Classification

Volker Blobel – University of Hamburg
March 2005

Given objects (e.g. particle tracks), which have certain features (e.g. momentum $\boldsymbol{p}$, specific energy loss $\mathrm{d}E/\,\mathrm{d}x$) and which belong to one of certain classes (e.g. particle types).

Classification = assigment of object to classes

1. Strategies for classification
2. Minimizing the probability for misclassification
3. Risk minimization

Christopher M. Bishop, Neural Networks for Pattern Recognition, Clarendon Press Oxford

Strategies for classification

In statistical classification problems an object belongs to one of several classes $\mathcal{C}_k$.

Each object has to be assigned to **one** of the classes on the basis of certain features of the object.

Example from particle physics analysis: identification of the particle type in a high energy particle reaction; here the classes are the particles types *pion, kaon, proton* ....

Classification strategies can be based on

- minimization of the probability of misclassification, or on
- risk minimization; this requirement applies to cases with rare classes, where rejecting the correct class has to be avoided.

Probabilities and classification

Formally **a-priori probabilities** $P(\mathcal{C}_k)$ of an object belonging to each of the classes $\mathcal{C}_k$ are introduced. This probability $P(\mathcal{C}_k)$ corresponds to the fractions of objects in each class in the limit of an infinite number of observations. If a classification is required, but nothing is known about the features of the objects, the best strategy is to assign it to the class having the highest a-priori probability. This minimizes the probability of misclassification, although some classifications must be wrong. In the particle identification example one would classify each particle as a *pion*, the most common particle created in a high energy reaction.

Information in terms of values x of certain features connected to each object is of statistical nature and has to be used to improve the classification. The distribution of values x follows a certain probability density function $p(x)$. The probability of x lying in an interval (a, b) of x is given by

$$P\left(x \in [a,b]\right) = \int_a^b p(x)\,\mathrm{d}x\ .$$

If there are several feature variables x_1, x_2, ..., they are grouped into a vector $\boldsymbol{x}$ corresponding to a point in multidimensional feature space. The probability of a multidimensional $\boldsymbol{x}$ lying in a region $\mathcal{R}$ of $\boldsymbol{x}$-space is given by

$$P\left(\boldsymbol{x} \in \mathcal{R}\right) = \int_{\mathcal{R}} p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\ .$$

Class-conditional probability

The class-conditional probability density $p(\boldsymbol{x}|\mathcal{C}_k)$ is the density of $\boldsymbol{x}$ *given* that it belongs to class $\mathcal{C}_k$. The unconditional density function $p(x)$ is the density function for $\boldsymbol{x}$, irrespective of the class, and is given by the sum

$$p(\boldsymbol{x}) = \sum_k p(\boldsymbol{x}|\mathcal{C}_k) \cdot P(\mathcal{C}_k) \ .$$

If, for a given object, the value $\boldsymbol{x}$ of the feature vector is known, one has to combine this information with the a-priori probabilities.

This is done by Bayes' theorem

$$\text{a-posteriori probability for class } k: \quad P(\mathcal{C}_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathcal{C}_k) \cdot P(\mathcal{C}_k)}{p(\boldsymbol{x})} \ .$$

The denominator $p(\boldsymbol{x})$, the unconditional density function, ensures the correct normalization of the a-posteriori probabilities $P(\mathcal{C}_k|\boldsymbol{x})$:

$$\text{Normalization:} \quad \sum_k P(\mathcal{C}_k|\boldsymbol{x}) = 1 \ .$$

Notation

Lower-case letters are used for probability densities, and upper-case letters for probabilities.

k-th class	$\mathcal{C}_k$
point in feature space	$x,\ \boldsymbol{x}$
probability density	$p(x),\ p(\boldsymbol{x})$
class-conditional probability density	$p(\boldsymbol{x}\|\mathcal{C}_k)$
a-priori probability for class k	$P(\mathcal{C}_k)$
a-posteriori probability for class k	$P(\mathcal{C}_k\|\boldsymbol{x})$
discriminant function for class $\mathcal{C}_k$	$y_k(\boldsymbol{x})$
loss matrix	$\boldsymbol{L}$ with elements L_{kj}
rejection threshold	P_{th}

Minimizing the probability of misclassification

After the observation of a feature vector $\boldsymbol{x}$, the a-posteriori probabilities $P(\mathcal{C}_k|\boldsymbol{x})$ give the probability of the pattern belonging to class $\mathcal{C}_k$.

