Design of the classifier in a pattern recognition system

The approach to be followed builds upon probabilistic arguments stemming from the statistical nature of the generated features.

This is due to the statistical variation of the patterns as well as to the noise in the measuring sensors.

Adopting this reasoning as our kickoff point, we will design classifiers that classify an unknown pattern in the most probable of the classes.

Thus, our task now becomes that of defining what "most probable" means.

Given a classification task of M classes, $\omega_1, \omega_2, \ldots, \omega_M$, and an unknown pattern, which is represented by a feature vector $\mathbf{x}$, we form the M conditional probabilities $P(\omega_i|\mathbf{x})$, i =1, 2, . . . , M.

Sometimes, these are also referred to as ***a posteriori probabilities***.

In words, each of them represents the probability that the unknown pattern belongs to the respective class ω_i, given that the corresponding feature vector takes the value $\boldsymbol{x}$.

Who could then argue that these conditional probabilities are not sensible choices to quantify the term *most probable*?

Indeed, the classifiers to be considered in this chapter compute either the maximum of these *M* values or, equivalently, the maximum of an appropriately defined function of them.

The unknown pattern is then assigned to the class corresponding to this maximum.

The first task we are faced with is the computation of the conditional probabilities.

The Bayes rule will once more prove its usefulness!

A major effort will be devoted to techniques for estimating probability density functions (pdf), based on the available experimental evidence, that is, the feature vectors corresponding to the patterns of the training set.

Bayes Decision Theory

We will initially focus on the two-class case. Let ω_1, ω_2 be the two classes in which our patterns belong.

In the sequel, we assume that the *a priori probabilities* $P(\omega_1)$, $P(\omega_2)$ are known.

This is a very reasonable assumption, because even if they are not known, they can easily be estimated from the available training feature vectors.

Indeed, if N is the total number of available training patterns, and N_1, N_2 of them belong to ω_1 and ω_2, respectively, then $P(\omega_1) \approx N_1/N$ and $P(\omega_2) \approx N_2/N$.

The other statistical quantities assumed to be known are the class-conditional probability density functions $p(\boldsymbol{x}|\omega_i)$, i=1, 2, describing the distribution of the feature vectors in each of the classes.

If these are not known, they can also be estimated from the available training data.

The pdf $p(\boldsymbol{x}|\omega_i)$ is sometimes referred to as the *likelihood function of* ω_i *with respect to* $\boldsymbol{x}$.

Here we should stress the fact that an implicit assumption has been made.

That is, the feature vectors can take any value in the l-dimensional feature space.

In the case that feature vectors can take only discrete values, density functions $p(\boldsymbol{x}|\omega_i)$ become probabilities and will be denoted by $P(\boldsymbol{x}|\omega_i)$.

We now have all the ingredients to compute our conditional probabilities, as stated in the introduction. To this end, let us recall from our probability course basics the *Bayes rule.*

$$P(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{p(\boldsymbol{x})}$$

where $p(\boldsymbol{x})$ is the pdf of $\boldsymbol{x}$ and for which we have:

$$p(\boldsymbol{x}) = \sum_{i=1}^{2} p(\boldsymbol{x}|\omega_i)P(\omega_i)$$

The *Bayes classification rule* can now be stated as

$$\begin{aligned} &\text{If} \quad P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x}), \quad \boldsymbol{x} \text{ is classified to } \omega_1 \\ &\text{If} \quad P(\omega_1|\boldsymbol{x}) < P(\omega_2|\boldsymbol{x}), \quad \boldsymbol{x} \text{ is classified to } \omega_2 \end{aligned}$$

The case of equality is detrimental and the pattern can be assigned to either of the two classes.

The decision can equivalently be based on the inequalities:

$$p(\boldsymbol{x}|\omega_1)P(\omega_1) \gtrless p(\boldsymbol{x}|\omega_2)P(\omega_2)$$

$p(\boldsymbol{x})$ is not taken into account, because it is the same for all classes and it does not affect the decision.

Furthermore, if the *a priori* probabilities are equal, that is, $P(\omega 1) = P(\omega 2) = 1/2$, then the above equation becomes:

$$p(\boldsymbol{x}|\omega_1) \gtrless p(\boldsymbol{x}|\omega_2)$$

Thus, the search for the maximum now rests on the values of the conditional pdfs evaluated at $\boldsymbol{x}$.

