



Maximum Likelihood and Bayesian Parameter Estimation

Pattern Classification by Richard O. Duda, Peter E. Hart and David G. Stork
Chapter 3 - Part 1

28 November 2018
Luciana Majercsik

Outline

Introduction

Maximum likelihood estimation

Bayesian estimation


Bayesian Parameter Estimation: Gaussian Case

Bayesian Parameter Estimation: General Theory

Introduction



- In the previous presentation (Chapter 2) we saw how we could design an optimal classifier if we knew the prior probabilities $P(\omega_i)$ and the class-conditional densities $p(x|\omega_i)$.
- In pattern recognition applications we rarely (if ever) have this kind of complete knowledge about the probabilistic structure of the problem. In a typical case we merely have some vague, general knowledge about the situation, together with a number of design samples or training data — particular representatives of the patterns we want to classify. The problem, then, is to find some way to use this information to design or train the classifier.

- 
- **Basic Idea:** use the samples to estimate the unknown probabilities and probability densities, and to use the resulting estimates as if they were the true values.
 - **How difficult is the estimation?**
 - The estimation of the prior probabilities - presents *no serious difficulties* .
 - The estimation of the class-conditional densities - *is quite difficult*. The number of available samples always seems too small, and serious problems arise when the dimensionality of the feature vector x is large .
 - **Possible solution:** Make an assumption regarding the form of the conditional densities.
 - Example: maybe we can reasonably assume that $p(x|\omega_i)$ is a normal density with mean μ_i and covariance matrix Σ_i , although we do not know these parameters.

Introduction



- Today's presentation addresses the parametric density estimation case where the *full probability structure* underlying the categories(data) *is not known*, but the general forms of their distributions are known (or assumed a priori) — i.e., the models.
- The uncertainty about a probability distribution is represented by the values of some unknown parameters, and we seek to determine these parameters to attain the best categorization.
- Pattern recognition algorithms are often categorized as *parametric* or *non-parametric* and

Parametric Methods



“A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called a **parametric model**. No matter how much data you throw at a parametric model, it won't change its mind about how many parameters it needs.”

[Artificial Intelligence: A Modern Approach](#)(page 737)

Nonparametric

“**Nonparametric methods** are good when you have a lot of data and no prior knowledge, and when you don't want to worry too much about choosing just the right features.”

[Artificial Intelligence: A Modern Approach](#) (page 757)



Parametric Methods

- Maximum likelihood
- Method of Moments
- Bayesian Estimation

Nonparametric

- Kernel Density Estimation
- Nearest neighbor rule

Advantages



Parametric

- **Simpler:** These methods are easier to understand and interpret results.
- **Speed:** Parametric models are very fast to learn from data.
- **Less Data:** They do not require as much training data and can work well even if the fit to the data is not perfect.

Nonparametric

- **Flexibility:** Capable of fitting a large number of functional forms.
- **Power:** No assumptions (or weak assumptions) about the underlying function.
- **Performance:** Can result in higher performance models for prediction.
- **Easy to apply** and to understand.

Disadvantages



Parametric

- **Constrained:** By choosing a functional form these methods are highly constrained to the specified form.
- **Limited Complexity:** The methods are more suited to simpler problems.
- **Poor Fit:** In practice the methods are unlikely to match the underlying mapping function.

Nonparametric

- **More data:** Require a lot more training data to estimate the mapping function.
- **Slower:** A lot slower to train as they often have far more parameters to train.
- **Overfitting:** More of a risk to overfit the training data and it is harder to explain why specific predictions are made.

Why do we still need both parametric and nonparametric methods?

Many times parametric methods are more efficient than the corresponding nonparametric methods.

Although this difference in efficiency is typically not that much of an issue, there are instances where we do need to consider which method is more efficient.



