An introduction to Bayesian statistics

Lennart Svensson

Signal Processing Group
Department of Signals and Systems
Chalmers University of Technology, Sweden

What is Bayesian statistics?

A statistical inference framework.

Can be used for estimation, classification, detection, model selection, etc.

Key characteristic: unknown quantities are described as random.

Applications of Bayesian statistics

- A medical application: analyze the disease of a patient.

– **Quantity of interest**: the disease, θ.

– **Observations**: blood samples, temperature, comments by patient, etc.

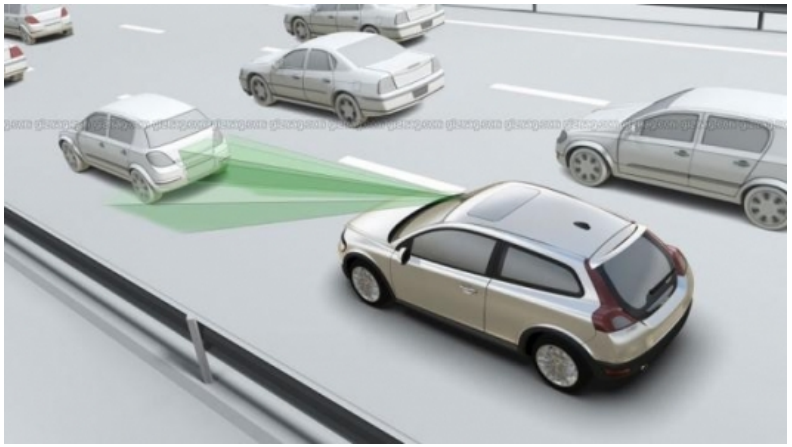
- In Bayesian statistics θ is described as random
  $\rightsquigarrow$ we can make statements like "patient has disease x with 97% probability". (e.g., x = flue)
- **Possible concern:** is the disease random?

Applications of Bayesian statistics

- Modern safety systems for cars rely on the ability to position surrounding vehicles.
- This enables the system to intervene, e.g., by braking, before a collision.

– **Quantity of interest:** relative position and velocity of other cars at time k.

– **Observations**: wheel speeds, radar detections (distance and angle), GPS measurements, etc.



- Bayesian statistics: vehicle motions are modelled statistically
  $\rightsquigarrow$ helps us to rule out unrealistic trajectories.
- **Possible concern:** are the vehicle motions random?

Comparison: Bayes vs Frequentist

- There are **two main strategies** to decision making: **Bayesian and frequentist statistics**.
- In **frequenstist statistics**, the quantities of interest are described as **unknown and deterministic**.

An example: suppose we wish to estimate the height of the eiffel tower. Is the height random or not?

- **Frequentist perspective:** the tower has a certain height and is therefore not random.
- **Bayesian perspective:** we describe our uncertainties in the height stochastically $\Rightarrow$ height is described as random!

Overview of the Bayesian strategy

Suppose we wish to estimate θ given measurements y.

Key steps in a Bayesian method:

1. **Modeling.** Model what we know about θ (using a prior $p(\theta)$) and the measurements y (using a density $p(y|\theta)$).
2. **Measurement update.** Combine what we knew before (the prior) with our measurement (with $p(y|\theta)$, also called the likelihood) to summarize what we know about θ ($p(\theta|y)$).
3. **Decision making.** Given what we know about θ (described by $p(\theta|y)$) and a loss function, we compute *an optimal decision*.

In the upcoming videos we first discuss how to perform step 2) and after that we look at step 3).

Self-assessment questions

Which of the following statements are correct:

- Bayesian methods can be used to solve many types of decision making problems including estimation, detection and classification.
- We can model the height of the Eiffel tower as random only if we think that there are many similar towers with different heights.
- In Bayesian statistics we describe what we know about θ (the quantity of interest) before observing any measurements.

Check all that apply.

