Bayesian Methods of Parameter Estimation

Aciel Eshky
University of Edinburgh
School of Informatics

Introduction

In order to motivate the idea of parameter estimation we need to first understand the notion of mathematical modeling.

What is the idea behind modeling real world phenomena? Mathematically modeling an aspect of the real world enables us to better understand it and better explain it, and perhaps enables us to reproduce it, either on a large scale, or on a simplified scale that characterizes only the critical parts of that phenomenon [1].

How do we model these real life phenomena? These real life phenomena are captured by means of distribution models, which are extracted or learned directly from data gathered about them.

So, what do we mean by parameter estimation? Every distribution model has a set of parameters that need to be estimated. These parameters specify any constants appearing in the model and provide a mechanism for efficient and accurate use of data [2].

Approaches to parameter estimation

Before discussing the Bayesian approach to parameter estimation it is important to understand the classical frequentest approach.

The frequentest approach

The frequentest approach is the classical approach to parameter estimation. It assumes that there is an unknown but objectively fixed parameter θ [3]. It chooses the value of θ which maximizes the likelihood of observed data [4], in other words, making the available data as likely as possible. A common example is the maximum likelihood estimator (MLE).

The frequentest approach is statistically driven, and defines probability as the frequency of successful trials over the number of total trials in an experiment. For example, in a coin toss experiment, we toss the coin 100 times and it comes out 25 times as heads and 75 times as tails. The probabilities are extracted directly from the given data as: $(P = heads) = 1/4$ and $(P = tails) = 3/4$.

Distribution models that use the frequentest approach to estimate their parameters are classified as *generative models* [5], which model the distribution of entire available data, assumed to have been generated with a fixed θ.

The Bayesian approach

In contrast, the Bayesian approach allows probability to represent subjective uncertainty or subjective belief [3]. It fixes the data and instead assumes possible values for θ.

Taking the same coin toss example, the probabilities would represent our subjective belief, rather than the number of successful trials over the total trials. If we believe that heads and tails are equally likely, the probabilities would become: $(P = heads) = 1/2$ and $(P = tails) = 1/2$.

Distribution models that use the Bayesian approach to estimate their parameters are classified as *conditional models*, also known as *discriminative models*, which do not require us to model much of the data and are rather only interested in how particular part of the data depends on the other parts [5].

The Bayesian paradigm

Basics of Bayesian inference

This description is attributed to the following reference [6]. Bayesian inference grows out of the simple formula known as *Bayes rule*. Assume we have two random variables A and B. A principle rule of probability theory known as the *chain rule* allows us to specify the joint probability of A and B taking on particular values a and b, $P(a, b)$, as the product of the conditional probability that A will take on value a given that B has taken on value b, $P(a|b)$, and the marginal probability that B takes on value b, $P(b)$. Which gives us:

Joint probability = Conditional Probability x Marginal Probability

Thus we have:

$$P(a, b) = P(a|b)P(b)$$

There is nothing special about our choice to marginalize B rather than A, and thus equally we have:

$$P(a, b) = P(b|a)P(a)$$

When combining the two we get:

$$P(a|b)P(b) = P(b|a)P(a)$$

rearranged as:

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

and can be equally written in a marginalized form as:

$$P(a|b) = \frac{P(b|a)P(a)}{\Sigma_{a'}^{A} P(b|a')P(a')}$$

This expression is Bayes Rule. Which indicates that we can compute the conditional probability of a variable A given the variable B from the conditional probability of B given A. This introduces the notion of prior and posterior knowledge.

Prior and posterior knowledge

A *prior* probability is the probability available to us beforehand, and before making any additional observations. A *posterior* probability is the probability obtained from the prior probability after making additional observation to the prior knowledge available [6]. In our example, the prior probability would be $P(a)$ and the posterior probability would be $P(a|b)$. The additional observation was observing that B takes on value b.

Utilizing Bayes rule for parameter estimation

Bayes rule obtains its strength from the assumptions we make about the random variables and the meaning of probability [7]. When dealing with parameter estimation, θ could be a parameter needed to be estimated from some given evidence or data d. The probability of data given the parameter is commonly referred to as the *likelihood*. And so, we would be computing the probability of a parameter given the likelihood of some data.

$$P(\theta|d) = \frac{P(d|\theta)P(\theta)}{\Sigma_{\theta'}^{\Theta} P(d|\theta')P(\theta')}$$

Bayesian parameter estimation specify how we should *update* our beliefs in the light of newly introduced evidence.

Summarizing the Bayesian approach

This summary is attributed to the following references [8, 4]. The Bayesian approach to parameter estimation works as follows:

1. Formulate our knowledge about a situation
2. Gather data
3. Obtain posterior knowledge that updates our beliefs

How do we formulate our knowledge about a situation?

a. Define a distribution model which expresses qualitative aspects of our knowledge about the situation. This model will have some unknown parameters, which will be dealt with as random variables [4].

b. Specify a prior probability distribution which expresses our subjective beliefs and subjective uncertainty about the unknown parameters, before seeing the data.