Parametric methods for density estimation

Maximum likelihood estimation

Method of moments

Bayesian estimation

Desirable properties for estimates



There are *several desirable properties* every good estimator should possess:

- **Consistency** : An estimator is consistent if the estimate it constructs is guaranteed to converge to the true parameter value θ as the quantity of data to which it is applied increases. If $\hat{\theta}_n$ represents the estimator based on n observations, then:

$$\hat{\theta}_n \xrightarrow{n \rightarrow \infty} \theta$$

- **Bias**: An estimator of a parameter is unbiased if the expected value of the estimate is the same as the true value of the parameters:

$$E[\hat{\theta}] = \theta$$

Desirable properties for estimates

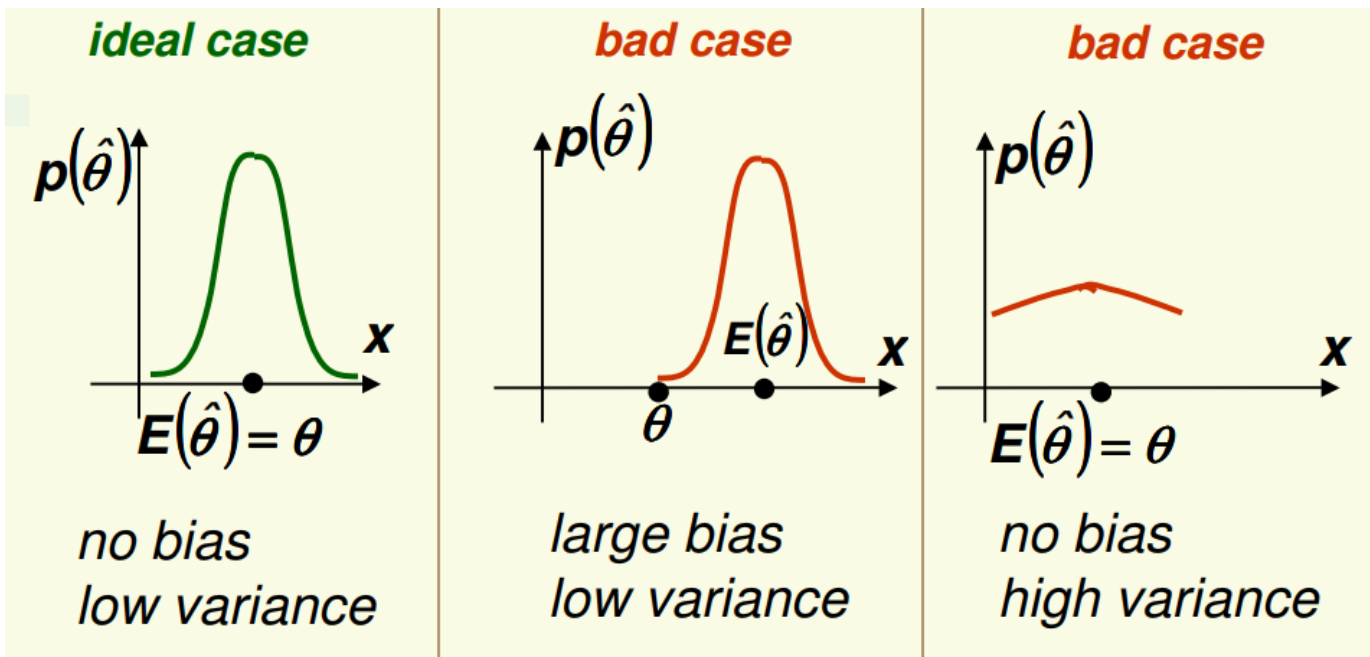


→ **Variance:** the variance measures how much the estimate changes for different datasets.


$$\text{var}(\theta) = E[(\hat{\theta} - E[\hat{\theta}])^2]$$

→ **Efficiency:** if we have two unbiased estimates, we prefer the estimate with a smaller variance because this means it's more precise in statistical terms. It's also important to note that the property of efficiency only applies in the presence of unbiasedness since we only consider the variances of unbiased estimates.

Desirable properties for estimates



Maximum Likelihood Estimation (MLE)



The principle of Maximum Likelihood Estimation was pioneered by Fisher in the 1920s and has a wide range of applications.

It is an indispensable tool for many statistical modeling techniques, and in particular, in non-linear modeling with non-normal data.

MLEs estimates are known for having both good computational and statistical properties, but they work with the premise that the *parameters being estimated does not change with time (they are FIXED)*, i.e., the distribution is assumed to be stationary.