Figure below presents an example of two equiprobable classes and shows the variations of $p(x|\omega_i)$, $i=1, 2$, as functions of x for the simple case of a single feature ($l=1$).

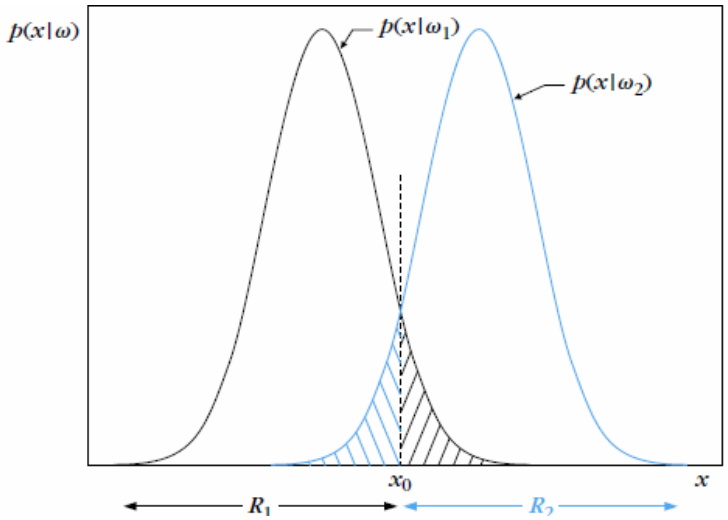

Figure: Example of the two regions *R*1 and *R*2 formed by the Bayesian classifier for the case of two equiprobable classes.

The dotted line at x_0 is a threshold partitioning the feature space into two regions, R_1 and R_2.

According to the Bayes decision rule, for all values of x in R_1 the classifier decides ω_1 and for all values in R_2 it decides ω_2.

However, it is obvious from the figure that decision errors are unavoidable.

Indeed, there is a finite probability for an x to lie in the R_2 region and at the same time to belong in class ω_1. Then our decision is in error. The same is true for points originating from class ω_2.

It does not take much thought to see that the total probability, P_e, of committing a decision error for the case of two equiprobable classes, is given by:

$$P_e = \frac{1}{2}\int_{-\infty}^{x_0} p(x|\omega_2)\,dx + \frac{1}{2}\int_{x_0}^{+\infty} p(x|\omega_1)\,dx$$

which is equal to the total shaded area under the curves in figure above.

We have now touched on a very important issue.

Our starting point to arrive at the Bayes classification rule was rather empirical, via our interpretation of the term *most probable*.

We will now see that this classification test, though simple in its formulation, has a sounder mathematical interpretation.

Minimizing the Classification Error Probability

We will show that *the Bayesian classifier is optimal with respect to minimizing the classification error probability*.

Indeed, the reader can easily verify, as an exercise, that moving the threshold away from x_0, in the above figure, always increases the corresponding shaded area under the curves. Let us now proceed with a more formal proof.

Proof: Let R_1 be the region of the feature space in which we decide in favor of ω_1 and R_2 be the corresponding region for ω_2. Then an error is made if $\boldsymbol{x} \in R_1$, although it belongs to ω_2 or if $\boldsymbol{x} \in R_2$, although it belongs to ω_1. That is,

$$P_e = P(\boldsymbol{x} \in R_2, \omega_1) + P(\boldsymbol{x} \in R_1, \omega_2)$$

where $P(\cdot, \cdot)$ is the joint probability of two events. Recalling, once more, our probability basics, this becomes:

$$\begin{aligned} P_e &= P(\boldsymbol{x} \in R_2|\omega_1)P(\omega_1) + P(\boldsymbol{x} \in R_1|\omega_2)P(\omega_2) \\ &= P(\omega_1)\int_{R_2} p(\boldsymbol{x}|\omega_1)\,d\boldsymbol{x} + P(\omega_2)\int_{R_1} p(\boldsymbol{x}|\omega_2)\,d\boldsymbol{x} \end{aligned}$$

or using the Bayes rule

$$P_e = \int_{R_2} P(\omega_1|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x} + \int_{R_1} P(\omega_2|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x}$$

It is now easy to see that the error is minimized if the partitioning regions R_1 and R_2 of the feature space are chosen so that:

$$R_1: P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x})$$
$$R_2: P(\omega_2|\boldsymbol{x}) > P(\omega_1|\boldsymbol{x})$$

Indeed, since the union of the regions R_1, R_2 covers all the space, from the definition of a probability density function we have that:

$$\int_{R_1} P(\omega_1|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x} + \int_{R_2} P(\omega_1|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x} = P(\omega_1)$$

Combining the above equations, we get:

$$P_e = P(\omega_1) - \int_{R_1} (P(\omega_1|\boldsymbol{x}) - P(\omega_2|\boldsymbol{x}))\,p(\boldsymbol{x})\,d\boldsymbol{x}$$

This suggests that the probability of error is minimized if R_1 is the region of space in which $P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x})$. Then, R_2 becomes the region where the reverse is true.