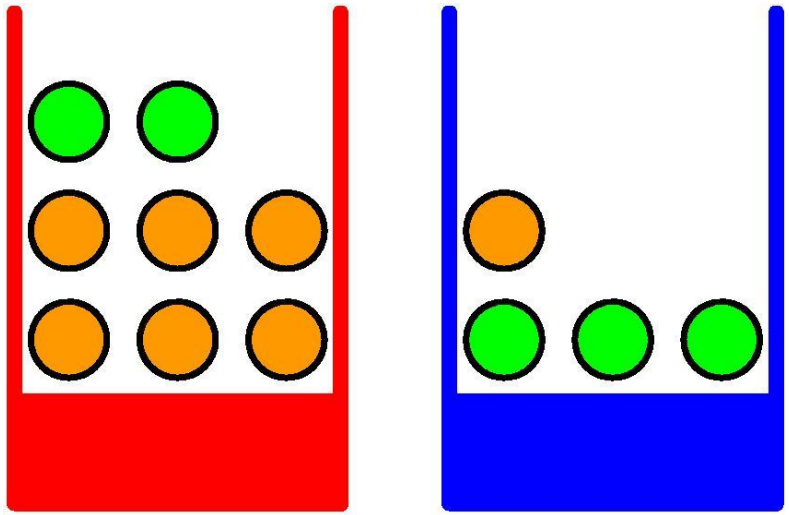
Bayes' rule – a first example

Lennart Svensson

Signal Processing Group
Department of Signals and Systems
Chalmers University of Technology, Sweden

Bayes' rule: a first example

- A box is selected at random (prob. 1/2, 1/2). From that box we pick a fruit.



- If fruit is orange, what is probability that box is red?

Probability theory

- Bayesian statistics is simple!
  We only need two rules:
  1. Conditional probability (product rule)

$$\Pr\{y, \theta\} = \Pr\{y|\theta\} \Pr\{\theta\}$$

  2. The law of total probability (sum rule)

$$\Pr\{y\} = \sum_{\theta} \Pr\{y, \theta\} \qquad \text{discrete variables}$$

$$p(y) = \int_{\theta} p(y, \theta)\, d\theta \qquad \text{continuous variables}$$

Probability theory – Bayes' rule

- Bayes' rule is a consequence of conditional probability,

$$p(y|\theta)p(\theta) = p(\theta|y)p(y).$$

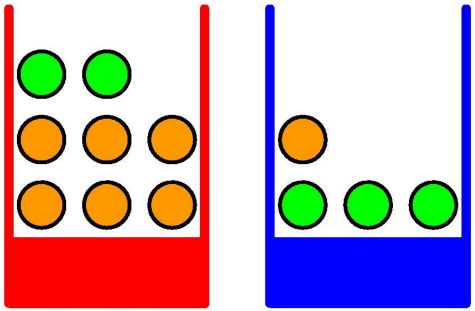
Bayes' rule

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- Usage of Bayes' rule: express a relation of interest, $p(\theta|y)$, in terms of the relation that we know $p(y|\theta)$.
- Note that $p(y) = \int_\theta p(y|\theta)p(\theta)\, d\theta$.

Bayes' rule: a first example

- Let $\theta \in \{r, b\}$ be color of box, and $y \in \{o, a\}$ be the fruit.



- Bayes' rule gives

$$\Pr\{\theta = r \mid y = o\} = \frac{\Pr\{y = o \mid \theta = r\} \cdot \Pr\{\theta = r\}}{\Pr\{y = o\}}$$

$$\Pr\{y = o, \theta = r\} + \Pr\{y = o, \theta = b\}$$

$$= \frac{3}{4} \cdot \frac{1}{2} \qquad \frac{1}{4} \cdot \frac{1}{2}$$

$$\Rightarrow \Pr\{\theta = r \mid y = o\} = \frac{3}{4}$$

Building blocks of Bayesian models

Lennart Svensson

Signal Processing Group
Department of Signals and Systems
Chalmers University of Technology, Sweden

Priors, posteriors and likelihoods

- We are interested in an unknown parameter

$$\theta$$

  such that $\theta \in \Theta$.

- **Common problem types** are *estimation* (e.g., $\Theta = \mathbb{R}^n$) and *detection* problems (e.g., $\Theta = \{-1, 1\}$).