MLE often can be simpler than alternate methods, such as Bayesian techniques or other methods presented in subsequent chapters.

General principle



Likelihood = a basic measure of the quality of a set of predictions with respect to observed data. In the context of parameter estimation, the likelihood is defined as the joint probability of a set of observations, conditioned on a choice for θ , and naturally is viewed as a function of the parameters θ to be estimated :

$$\text{Like}(\theta; D) \equiv p(D|\theta) = p(x_1, x_2, \dots, x_n | \theta)$$

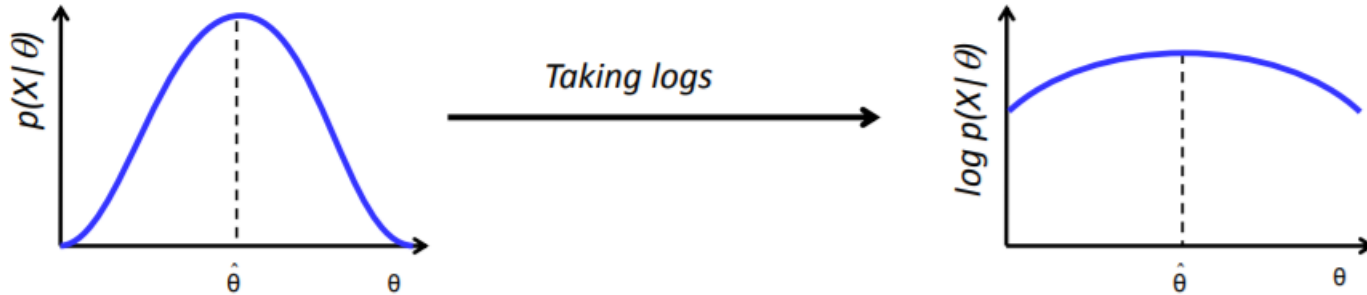
Since good predictions are better, a natural approach to parameter estimation is to ***choose the set of parameter values that yields the best predictions***—that is, the parameter that maximizes the likelihood of the observed data. This value is called the maximum likelihood estimate (MLE), defined formally as:

$$\theta_{\text{MLE}} = \theta = \arg \max \text{Like}(\theta; D)$$

General principle

For easier computations we will work (when convenient) with the log likelihood . Because the logarithm is a monotonic function, then:

$$\hat{\theta} = \arg \max p(D | \theta) = \arg \max \left[\ln \left(p(D | \theta) \right) \right]$$



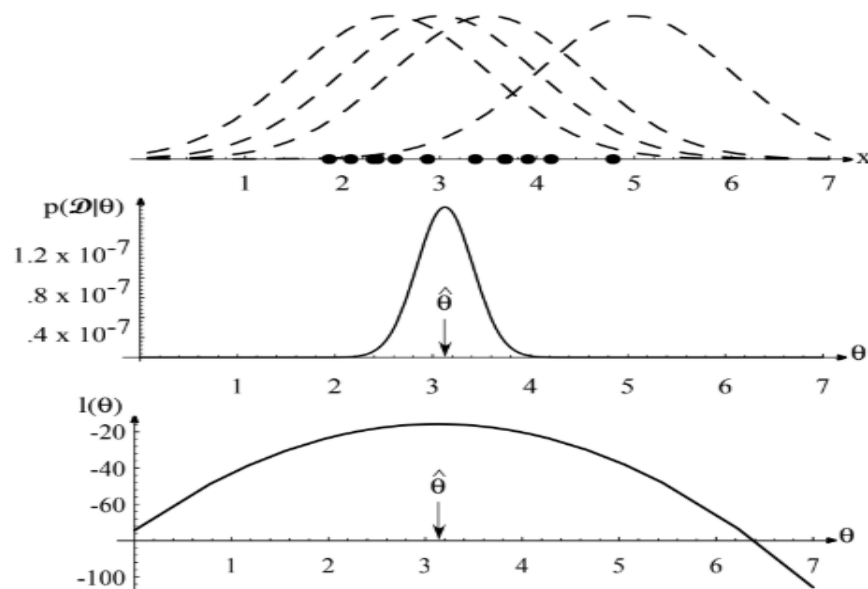


Figure 3.1: The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figures shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood — i.e., the log-likelihood $l(\theta)$, shown at the bottom. Note especially that the likelihood lies in a different space from $p(x|\hat{\theta})$, and the two can have different functional forms.