So far, we have dealt with the simple case of two classes.

Generalizations to the multiclass case are straightforward. In a classification task with M classes, $\omega_1, \omega_2, \ldots, \omega_M$, an unknown pattern, represented by the feature vector $\boldsymbol{x}$, is assigned to class ω_i if:

$$P(\omega_i|\boldsymbol{x}) > P(\omega_j|\boldsymbol{x}) \quad \forall j \neq i$$

It turns out that such a choice also minimizes the classification error probability.

Discriminant Functions and Decision Surfaces

It is by now clear that minimizing either the risk or the error probability or the Neyman-Pearson criterion is equivalent to partitioning the feature space into M regions, for a task with M classes.

If regions R_i, R_j happen to be contiguous, then they are separated by a ***Decision Surface*** in the multidimensional feature space.

For the minimum error probability case, this is described by the equation:

$$P(\omega_i|\boldsymbol{x}) - P(\omega_j|\boldsymbol{x}) = 0$$

From the one side of the surface this difference is positive, and from the other it is negative.

Sometimes, instead of working directly with probabilities (or risk functions), it may be more convenient, from a mathematical point of view, to work with an equivalent function of them, for example, $g_i(\boldsymbol{x}) \equiv f(P(i|\boldsymbol{x}))$,where $f(\cdot)$ is a monotonically increasing function.

$g_i(\boldsymbol{x})$ is known as a *discriminant function*. The decision test is now stated as

$$\text{classify } \boldsymbol{x} \text{ in } \omega_i \quad \text{if} \quad g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}) \quad \forall j \neq i$$

The decision surfaces, separating contiguous regions, are described by:

$$g_{ij}(\boldsymbol{x}) \equiv g_i(\boldsymbol{x}) - g_j(\boldsymbol{x}) = 0, \quad i,j = 1,2,\ldots,M, \quad i \neq j$$

So far, we have approached the classification problem via Bayesian probabilistic arguments and the goal was to minimize the classification error probability or the risk.

However, as we will soon see, not all problems are well suited to such approaches.

For example, in many cases the involved pdfs are complicated and their estimation is not an easy task.

In such cases, it may be preferable to compute decision surfaces *directly by means of alternative costs*.

Such approaches give rise to discriminant functions and decision surfaces, which are entities with no (necessary) relation to Bayesian classification, and they are, in general, suboptimal with respect to Bayesian classifiers.

In the following we will focus on a particular family of decision surfaces associated with the Bayesian classification for the specific case of Gaussian density functions.

Generalized Linear Models (GLM)

In statistics, a GLM is a flexible generalization of ordinary LR that allows for response varaibles that have error distribution models other than a normal distribution.

The GLM generalizes LR by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of the predicted value.

These models were formulated by John Nelder and Robert Wedder bur as a way of unifying various other statistical models, including LR, Logistic Regerssion and Poisson Regression.

They proposed an iteratively reweighted least squares method for maximum likelihood estimation of the model parameters.

Maximum likelihood estimation remains popular and is the default method on many statistical computing packages.

Ordinary LR predicts the expected value of a given unknown quantity (the response variable, a random variable) as a linear combination of a set of observed values (predictors) .

This implies that a constant change in a predictor leads to a constant change in the response variable (i.e. a linear response model).

This is appropriate when the response variable has a normal distribution (intuitively, when a response variable can vary essentially indefinitely in either direction with no fixed zero value or more generally for any quantity that only varies by a relatively small amount e.g. human heights).

However these assumptions are inappropriate for some types of response variables.
Example: In cases where the response variable is expected to be always positive and varying over a wide range, constant input changes lead to geometrically varying rather than constantly varying, output changes.

As an **example**, a prediction model might predict that 10 degrees temperature decrease would lead to 1000 fewer people visiting the beach is unlikely to generalize well over both small beaches (e.g. those where the expected attendance was 50 at a particular temperature) and large beaches (e.g. those where the expected attendance was 10,000 at a low temperature).