- The observed data, y, is distributed as

$$y \sim p(y|\theta),$$

  where p is a known distribution.

Priors, posteriors and likelihoods

- Since data y is observed, we often view $p(y|\theta)$ as a function of θ,

$$l(\theta|y) = p(y|\theta),$$

where $l(\theta|y)$ is called the likelihood function.
Note: the likelihood function is *not* a density w.r.t. θ.

- In Bayesian statistics we have a prior distribution $p(\theta)$ on θ.
Prior means *earlier*, or before, and $p(\theta)$ describes what we know *before* observing y.

- One objective in Bayesian statistics is to compute the posterior

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto l(\theta|y)p(\theta)$$

Posterior means *after* and $p(\theta|y)$ describes what we know *after observing* y.

Priors, posteriors and likelihoods

- We summarize this as

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

- Given the posterior, $p(\theta|y)$ we can answer, e.g.,
  - What is the most probable θ?
  - What is the probability that $\theta \in \mathcal{A}$?
  - What is the posterior mean of θ?

- We can also minimize expected costs in a decision theoretic manner (see later videos).

Example: scalar in Gaussian noise

- Suppose we observe

$$y = \theta + v, \qquad v \sim \mathcal{N}(0, \sigma^2)$$

such that $p(y|\theta) = \mathcal{N}(y; \theta, \sigma^2) \propto \exp\{-(y-\theta)^2/(2\sigma^2)\}$.

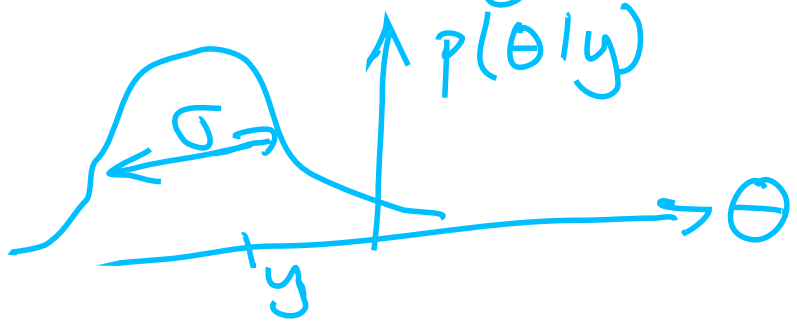
- A common noninformative prior on θ is

$$p(\theta) \propto 1.$$

- The posterior is $p(\theta|y) \propto p(\theta)p(y|\theta)$

$\propto 1 \cdot \exp(-(y-\theta)^2/2\sigma^2) \propto \mathcal{N}(\theta; y, \sigma^2)$

$\Rightarrow p(\theta|y) = \mathcal{N}(\theta; y, \sigma^2)$



Example: scalar in Gaussian noise

- Suppose we observe

$$y = \theta + v, \qquad v \sim \mathcal{N}(0, \sigma^2)$$

such that $p(y|\theta) = \mathcal{N}(y; \theta, \sigma^2) \propto \exp\{-(y-\theta)^2/(2\sigma^2)\}$.

- A common noninformative prior on θ is

$$p(\theta) \propto 1.$$

- The posterior is

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \propto \exp\{-(y-\theta)^2/(2\sigma^2)\} \\ &\propto \mathcal{N}(\theta; y, \sigma^2) \end{aligned}$$

$$\Rightarrow p(\theta|y) = \mathcal{N}(\theta; y, \sigma^2)$$

Bayesian approach to sensor fusion

- Suppose we collect measurements from **two types of sensors**:

  $$y_1 \quad \text{and} \quad y_2,$$

  say a laser scanner and a camera.

- **Bayesian sensor fusion** is simply Bayesian statistics. As always we seek the posterior distribution:

  $$p(\theta | \underbrace{y_1, y_2}_{y}) \propto p(\theta) p(\underbrace{y_1, y_2}_{y} | \theta).$$

- It is often reasonable to assume that

  $$p(y_1, y_2 | \theta) \approx p(y_1 | \theta) p(y_2 | \theta),$$

  i.e., that measurements are **conditionally independent**.