General principle



Assume we have dataset $D = \{x_1, x_2, \dots, x_n\}$ with the variables x_k drawn independently from the distribution $p(x|\theta)$ (an i.i.d. set).

The independence property allows us to write the joint probability distribution as:

$$p(D | \theta) = p(x_1, x_2, \dots, x_n | \theta) = \prod_{k=1}^n p(x_k | \theta)$$

Hence, the ML estimate of θ can be written as:

$$\hat{\theta} = \arg \max \left[\ln \prod_{k=1}^n p(x_k | \theta) \right] = \arg \max \sum_{k=1}^n \ln (p(x_k | \theta))$$

This simplifies the problem, since now we have to maximize a sum of terms rather than a product.

A General MLE strategy

- Suppose $\theta = (\theta_1, \theta_2, \dots, \theta_m)$ is a vector of parameters.
- Task: Find MLE θ assuming known form for $p(D | \theta)$

1. Write $\ln p(D | \theta)$

2. Compute $\partial \ln(p)/\partial \theta_k$ for any $k \in \{1, 2, \dots, m\}$

3. Solve the set of simultaneous equations

$$\frac{\partial \ln(p)}{\partial \theta_1} = 0$$

$$\frac{\partial \ln(p)}{\partial \theta_2} = 0$$

$$\vdots$$

$$\frac{\partial \ln(p)}{\partial \theta_m} = 0$$

4. Check that you're at a maximum

Examples

MLE for univariate Gaussian - unknown mean



- Suppose you have $x_1, x_2, \dots, x_n \sim (\text{i.i.d}) \mathcal{N}(\mu, \sigma^2)$
- But you don't know μ (you do know σ^2)
- MLE: For which μ is x_1, x_2, \dots, x_n most likely?

$$\begin{aligned}
\theta = \mu &\Rightarrow \hat{\theta} = \arg \max_{\theta} \ln p(x_1, x_2, \dots, x_n | \theta) \\
&= \arg \max_{\theta} \ln \left(\prod_{k=1}^n p(x_k | \theta) \right) = \arg \max_{\theta} \sum_{k=1}^n \ln p(x_k | \theta) \\
&= \arg \max_{\theta = \mu} \sum_{k=1}^n \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{1}{2\sigma^2} (x_k - \mu)^2 \right) \right) = \\
&= \arg \max_{\mu} \sum_{k=1}^n \left[\ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} (x_k - \mu)^2 \right]
\end{aligned}$$

The maximum of a function are defined by the zeros of its derivative

$$\frac{\partial}{\partial \mu} \sum_{k=1}^n \left[\ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} (x_k - \mu)^2 \right] = 0$$

The former eq. is equivalent to

$$\sum_{k=1}^n \frac{1}{\sigma^2} (x_k - \hat{\mu}) = 0$$

By solving the equation, we get the ML estimate of the mean

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n x_k$$

Interpretation: the maximum likelihood estimate for the unknown population mean is just the arithmetic average of the training samples (a very intuitive result)

Geometrically, if we think of the n samples as a cloud of points, the sample mean is the centroid of the cloud.

The sample mean has a number of desirable statistical properties as well, and one would be inclined to use this rather obvious estimate even without knowing that it is the maximum likelihood solution

MLE for univariate Gaussian - unknown mean and std

- Suppose you have $x_1, x_2, \dots, x_n \sim (\text{i.i.d}) \mathcal{N}(\mu, \sigma^2)$
- But you don't know neither μ nor σ^2
- MLE: For which $\theta = (\mu, \sigma^2)$ is x_1, x_2, \dots, x_n most likely?