The problem with this kind of prediction model would imply a temperature drop of 10 degrees would lead to 1000 fewer people visiting the beach, a beach whose expected attendance was 50 at a higher temperature would now be predicted to have an impossible attendance of -950.

Logically, a more realistic model would instead predict a constant rate of increased beach attendance (e.g. an increase in 10 degrees leads to a doubling in beach attendance, and a drop in 10 degrees leads to halving the attendance).

Such a model is termed an Exponential Response Model (or log-linear model, since the logarithm of the response is predicted to vary linearly).

Similarly a model that predicts the probability of making a yes/no choice (a Bernoulli Variable) is even less suitable as a linear response model, since the probabilities are bounded on both ends (they must between 0 and 1).

Example: Imagine a model that predicts the likelihood of a given person going to the beach as a function of temperature. A reasonable model might predict, for example, that a change in 10 degrees makes a person two times more or less double the probability value(e.ge. 50% becomes 100%, 75% becomes 150%, etc).

Rather it is the odds that are doubling: from 2:1 odds, to 4:1 odds, to 8: 1 odds, etc. Such a model is called the log-odds model.

In a GLM, each outcome **Y** of the dependent variables is assumed to be generated from a particular distribution in the exponential family, a large range of probability distributions that includes the Normal, Binomial, Poisson and Gamma distributions, among others.

The mean μ of the distribution depends on the independent variables, **X,** through:

$$\mathbf{E(Y)} = \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$$

Where E(**Y**) is the expected value of Y, **Xβ** is the Linear Predictor, a linear combination of unknown parameters **β**, **g** is the Link function.

In this framework, the variance is typically a function, **V** of the mean:

$$\mathbf{Var(Y)} = \mathbf{V}(\boldsymbol{\mu}) = \mathbf{V}(g^{-1}(\mathbf{X}\boldsymbol{\beta})).$$

It is convenient if **V** follows from the exponential family distribution, but it may simply be that the variance is a function of the predicted value.

The unknown parameters, **β**, are typically estimated with maximum-likelihood, maximum quasi-likelihood, or Bayesian techniques.

Model Components

The GLM consists of 3 elements:

1. A probability distribution from the exponential family.
2. A linear predictor $\eta = \mathbf{X}\boldsymbol{\beta}$.
3. A link function g such that $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu} = g^{-1}(\eta)$.

The **overdispersed exponential family** of distributions is a generalization of the exponential family and exponential dispersion model of distributions and includes those probability distributions, parameterized by θ and τ, whose density functions *f* (or probability mass function, for the case of a discrete distribution) can be expressed in the form:

$$f_Y(\mathbf{y} \mid \boldsymbol{\theta}, \tau) = h(\mathbf{y}, \tau) \exp\left(\frac{\mathbf{b}(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{T}(y) - A(\boldsymbol{\theta})}{d(\tau)}\right)$$

T, called the ***Dispersion Parameter***, typically is known and is usually related to the variance of the distribution. The functions **h(y,τ)**, **b(θ)**, **T(y)**, **A(θ)** , and **d(τ)** are known. Many common distributions are in this family.

For scalar Y and θ, this reduces to:

$$f_Y(y \mid \theta, \tau) = h(y, \tau) \exp\left(\frac{b(\theta)T(y) - A(\theta)}{d(\tau)}\right)$$

θ is related to the mean of the distribution.

If **b(θ)** is the identity function, then the distribution is said to be in canonical form (or *natural form*).

Note that any distribution can be converted to canonical form by rewriting **θ** as **θ'** and then applying the transformation **θ = b(θ')**.

It is always possible to convert **A(θ)** in terms of the new parametrization, even if is not a one-to-one function; see comments in the page on the exponential family.

If, in addition **T(y)** is the identity and is known, then **θ** is called the ***Canonical Parameter*** **(or *Natural Parameter*)** and is related to the mean through:

$$\boldsymbol{\mu} = \mathrm{E}(\mathbf{Y}) = \nabla A(\boldsymbol{\theta})$$

For scalar Y and **θ**, this reduces to:

$$\mu = \mathrm{E}(Y) = A'(\theta).$$

Under this scenario, the variance of the distribution can be shown to be:

$$\mathrm{Var}(\mathbf{Y}) = \nabla\nabla^{\mathrm{T}} A(\boldsymbol{\theta}) d(\tau).$$

For scalar Y and **θ**, this reduces to

$$\mathrm{Var}(Y) = A''(\theta) d(\tau).$$

Linear predictor

The linear predictor is the quantity which incorporates the information about the independent variables into the model.