Self-assessment

The posterior distribution is $p(\theta|y) \propto p(y|\theta)p(\theta)$.
It is also true that:

- The normalization factor is **not always unique**?
- The posterior $p(\theta|y)$ can **always be uniquely determined** from the fact that $\int p(\theta|y)\, d\theta = 1$?
- The posterior distribution can **only be uniquely determined if** it is proportional to a well known distribution, e.g., a Gaussian.

Only one statement is correct.

Bayesian Decision Theory

Lennart Svensson

Signal Processing Group
Department of Signals and Systems
Chalmers University of Technology, Sweden

Bayesian decision principle

- How can we use $p(\theta|y)$ to make decisions?
- **Examples** of decision problems
  - How to invest money.
  - Select medicine to give to a patient
  - Estimate a parameter vector (may represent temperature, distance, etc).

Basic principle of Bayesian decision theory

- Minimize expected loss

  or, equivalently,

- Maximize expected utility.

Decision theory – a toy example

- A student wants to decide whether to take a course or not.
- Suppose $\theta \in \{\text{good course, fair course, bad course}\}$ and

	good course	fair course	bad course
$\Pr\{\theta \mid y\}$	0.3	0.3	0.4

- If the loss function is

	good course	fair course	bad course
Taking course	0	5	10
Not taking course	20	5	0

should he/she then take the course?

Decision theory – a toy example

- We can easily compute the expected loss for the two possible decisions.
  - Takes the course:
    $0 \times 0.3 + 5 \times 0.3 + 10 \times 0.4 = 5.5$.
  - Does not take the course:
    $20 \times 0.3 + 5 \times 0.3 + 0 \times 0.4 = 7.5$.

- It is thus better to take the course. (Of course, this conclusion has nothing to do with this course.)

Minimum posterior expected loss

- We often study loss functions

$$C(\theta, \hat{\theta})$$

  instead of utilities. (Typically, $C \geq 0$.)
- Let $\hat{\theta}$ denote an estimate of θ.

Optimal Bayesian decisions

Minimize the posterior expected loss

$$\hat{\theta} = \arg\min_a \mathbb{E}\left\{ C(\theta, a) \middle| y \right\}$$

where $\mathbb{E}\left\{ C(\theta, a) \middle| x \right\} = \int_\Theta C(\theta, a) p(\theta \middle| y)\, d\theta$

- Note: y is given (fixed) and θ is random.

Self-assessment

To make an optimal Bayesian decision it is sufficient to know:

- The prior, $p(\theta)$, the likelihood, $p(y|\theta)$, and a loss function $C(\theta, a)$.
- The likelihood, $p(y|\theta)$, and a loss function $C(\theta, a)$.
- The posterior distribution, $p(\theta|y)$, and a loss function, $C(\theta, a)$.

Check all statements that apply.

Comparison: Bayes vs Frequentist

Frequentist	Bayes
θ is fixed and unknown $\Rightarrow \theta$ is deterministic	Uncertainties in θ are described stochastically $\Rightarrow \theta$ is random
Maximum likelihood (ML) most famous estimator $\hat{\theta}_{ML} = \arg\max_\theta l(\theta \mid y)$	Minimum mean square error and maximum a posteriori estimators, e.g. $\hat{\theta}_{MAP} = \arg\max_\theta p(\theta) l(\theta \mid y)$
Study performance by averaging over y for fixed θ	Make decisions conditioned on the observation y.

- Note 1: most Bayesians also study frequentist performance.
- Note 2: many frequentists agree that parameters may be random in some situations.

Cost functions in Bayesian decision theory

Lennart Svensson

Signal Processing Group
Department of Signals and Systems
Chalmers University of Technology, Sweden

Bayesian decision theory – summary

- Bayesian decision theory relies on
  1. Likelihood: $p(y|\theta)$
  2. Prior distribution: $p(\theta)$
  3. Loss function: $C(\theta, a)$.