As in the previous case we have:

$$\begin{aligned}\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} &\Rightarrow \hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} \ln p(x_1, x_2, \dots, x_n | \boldsymbol{\theta}) \\ &= \arg \max_{(\mu, \sigma^2)} \sum_{k=1}^n \left[\ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} (x_k - \mu)^2 \right]\end{aligned}$$

But the derivative becomes a gradient, since we have 2 variables:

$$\nabla_{\boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \sum_{k=1}^n \ln p(x_k | \boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_2} \sum_{k=1}^n \ln p(x_k | \boldsymbol{\theta}) \end{bmatrix} = \sum_{k=1}^n \begin{bmatrix} \frac{\partial}{\partial \mu} \ln p(x_k | \boldsymbol{\theta}) \\ \frac{\partial}{\partial \sigma^2} \ln p(x_k | \boldsymbol{\theta}) \end{bmatrix} = \sum_{k=1}^n \begin{bmatrix} \frac{1}{\sigma^2} (x_k - \mu) \\ -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} (x_k - \mu)^2 \end{bmatrix}$$

Solving the system of equations:

$$\nabla_{\boldsymbol{\theta}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^n x_k \quad ; \quad \sigma^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \hat{\mu})^2$$

Therefore, the ML estimate of the variance is the sample variance of the dataset, again a very intuitive result

Properties of the estimates



The bias of the ML estimate of the mean is:

$$E[\hat{\mu}] - \mu = E\left[\frac{1}{n} \sum_{k=1}^n x_k\right] - \mu = \mu - \mu = 0$$

Therefore, the mean is an unbiased estimate.

The bias of the ML estimate of the variance is:

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n} \sum_{k=1}^n (x_k - \hat{\mu})^2\right] = \frac{n-1}{n} \sigma^2 \neq \sigma^2$$

Thus the estimate of the variance is BIASED. This is because the ML estimate of variance uses μ instead of $\hat{\mu}$

Properties of the estimates



Question: How “bad” is this bias?

We can see that:

a) for $N \rightarrow \infty$ the bias becomes zero asymptotically

b) The bias is only noticeable when we have very few samples, in which case we should not be doing statistics in the first place!

c) We can choose:
$$\hat{\sigma}_{UNBIASED}^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \hat{\mu})^2$$

Variance for MLE mean and STD

- The variance of the ML estimate of the mean is

$$\begin{aligned} E[(\hat{\mu} - \mu)^2] &= E\left[\frac{1}{n} \sum_{k=1}^n x_k - \mu\right]^2 = E\left[\frac{1}{n} \sum_{k=1}^n (x_k - \mu)\right]^2 = \\ &= \frac{1}{n^2} E\left[\sum_{j=1}^n \sum_{k=1}^n (x_j - \mu)(x_k - \mu)\right] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

Thus, for a large number of samples, the variance is small which means that the ML estimate of the mean **is a very good estimator**.

Similarly it can be shown that the variance of ML estimate of the variance goes to 0 as n goes to infinity. So, the ML estimate of the variance **is a good estimator**

The multivariate Gaussian Case



In a similar manner (but with considerable computations), it can be shown that the ML estimates for the multivariate Gaussian are the sample mean vector and sample covariance matrix

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T$$

The exponential distribution



Suppose that the lifetime of some brand of light bulbs is modeled by an exponential distribution with (unknown) parameter λ . We test 5 bulbs and find they have lifetimes of 2, 3, 1, 3, and 4 years, respectively. What is the MLE for λ ?

Let X_i be the lifetime of the i -th bulb and let x_i be the value X_i takes. Then each X_i has exponential pdf

$$p(x_i | \lambda) = \lambda e^{-\lambda x_i}$$

We assume the lifetimes of the bulbs are independent, so the joint pdf is the product of the individual densities:

$$\begin{aligned} p(x_1, \dots, x_n | \lambda) &= \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \cdot \lambda e^{-\lambda x_3} \cdot \lambda e^{-\lambda x_4} \cdot \lambda e^{-\lambda x_5} = \\ &= \lambda^5 e^{-\lambda (\sum_{i=1}^5 x_i)} \end{aligned}$$

The exponential distribution



Replacing the values of the variables, the likelihood function is:

$$p(2,3,1,3,4 | \lambda) = \lambda^5 e^{-13\lambda}$$

That means that the log likelihood function is:

$$\ln p(2,3,1,3,4 | \lambda) = 5 \ln \lambda - 13\lambda$$

The derivative of this function is

$$\frac{\partial}{\partial \lambda} \ln p(2,3,1,3,4 | \lambda) = \frac{5}{\lambda} - 13 = 0 \Rightarrow \hat{\lambda} = \frac{5}{13}$$

The MLE for λ turned out to be the reciprocal of the sample mean.