The symbol η denotes a linear predictor.

It is related to the expected value of the data (thus, "predictor") through the link function.

η is expressed as linear combinations (thus, "linear") of unknown parameters $\boldsymbol{\beta}$.

The coefficients of the linear combination are represented as the matrix of independent variables **X**. η can thus be expressed as

$$\boldsymbol{\eta} = \mathbf{X} * \boldsymbol{\beta}$$

To model the probability distribution $p(\mathbf{x})$ of a random variable $\mathbf{x}$, given a finite set $\mathbf{x}1, \ldots, \mathbf{x}N$ of observations.

This problem is known as **Density Estimation**.

We shall assume that the data points are independent and identically distributed.

It should be emphasized that the problem of density estimation is fundamentally ill-posed, because there are infinitely many probability distributions that could have given rise to the observed finite data set.

Indeed, any distribution $p(\mathbf{x})$ that is nonzero at each of the data points $\mathbf{x}1, \ldots, \mathbf{x}N$ is a potential candidate.

The issue of choosing an appropriate distribution relates to the problem of model selection.

We begin by considering the binomial and multinomial distributions for discrete random variables and the Gaussian distribution for continuous random variables.

These are specific examples of *parametric* distributions, so-called because they are governed by a small number of adaptive parameters, such as the mean and variance in the case of a Gaussian for example.

To apply such models to the problem of density estimation, we need a procedure for determining suitable values for the parameters, given an observed data set. In a frequentist treatment, we choose specific values for the parameters by optimizing some criterion, such as the likelihood function.

By contrast, in a Bayesian treatment we introduce prior distributions over the parameters and then use Bayes' theorem to compute the corresponding posterior distribution given the observed data.

We shall see that an important role is played by *conjugate* priors, that lead to posterior distributions having the same functional form as the prior, and that therefore lead to a greatly simplified Bayesian analysis.

For example, the conjugate prior for the parameters of the multinomial distribution is called the *Dirichlet* distribution, while the conjugate prior for the mean of a Gaussian is another Gaussian.

All of these distributions are examples of the *exponential family* of distributions, which possess a number of important properties.

One limitation of the parametric approach is that it assumes a specific functional form for the distribution, which may turn out to be inappropriate for a particular application.

An alternative approach is given by *nonparametric* density estimation methods in which the form of the distribution typically depends on the size of the data set.

Such models still contain parameters, but these control the model complexity rather than the form of the distribution.

We end this chapter by considering three nonparametric methods based respectively on histograms, nearest-neighbors, and kernels.

Binary Variables:

We begin by considering a single binary random variable $x \in \{0, 1\}$.

For example, x might describe the outcome of flipping a coin, with $x = 1$ representing 'heads', and $x = 0$ representing 'tails'.

We can imagine that this is a damaged coin so that the probability of landing heads is not necessarily the same as that of landing tails.

The probability of $x = 1$ will be denoted by the parameter μ so that: $p(x = 1|\mu) = \mu$
where $0 \leq \mu \leq 1$, from which it follows that $p(x = 0|\mu) = 1 - \mu$.

The probability distribution over x can therefore be written in the form:

$$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x}$$

which is known as the *Bernoulli* distribution.

It is easily verified that this distribution is normalized and that it has mean and variance given by

$$\mathbf{E}[x] = \mu$$

$$\mathbf{var}[x] = \mu(1 - \mu).$$

Now suppose we have a data set $D = \{x1, \ldots, xN\}$ of observed values of x.

We can construct the likelihood function, which is a function of μ, on the assumption that the observations are drawn independently from $p(x|\mu)$, so that:

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

In a frequentist setting, we can estimate a value for μ by maximizing the likelihood function, or equivalently by maximizing the logarithm of the likelihood. In the case of the Bernoulli distribution, the log likelihood function is given by:

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n)\ln(1-\mu)\}$$

At this point, it is worth noting that the log likelihood function depends on the N observations x_n only through their sum $\Sigma_n x_n$. This sum provides an example of a *sufficient statistic* for the data under this distribution, and we shall study the important role of sufficient statistics in some detail. If we set the derivative of ln $p(D|\mu)$ with respect to μ equal to zero, we obtain the maximum likelihood estimator:

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Probability Densities

As well as considering probabilities defined over discrete sets of events, we also wish to consider probabilities with respect to continuous variables.