The probability of misclassification is minimized by selecting class $\mathcal{C}_k$ if

$$P\left(\mathcal{C}_k|\boldsymbol{x}\right) > P\left(\mathcal{C}_j|\boldsymbol{x}\right) \qquad \text{for all} \quad j \neq k$$

For this comparison the unconditional density $p(\boldsymbol{x})$ may be removed from the Bayes' formula, and the criterion can be rewritten in the form

$$\boxed{P\left(\boldsymbol{x}|\mathcal{C}_k\right) \cdot P\left(\mathcal{C}_k\right) > P\left(\boldsymbol{x}|\mathcal{C}_j\right) \cdot P\left(\mathcal{C}_j\right) \qquad \text{for all} \quad j \neq k}\ .$$

This is the fundamental decision formula, which minimizes the *probability of misclassification*.

Each point of feature space is assigned to one of c classes. The feature space is divided up into c decision regions $\mathcal{R}_1$, $\mathcal{R}_2$ $\ldots$ $\mathcal{R}_c$; a feature vector $\boldsymbol{x}$ falling in region $\mathcal{R}_k$ is assigned to class $\mathcal{C}_k$.

The decision regions $\mathcal{R}_k$ for a particular class $\mathcal{C}_k$ may be contiguous or may be divided into disjoint regions.

The boundaries between these regions are called decision boundaries (surfaces).

Decision probabilities

The probability for a correct decision is given by

$$\begin{aligned} P\,(\text{correct}) &= \sum_k P(\boldsymbol{x} \in \mathcal{R}_k,\, \mathcal{C}_k) = \sum_k P(\boldsymbol{x} \in \mathcal{R}_k | \mathcal{C}_k) \cdot P\,(\mathcal{C}_k) \\ &= \sum_k \int_{\mathcal{R}_k} p(\boldsymbol{x} \in \mathcal{R}_k | \mathcal{C}_k) \cdot P\,(\mathcal{C}_k)\; \mathrm{d}\boldsymbol{x}\;. \end{aligned}$$

This probability is maximized by choosing the regions $\{\mathcal{R}_k\}$ such that each $\boldsymbol{x}$ is assigned to the class for which the integrand is a maximum.

Error probabilities in the case of only two classes:

$$\begin{aligned} P(\text{error}) &= P(\boldsymbol{x} \in \mathcal{R}_2,\, \mathcal{C}_1) + P(\boldsymbol{x} \in \mathcal{R}_1,\, \mathcal{C}_2) = P(\boldsymbol{x} \in \mathcal{R}_2 | \mathcal{C}_1) \cdot P(\mathcal{C}_1) + P(\boldsymbol{x} \in \mathcal{R}_1 | \mathcal{C}_2) \cdot P(\mathcal{C}_2) \\ &= \int_{\mathcal{R}_2} p(\boldsymbol{x} | \mathcal{C}_1) \cdot P(\mathcal{C}_1)\, \mathrm{d}\boldsymbol{x} + \int_{\mathcal{R}_1} p(\boldsymbol{x} | \mathcal{C}_2) \cdot P(\mathcal{C}_2)\, \mathrm{d}\boldsymbol{x} \end{aligned}$$

$P(\boldsymbol{x} \in \mathcal{R}_2,\, \mathcal{C}_1)$ is the joint probability of $\boldsymbol{x}$ being assigned to class $\mathcal{C}_2$, and the true class being $\mathcal{C}_1$.

If $p(\boldsymbol{x}|\mathcal{C}_1) \cdot P(\mathcal{C}_1) > p(\boldsymbol{x}|\mathcal{C}_2) \cdot P(\mathcal{C}_2)$ for a given $\boldsymbol{x}$, the regions $\mathcal{R}_1$ and $\mathcal{R}_2$ have to be chosen such that $\boldsymbol{x}$ is in region $\mathcal{R}_1$, since this gives the smaller contribution to the error.

Discriminant functions

Decision on class membership has been based solely on the relative sizes of the probabilities. The classification can be reformulated in terms of a set of *discriminant functions* $y_1(\boldsymbol{x})$, $y_2(\boldsymbol{x}) \ldots$.

A feature vector $\boldsymbol{x}$ is assigned to class $\mathcal{C}_k$ if

$$y_k(\boldsymbol{x}) > y_j(\boldsymbol{x}) \qquad \text{for all} \quad j \neq k \; .$$

The decision rule for minimizing the probability of misclassification can be made by choosing

$$y_k(\boldsymbol{x}) = P(\mathcal{C}_k|\boldsymbol{x}) \; .$$

Since the unconditional density $p(\boldsymbol{x})$ does not affect the classification decision an equivalent choice is

$$y_k(\boldsymbol{x}) = p(\boldsymbol{x}|\mathcal{C}_k) \cdot P(\mathcal{C}_k) \; .$$

Furthermore, since only the relative magnitude of the discriminat function are important in determinating the class, replacing $y(\boldsymbol{x})$ by $g(y(\boldsymbol{x}))$, where $g(\cdot)$ is any monotonic function, will not modify the classification.