The uniform distribution



Suppose our data x_1, x_2, \dots, x_n are independently drawn from a uniform distribution $U[0, \theta]$. We want to find the MLE estimate for θ . The density for a uniform distribution is:

$$p(x) = \begin{cases} \frac{1}{\theta}, & x \in [0, \theta] \\ 0, & \text{otherwise} \end{cases}$$

Therefore the likelihood function is

$$p(x_1, \dots, x_n \mid \theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n, & x_k \in [0, \theta] \quad \forall k = 1, \dots, n \\ 0, & \text{otherwise} \end{cases}$$

The uniform distribution



Consequently, the joint density is 0 whenever any of the $x_k > \theta$.


Restating this in terms of likelihood: no value of θ is possible that is less than any of the x_k . Consequently, any value of θ less than any of the x_k has likelihood 0. If

$$x_{\max} = \max(x_1, \dots, x_n)$$

The likelihood is 0 on the interval $(0, x_{\max})$ and is positive and decreasing on the interval $[x_{\max}, \infty)$. Thus, to maximize it, we should take the minimum value of θ on this interval:

$$\hat{\theta} = \max(x_1, \dots, x_n)$$

Bayesian Estimation (Learning)

- 
- While in the maximum likelihood method the true parameter vector we seek, θ , is considered to be fixed, in **Bayesian learning** we consider θ to be a **RANDOM VARIABLE**, and training data allows us to convert a distribution on this variable into a posterior probability density.
 - The answers we get by this method will generally be nearly identical to those obtained by maximum likelihood.
 - At the heart of Bayesian classification lies the computation of the posterior probabilities $P(\omega_i|x)$. Bayes' formula allows us to compute these probabilities from the prior probabilities $P(\omega_i)$ and the class-conditional densities $p(x|\omega_i)$.
 - **Problem:** how can we proceed when these quantities are unknown?



→ **Answer:** The best we can do is to compute $P(\omega_i|x)$ using *all available information*.

→ Available information might be:

- prior knowledge, such as knowledge of the functional forms for unknown densities and ranges for the values of unknown parameters

- a set of training samples

→ Let D denote the set of samples. We can emphasize the role of the samples by saying that our goal is to compute the posterior probabilities $P(\omega_i|x, D)$. From these probabilities we can obtain the Bayes classifier.



Given the sample D , Bayes's formula gives us:

$$p(\omega_i | x, D) = \frac{p(x | \omega_i, D_i) P(\omega_i)}{\sum_{j=1}^c p(x | \omega_j, D_j) P(\omega_j)}$$

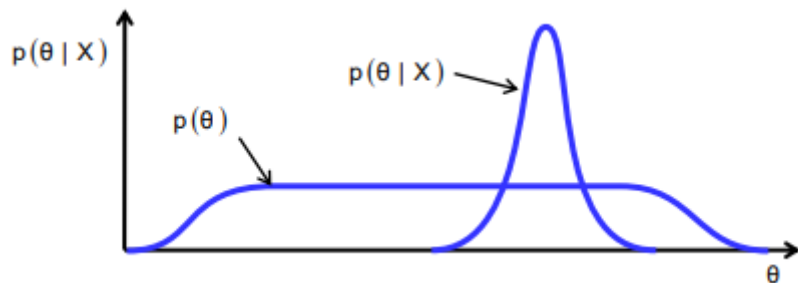
Because each class can be treated independently, we have c separate problems of the following form: use a set D of samples drawn independently according to the fixed but unknown probability distribution $p(x)$ to determine $p(x|D)$. This is the central problem of Bayesian learning.



Although the desired probability density $p(x)$ is unknown, we assume that it has a known parametric form. The only thing assumed unknown is the value of a parameter vector θ .