We shall limit ourselves to a relatively informal discussion.

If the probability of a real-valued variable x falling in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta x \to 0$, then $p(x)$ is called the ***Probability Density*** over x.

The probability that x will lie in an interval (a, b) is then given by:

$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x.$$

The probability density can be expressed as the derivative of a cumulative distribution function $P(x)$.

Because probabilities are nonnegative, and because the value of x must lie somewhere on the real axis, the probability density $p(x)$ must satisfy the two conditions:

$$\begin{aligned} p(x) &\geqslant 0 \\ \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x &= 1. \end{aligned}$$

Expectations and covariances

One of the most important operations involving probabilities is that of finding weighted averages of functions. The average value of some function $f(x)$ under a probability distribution $p(x)$ is called the *expectation* of $f(x)$ and will be denoted by $\mathrm{E}[f]$. For a discrete distribution, it is given by:

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

so that the average is weighted by the relative probabilities of the different values of x.

In the case of continuous variables, expectations are expressed in terms of an integration with respect to the corresponding probability density.

$$\mathbb{E}[f] = \int p(x) f(x)\,\mathrm{d}x$$

In either case, if we are given a finite number N of points drawn from the probability distribution or probability density, then the expectation can be approximated as a finite sum over these points.

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

Sometimes we will be considering expectations of functions of several variables, in which case we can use a subscript to indicate which variable is being averaged over, so that for instance.

$$\mathbb{E}_x[f(x,y)]$$

denotes the average of the function $f(x,y)$ with respect to the distribution of x. Note that $\mathbb{E}_x[f(x,y)]$ will be a function of y.

We can also consider a *conditional expectation* with respect to a conditional distribution, so that:

$$\mathbb{E}_x[f|y] = \sum_x p(x|y)f(x)$$

with an analogous definition for continuous variables.

The *variance* of *f*(*x*) is defined by:

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right.$$

and provides a measure of how much variability there is in $f(x)$ around its mean value $\mathbb{E}[f(x)]$. Expanding out the square, we see that the variance can also be written in terms of the expectations of $f(x)$ and $f(x)^2$

$$\mathrm{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

In particular, we can consider the variance of the variable *x* itself, which is given by:

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2.$$

For two random variables x and y, the *covariance* is defined by

$$\begin{aligned}\mathrm{cov}[x,y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

One of the most important probability distributions for continuous variables, called the *normal* or *Gaussian* distribution.

For the case of a single real-valued variable *x*, the Gaussian distribution is defined by:

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

which is governed by two parameters: μ, called the *mean*, and σ^2, called the *variance*. The square root of the variance, given by σ, is called the *standard deviation*, and the reciprocal of the variance, written as $\beta = 1/\sigma^2$, is called the *precision*.

The Gaussian distribution satisfies:

$$\mathcal{N}(x|\mu, \sigma^2) > 0.$$

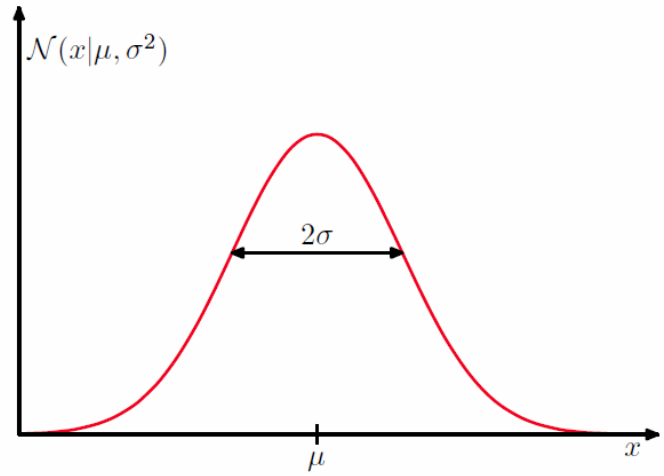

Figure: Plot of the univariate Gaussian showing the mean μ and the standard deviation σ.

Also it is straightforward to show that the Gaussian is normalized, so that:

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right)\, \mathrm{d}x = 1.$$

We can readily find expectations of functions of x under the Gaussian distribution. In particular, the average value of x is given by:

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x\, \mathrm{d}x = \mu$$

Because the parameter μ represents the average value of x under the distribution, it is referred to as the mean. Similarly, for the second order moment.