Taking the logarithms:

$$y_k(\boldsymbol{x}) = \ln p(\boldsymbol{x}|\mathcal{C}_k) + \ln P(\mathcal{C}_k) \; .$$

Decision boundaries are given by the regions where the discriminant functions are equal. If the regions $\mathcal{R}_k$ and $\mathcal{R}_j$ are contiguous then the decision boundary is given by $y_k(\boldsymbol{x}) = y_j(\boldsymbol{x})$.

Two-class discriminant functions

In this special case the two discriminant functions $y_1(\boldsymbol{x})$ and $y_2(\boldsymbol{x})$ can alternatively be combined to a single discriminant function

$$y(\boldsymbol{x}) = y_1(\boldsymbol{x}) - y_2(\boldsymbol{x}) \; .$$

The pattern $\boldsymbol{x}$ is assigned to class $\mathcal{C}_1$ if $y(\boldsymbol{x}) > 0$ and to class $\mathcal{C}_2$ if $y(\boldsymbol{x}) < 0$. Alternative forms of the single discriminant function are:

$$\begin{aligned} & y(\boldsymbol{x}) &=& P(\mathcal{C}_1|\boldsymbol{x}) - P(\mathcal{C}_2|\boldsymbol{x}) \\ \text{and} \qquad & y(\boldsymbol{x}) &=& \ln \frac{p(\boldsymbol{x}|\mathcal{C}_1)}{p(\boldsymbol{x}|\mathcal{C}_2)} + \ln \frac{P(\mathcal{C}_1)}{P(\mathcal{C}_2)} \; . \end{aligned}$$

Risk minimization

In many applications the minimization of the probability of misclassifying is not the most appropriate criterion, especially if there are rare processes, where rejection of the correct class has serious consequences. This may be taken into account in the following way.

A loss matrix $\boldsymbol{L}$ is defined. where the elements L_{kj} have the meaning of a **penalty** associated with assigning a pattern to class $\mathcal{C}_j$, when it in fact belongs to class $\mathcal{C}_k$. Then the expected (average) loss for those patterns is given by

$$R_k = \sum_k L_{kj} \int_{\mathcal{R}_j} p\left(\boldsymbol{x}|\mathcal{C}_k\right) \, \mathrm{d}\boldsymbol{x} \; .$$

The overall expected loss, or *risk*, for patterns from all classes is

$$R = \sum_k R_k \cdot P\left(\mathcal{C}_k\right) = \sum_j \int_{\mathcal{R}_j} \left\{ \sum_k L_{kj} \, p\left(\boldsymbol{x}|\mathcal{C}_k\right) \cdot P\left(\mathcal{C}_k\right) \right\} \mathrm{d}\boldsymbol{x} \; .$$

The *risk is minimized* if the integrand is minimized at each point $\boldsymbol{x}$, i.e. if the regions $\mathcal{R}_j$ are chosen such that $\boldsymbol{x} \in \mathcal{R}_j$ when

$$\boxed{\sum_k L_{kj} \, p(\boldsymbol{x}|\mathcal{C}_k) \cdot P(\mathcal{C}_k) < \sum_k L_{ki} \, p(\boldsymbol{x}|\mathcal{C}_k) \cdot P(\mathcal{C}_k) \qquad \text{for all} \quad i \neq j} \; .$$

Loss matrix

This is a generalization of the decision rule for minimization the probability of misclassification. The general formula reduces to the special formula for the special assigment

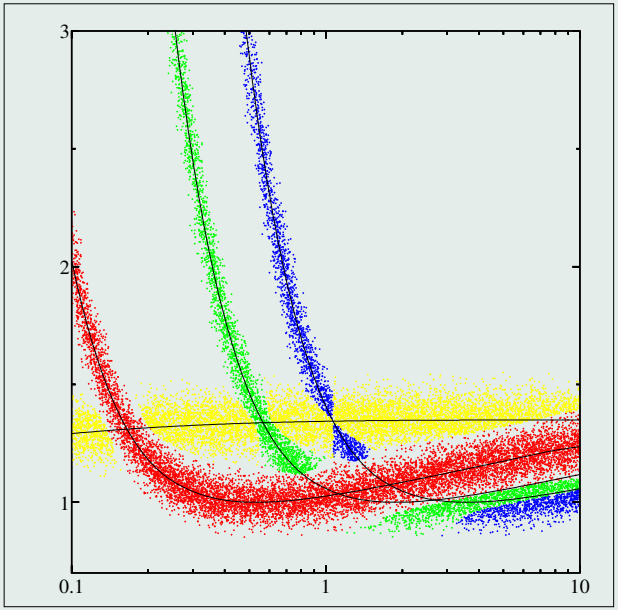
$$L_{kj} = \begin{cases} 0 & \text{for} \quad k = j \quad \text{(correct class)} \\ 1 & \text{for} \quad k \neq j \quad \text{(wrong class)} \end{cases} \quad .$$

In general the elements of the loss matrix have to be estimated based on **experience**. If class $\mathcal{C}_j$ is a **rare**, but important class, which should have a small loss, then for $k \neq j$ the elements L_{kj} should be small; otherwise the acceptance of class $\mathcal{C}_j$ is small because of the small value of the a-priori probability $P(\mathcal{C}_j)$ for rare classes.