We shall express the fact that $p(x)$ is unknown but has known parametric form by saying that the function $p(x|\theta)$ is completely known.

Any information we might have about θ prior to observing the samples is assumed to be contained in a known prior density $p(\theta)$. And the observation of the samples converts this to a posterior density $p(\theta|D)$, which, we hope, is sharply peaked about the true value of θ .





Basic goal: is to compute $p(x|D)$, which is as close as we can come to obtaining the unknown $p(x)$.

How ?: By integrating the joint density $p(x, \theta | D)$ over θ .

$$p(x | D) = \int p(x, \theta | D) d\theta$$

We can write the equation in the equivalent form

$$p(x | D) = \int p(x | \theta, D) p(\theta | D) d\theta = \int p(x | \theta) p(\theta | D) d\theta$$

Since the selection of x and that of the training samples in D is done independently, the first factor is merely $p(x|\theta)$.

That is, the distribution of x is known completely once we know the value of the parameter vector.
This key equation links the desired class-conditional density $p(x|D)$ to the posterior density $p(\theta|D)$ for the unknown parameter vector.

Bayesian inference for the Gaussian




Assume a univariate density where our random variable x is generated from a normal distribution $N(\mu, \sigma^2)$ with known standard deviation .

Goal: find the mean μ of the distribution given some i.i.d. data points $D = \{x_1, x_2, \dots, x_n\}$.

To capture our knowledge about $\theta = \mu$, we assume that it also follows a normal density $N(\mu_0, \sigma_0^2)$


$$p_0(\theta) = p_0(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(\theta - \mu_0)^2\right)$$

We use Bayes rule to develop an expression for the posterior $p(\theta|D)$


$$\begin{aligned} p(\theta | D) &= \frac{p(D | \theta) p(\theta)}{p(D)} = \frac{p(x_1, x_2, \dots, x_n | \theta) p(\theta)}{p(D)} = \frac{p_0(\theta)}{p(D)} \left(\prod_{k=1}^n p(x_k | \theta) \right) = \\ &= \frac{1}{p(D)} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2} (\theta - \mu_0)^2\right) \prod_{k=1}^n \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (x_k - \theta)^2\right) \right) \end{aligned}$$

To understand **how Bayesian estimation changes the posterior** as more data becomes available, we will find the maximum of $p(\theta|X)$. The technique is the same with what we used in MLE.

We compute the partial derivative of the $\log p(\theta|X)$ with respect to $\theta = \mu$.



$$\frac{\partial}{\partial \theta} \ln p(\theta | D) = 0 \Rightarrow \frac{\partial}{\partial \mu} \left[-\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 - \sum_{k=1}^n \frac{1}{2\sigma^2} (x_k - \mu)^2 \right] = 0$$

After computing the derivative and doing some algebraic computation, we obtain for the estimate of the mean based on n available observations:

$$\mu_n = \underbrace{\frac{\sigma^2}{\sigma^2 + n\sigma_0^2} \mu_0}_{PRIOR} + \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2} \underbrace{\frac{1}{n} \sum_{k=1}^n x_k}_{ML}$$

Remark: 1. the mean of the posterior distribution given by the above formula is a compromise between the prior mean μ_0 and the maximum likelihood solution μ_{ML} .

2. If the number of observed data points $n=0$, then μ_n reduces to the prior mean as expected. For $n \rightarrow \infty$, the posterior mean is given by the maximum likelihood solution.



Similarly, the standard deviation can be found to be:

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$$

Remarks:

