli distribution.
- Take π to be the uniform distribution on $\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$.

Example

- Suppose our data consists of one binary outcome y.
- Consider the model $\mathcal{M} = \{P_\theta \mid \theta \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}\}$, where $P_\theta(y) = \theta^y(1-\theta)^{1-y}$ is a Bernoulli distribution.
- Take π to be the uniform distribution on $\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$.

$$\begin{aligned} P_{\text{Bayes}}\left(y = 1, \theta = \frac{1}{2}\right) &= P_{\text{Bayes}}\left(y = 1 \mid \theta = \frac{1}{2}\right) P_{\text{Bayes}}\left(\theta = \frac{1}{2}\right) \\ &= P_{\frac{1}{2}}(1) \cdot \pi\left(\frac{1}{2}\right) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} \end{aligned}$$

$$\begin{aligned} P_{\text{Bayes}}\left(y = 0, \theta = \frac{1}{4}\right) &= P_{\text{Bayes}}\left(y = 0 \mid \theta = \frac{1}{4}\right) P_{\text{Bayes}}\left(\theta = \frac{1}{4}\right) \\ &= P_{\frac{1}{4}}(0) \cdot \pi\left(\frac{1}{4}\right) = \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{4} \end{aligned}$$

Different Interpretations of Probability

- Suppose P is a distribution on Ω and $A \subseteq \Omega$ is an event.

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

Subjective Bayesian:[4] Before observing the outcome of the experiment, $P(A)$ is our **degree of belief** that we will get an outcome $\omega \in A$.

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

Different Interpretations of Probability

- Suppose P is a distribution on Ω and $A \subseteq \Omega$ is an event.

Frequentist: If we perform this same experiment n times, then the **relative frequency** of observing an outcome $\omega \in A$ 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:[4] Before observing the outcome of the experiment, $P(A)$ is our **degree of belief** that we will get an outcome $\omega \in A$.

- 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.

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

Overview

- Rogier: Weka Demonstration
- Organisational Matters
- Naive Bayes Continued
- Probability Theory
  - I.I.D. Distributions
  - Distributions on $\mathbb{R}$
- Models
- Maximum Likelihood Parameter Estimation
- Bayesian Learning (Part 1)

References

- P. Domingos and M. Pazzani, “Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier”, Proceedings of the 13th International Conference on Machine Learning, 1996
- A.N. Shiryaev, “Probability”, Second Edition, 1996
- P. Grünwald, “The Minimum Description Length Principle”, 2007
- T.M. Mitchell, “Machine Learning”, McGraw-Hill, 1997