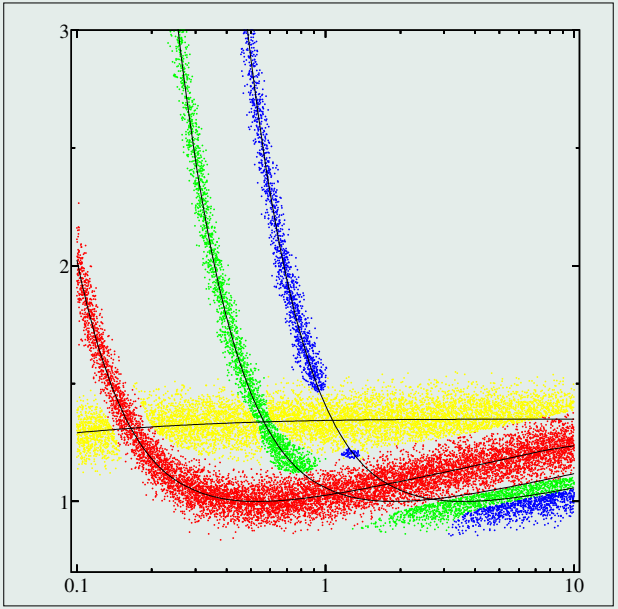
In any case the classification has to be studied carefully in order to check the values of the a-priori probabilities and the loss matrix $\boldsymbol{L}$.

Example: particle identification

$\mathrm{d}E/\,\mathrm{d}x$-curves for π (red), $K^{\pm}$ (green), p (blue) and electron (yellow) with a-priori probabilities 100 : 10 : 5 : 5



Default loss matrix



Loss matrix elements in favour of electron with element 0.2 instead of 1 for p, 0.8 for π and $K^{\pm}$.

Rejection thresholds

Most classification errors are expected to occur in those regions of $\boldsymbol{x}$-space where the largest of the a-posteriori probabilities is relatively **low**, due to a strong overlap between different classes.

It may be better **not** to make a classification decision in those case, and to reject the pattern instead.

Rejection can be made on the basis of a threshold P_{th} in the range 0 to 1, according to

$$\text{if} \quad \max_k P(\mathcal{C}_k|\boldsymbol{x}) \quad \begin{cases} \geq P_{\mathrm{th}}, & \text{then classify } \boldsymbol{x} \\ < P_{\mathrm{th}}, & \text{then reject } \boldsymbol{x} \ . \end{cases}$$

Normal distribution and discriminant functions

If the class-conditional density functions are independent normal distributions, then the discriminant function has the form

$$y_k(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{V}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) - \frac{1}{2}\ln|\boldsymbol{V}_k| + \ln P(\mathcal{C}_k)\ ,$$

where constant terms have been dropped. Decision boundaries given by conditions $y_k(\boldsymbol{x}) = y_j(\boldsymbol{x})$ are general quadratic functions in the feature space.

The discriminant functions are simplified if the covariance matrices for the various classes are equal ($\boldsymbol{V}_k \equiv \boldsymbol{V}$). Then the terms with the determinant $|\boldsymbol{V}|$ are identical for all classes and can be dropped. The quadratic contribution $\boldsymbol{x}^T\boldsymbol{V}^{-1}\boldsymbol{x}$ is class-independent too and can be dropped. Because $\boldsymbol{x}^T\boldsymbol{V}^{-1}\boldsymbol{\mu}_k = \boldsymbol{\mu}_k^T\boldsymbol{V}^{-1}\boldsymbol{x}$ the discriminant functions can be simplified to the form

$$y_k(\boldsymbol{x}) = \boldsymbol{w}_k^T\boldsymbol{x} + w_{k0} \qquad \text{with} \qquad \begin{aligned} \boldsymbol{w}_k^T &= \mu_k^T\boldsymbol{V}^{-1} \\ w_{k0} &= -\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{V}^{-1}\boldsymbol{\mu}_k + \ln P(\mathcal{C}_k)\ . \end{aligned}$$

The functions are *linear* in $\boldsymbol{x}$; decision boundaries corresponding to $y_k(\boldsymbol{x}) = y_j(\boldsymbol{x})$ are then hyperplanes.

Classification