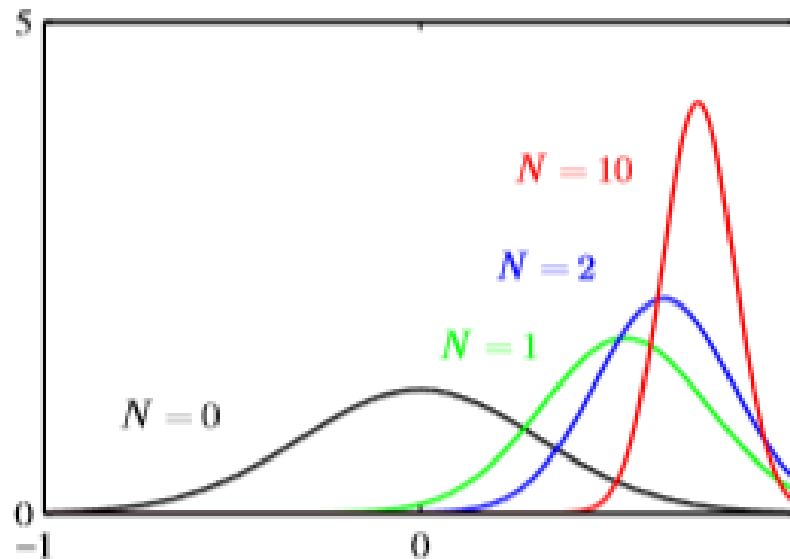
1. We see that this expression is given in terms of the inverse variance, which is called the **precision**. Furthermore, the precisions are additive, so that the precision of the posterior is given by the precision of the prior plus one contribution of the data precision from each of the observed data points. As we increase the number of observed data points, the precision steadily increases, corresponding to a posterior distribution with steadily decreasing variance.
2. With no observed data points, we have the prior variance, whereas if the number of data points $n \rightarrow \infty$, the variance σ_n^2 goes to zero and the posterior distribution becomes infinitely peaked around the maximum likelihood solution.

Illustration of Bayesian inference for the mean μ of a Gaussian distribution (Bishop)

Assume that the data points are generated from a Gaussian of mean 0.8 and variance 0.1

In reality we would not know the true mean; we are just making estimations.

We generate a number of examples from this distribution. To capture our lack of knowledge about the mean, we assume a normal prior $p_0(\theta_0)$, with $\mu_0 = 0.0$ and $\sigma_0 = 0.1$



The figure shows the posterior $p(\mu|D)$. As N increases, the estimate μ_N approaches its true value ($\mu = 0.8$) and the spread (or uncertainty in the estimate) decreases.

The curves show the prior distribution over μ (the curve labelled $N=0$), which in this case is itself Gaussian, along with the posterior distribution given by for increasing numbers N of data points.

MLE versus Bayesian Parameter Estimation

- ❖ The maximum likelihood method seeks to find the parameter value that is best supported by the training data, i.e., maximizes the probability of obtaining the samples actually observed. (In practice, for computational simplicity one typically uses log-likelihood.)
- ❖ In Bayesian estimation the parameters are considered random variables having a known a priori density; the training data convert this to an a posteriori density.
- ❖ The Bayesian estimate will approximate the ML solution, and when $n \rightarrow \infty$ the Bayesian estimate of $p(x)$ will approach the ML solution.
- ❖ In practice, only when we have a limited number of observations will the two approaches yield different results

MLE versus Bayesian Parameter Estimation

- ❖ **Computational complexity** : the maximum likelihood methods are often to be preferred since they require merely differential calculus techniques or gradient search for θ , rather than a possibly complex multidimensional integration needed in Bayesian estimation.
- ❖ **Interpretability**: In many cases the maximum likelihood solution will be easier to interpret and understand since it returns the single best model from the set the designer provided. In contrast Bayesian methods give a weighted average of models (parameters), often leading to solutions more complicated and harder to understand. The Bayesian approach reflects the remaining uncertainty in the possible models.
- ❖ **Classification**: When designing a classifier by either of these methods, we determine the posterior densities for each category, and classify a test point by *the maximum posterior*. (If there are costs, summarized in a cost matrix, these can be incorporated as well)

Sources of Classification Error

There *are three sources of classification error* :

Bayes or indistinguishability error: the error due to overlapping densities $p(x|\omega_i)$ for different values of i . This error is an inherent property of the problem and can never be eliminated.

Model error: the error due to having an incorrect model. This error can only be eliminated if the designer specifies a model that includes the true model which generated the data. Designers generally choose the **model based on knowledge of the problem domain** rather than on the subsequent estimation method, and thus the model error in maximum likelihood and Bayes methods rarely differ.

Estimation error: the error arising from the fact that the parameters are estimated from a finite sample. This error can best be reduced by increasing the training data.