After gathering the data, how do we obtain posterior knowledge?

c. Compute posterior probability distribution which estimates the unknown parameters using the rules of probability and given the observed data, presenting us with updated beliefs.

The problem of visual perception

To illustrate this Bayesian paradigm of parameter estimation, let us apply it to a simple example concerning visual perception. The example is attributed to the following reference [9].

Formulating the Problem

The perception problem is modeled using observed image data, denoted as d. The observable scene properties, denoted as θ, constitute the parameters needed to be estimated for this model. We can define probabilities as follows:

$P(\theta)$ represents the probability distribution of observable scene properties, which are the parameters we need to estimate, or in other words, update in the light of new data. This probability constitutes the *prior probability*.

$P(d|\theta)$ represents the probability distribution of the images given the observable scene properties. This probability constitutes the *likelihood*.

$P(d)$ represents the probability of the images, which are constants that can be normalized over.

$P(\theta|d)$ represents the probability distribution of the observable scene properties given the images. This probability constitute the *posterior probability* of the estimated parameters.

By applying Bayes theorem we arrive at:

$$P(\theta|d) = \frac{P(d|\theta)P(\theta)}{P(d)}$$

And equally:

$$P(\theta|d) = \frac{P(d|\theta)P(\theta)}{\Sigma_{\theta'}^{\Theta} P(d|\theta')P(\theta')}$$

The denominator can be consider as a normalizing constant:

$$P(\theta|d) = k * P(d|\theta)P(\theta)$$

An example

Consider the following problem. Given the silhouette of an object, we need to infer what that object is.

The prior distribution of objects, $P(Object) = P(\theta)$, is:

Object	Probability
Cube	0.3
Cylinder	0.2
Sphere	0.1
Prism	0.4

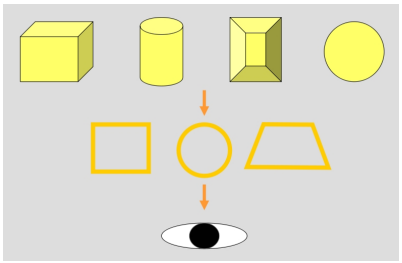

Figure 1: Objects and Silhouette [9]

The likelihood of a silhouette given an object, $P(Silhouette|Object) = P(d|\theta)$, is:

	Cube	Cylinder	Sphere	Prism
Square	1.0	0.6	0.0	0.4
Circle	0.0	0.4	1.0	0.0
Trapezoid	0.0	0.0	0.0	0.6

The normalization constant k is given as 1.85.

The posterior distribution of objects given the silhouettes, $P(Object|Silhouette) = P(\theta|d)$ can then be computed. For example, given $\theta = Square$

$$P(Cube|Square) = k * 0.2 * 1.0 = 0.37$$

$$P(Cylinder|Square) = k * 0.3 * 0.6 = 0.333$$

$$P(Sphere|Square) = k * 0.1 * 0.0 = 0.0$$

$$P(Prism|Square) = k * 0.4 * 0.4 = 0.296$$

And thus we have updated our beliefs in the light of newly introduced data.

References

[1] Amos Storkey. Mlpr lectures: Distributions and models. http://www.inf.ed.ac.uk/teaching/courses/mlpr/lectures/distnsandmodels-print4up.pdf, 2009. School of Informatics, University of Edinburgh.

[2] J.V. Beck and K.J. Arnold. *Parameter estimation in engineering and science. Wiley series in probability and mathematical statistics.* J. Wiley, New York, 1977.

[3] Algorithms for graphical models (agm) bayesian parameter estimation. www-users.cs.york.ac.uk/ jc/teaching/agm/lectures/lect14/lect14.pdf, November 2006. University of York, Department of Computer Science.

[4] Chris Williams. Pmr lectures: Bayesian parameter estimation. http://www.inf.ed.ac.uk/teaching/courses/pmr/slides/bayespe-2x2.pdf, 2008. School of Informatics, University of Edinburgh.

[5] Amos Storkey. Mlpr lectures: Naive bayes and bayesian methods. http://www.inf.ed.ac.uk/teaching/courses/mlpr/lectures/naiveandbayes-print4up.pdf, 2009. School of Informatics, University of Edinburgh.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.

[7] Thomas L. Griffiths, Charles Kemp, and Joshua B. Tenenbaum. Bayesian models of cognition. Technical report.

[8] Radford M. Neal. Bayesian methods for machine learning. www.cs.toronto.edu/pub/radford/bayes-tut.pdf, December 2004. NIPS Tutorial, University of Toronto.

[9] Olivier Aycard and Luis Enrique Sucar. Bayesian techniques in vision and perception. http://homepages.inf.ed.ac.uk/rbf/IAPR.