- Combining likelihood and prior gives posterior

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

Posterior and loss gives decisions

$$\hat{\theta} = \arg\min_a \int_\Theta C(\theta, a)p(\theta|y)\, d\theta.$$

The quadratic loss function

- **Example:** parameter estimation, $\theta \in \Theta = \mathbb{R}^n$.
- Most common loss function is the **quadratic loss**

$$C(\theta, a) = \left\|\theta - a\right\|_2^2 = (\theta - a)^T(\theta - a)$$

- Let: $\bar{\theta} = \mathbb{E}\left\{\theta \middle| y\right\}$, $\mathbf{P} = \mathrm{Cov}\{\theta|y\} = \mathbb{E}\left\{(\theta - \bar{\theta})(\theta - \bar{\theta})^T \middle| y\right\}$
- **Optimal estimator:**

$$\mathbb{E}\{C(\theta,a)|y\} = \mathbb{E}\{(\theta-a)^T(\theta-a)|y\}$$

$$= \mathbb{E}\{(\theta-\bar{\theta}+\bar{\theta}-a)^T(\theta-\bar{\theta}+\bar{\theta}-a)|y\} = \mathbb{E}\{(\theta-\bar{\theta})^T(\theta-\bar{\theta})|y\}$$

(zero mean, determ.)

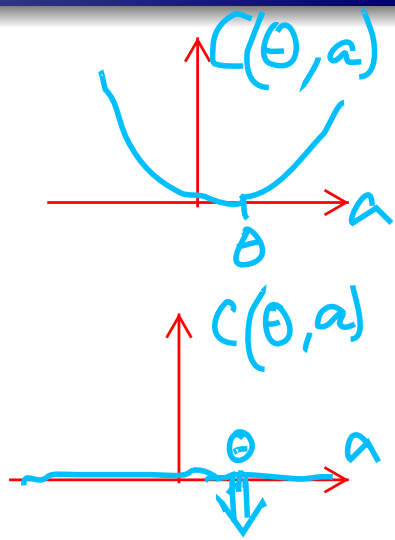
$$+ \mathbb{E}\{(\theta-\bar{\theta})^T|y\}(\bar{\theta}-a) + 0 + (\bar{\theta}-a)^T(\bar{\theta}-a)$$

$(=0)$ $\qquad \mathrm{tr}\{P\}$

$$\hat{\theta} = \underset{a}{\mathrm{argmin}}\, \mathbb{E}\{C(\theta,a)|y\} = \bar{\theta}$$

This is the **minimum mean squared error** (MMSE) estimator, often denoted $\hat{\theta}_{\mathrm{MMSE}}$.

The $0-1$ loss function



- **Example:** parameter estimation, $\theta \in \Theta = \mathbb{R}^n$.
- Another common choice is the $0-1$ **loss** function

$$C(\theta, a) = -\delta(\theta - a),$$

$\delta(\cdot)$ is the Dirac's delta function.

- **Optimal estimator:** $\mathbb{E}\{C(\theta, a)\big|y\} = \mathbb{E}\{-\delta(\theta - a)\big|y\}$

$$= -\int p(\theta\big|y)\delta(\theta - a)\, d\theta = -p(\theta\big|y)\Big|_{\theta=a}$$

$$\Rightarrow \quad \hat{\theta} = \arg\min_a -p(\theta\big|y)\Big|_{\theta=a}$$

$$= \arg\max_\theta p(\theta\big|y)$$

This is the **maximum a posteriori** (MAP) estimator, often denoted $\hat{\theta}_{MAP}$.

Self-assessment

Suppose $p(\theta|y) = \mathcal{N}(\theta; \bar{\theta}, \mathbf{P})$. The MMSE and MAP estimators are, respectively,

- $\bar{\theta} + \text{tr}\{\mathbf{P}\}$ and $\bar{\theta}$.
- $\bar{\theta}$ and $\mathcal{N}(\theta; \bar{\theta}, \mathbf{P})$.
- $\bar{\theta}$ and $\bar{\theta}$.
- $\bar{\theta} + \text{tr}\{\mathbf{P}\}$ and $\mathcal{N}(\theta; \bar{\theta}, \mathbf{P})$.

Only one statement is correct.