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x^2\, \mathrm{d}x = \mu^2 + \sigma^2.$$

From the above 2 equations, it follows that the variance of x is given by:

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

and hence σ2 is referred to as the variance parameter. The maximum of a distribution is known as its mode. For a Gaussian, the mode coincides with the mean.

We are also interested in the Gaussian distribution defined over a D-dimensional vector $\mathbf{x}$ of continuous variables, which is given by:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

where the D-dimensional vector $\boldsymbol{\mu}$ is called the mean, the $D \times D$ matrix $\boldsymbol{\Sigma}$ is called the covariance, and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

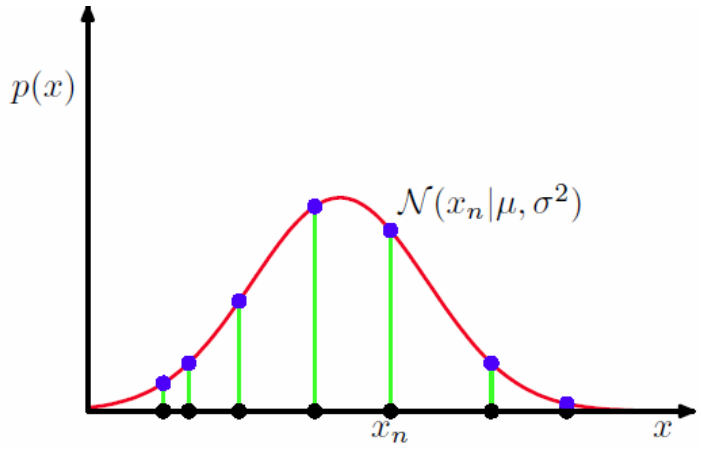

Figure: Illustration of the likelihood function for a Gaussian distribution, shown by the red curve. Here the black points denote a data set of values $\{x_n\}$, and the likelihood function corresponds to the product of the blue values. Maximizing the likelihood involves adjusting the mean and variance of the Gaussian so as to maximize this product.

Now suppose that we have a data set of observations $\mathbf{x} = (x_1, \ldots, x_N)^{\mathrm{T}}$, representing N observations of the scalar variable x. Note that we are using the typeface $\mathbf{x}$ to distinguish this from a single observation of the vector-valued variable $(x_1, \ldots, x_D)^{\mathrm{T}}$, which we denote by $\mathbf{x}$. We shall suppose that the observations are drawn independently from a Gaussian distribution whose mean μ and variance σ^2 are unknown, and we would like to determine these parameters from the data set.

Data points that are drawn independently from the same distribution are said to be *independent and identically distributed*, which is often abbreviated to i.i.d. We have seen that the joint probability of two independent events is given by the product of the marginal probabilities for each event separately. Because our data set $\mathbf{x}$ is i.i.d., we can therefore write the probability of the data set, given μ and σ^2, in the form.

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu, \sigma^2\right)$$

When viewed as a function of μ and σ^2, this is the likelihood function for the Gaussian and is interpreted diagrammatically in Figure above.

One common criterion for determining the parameters in a probability distribution using an observed data set is to find the parameter values that maximize the likelihood function.

This might seem like a strange criterion because, from our foregoing discussion of probability theory, it would seem more natural to maximize the probability of the parameters given the data, not the probability of the data given the parameters.

In fact, these two criteria are related, as we shall discuss in the context of curve fitting.

For the moment, however, we shall determine values for the unknown parameters μ and $\sigma 2$ in the Gaussian by maximizing the likelihood function.

In practice, it is more convenient to maximize the log of the likelihood function.

Because the logarithm is a monotonically increasing function of its argument, maximization of the log of a function is equivalent to maximization of the function itself.

Taking the log not only simplifies the subsequent mathematical analysis, but it also helps numerically because the product of a large number of small probabilities can easily underflow the numerical precision of the computer, and this is resolved by computing instead the sum of the log probabilities.

The log likelihood function can be written in the form:

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

Maximizing the above equation with respect to μ, we obtain the maximum likelihood solution is given by:

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n$$

which is the *sample mean*, i.e., the mean of the observed values *{xn}*. Similarly, maximizing (1.54) with respect to σ^2, we obtain the maximum likelihood solution for the variance in the form:

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

which is the *sample variance* measured with respect to the sample mean μ_{ML}.