TWO STATISTICAL MODELS

Václav Hlaváč

Czech Technical University, Faculty of Electrical Engineering
Department of Cybernetics, Center for Machine Perception
121 35 Praha 2, Karlovo nám. 13, Czech Republic

`hlavac@fel.cvut.cz`, http://cmp.felk.cvut.cz

LECTURE PLAN

1. Conditional independence of features.
2. Gaussian probability distribution.
3. Straightening of the feature space $\Longrightarrow$ linear classification.

TWO OFTEN USED STATISTICAL MODELS

Two simple special cases of $p_{X|K}\colon X \times K \rightarrow \mathbb{R}$ of conditional probabilities of observations $x \in X$, under the condition that the object is in a state $k \in K$ are often used as statistical models of the recognized objects.

- Conditional independence of features.
- Gaussian probability distribution.

CONDITIONAL INDEPENDENCE OF FEATURES

- An observation $x=(x_1, x_2, \ldots, x_n)$. Each feature $x_i \in X_i$, $i \in I$.
- The set of observations X is a Cartesian product $X = X_1 \times X_2 \times \ldots \times X_n$.
- It is assumed that the probabilities $p_{X|K}(x \,|\, k)$ have the form

$$p_{X|K}(x \,|\, k) = \prod_{i=1}^{n} p_{X_i|K}(x_i \,|\, k) \; .$$

- Features become mutually independent at the fixed state k.
- The object's features are dependent on each other but all the dependence is realized via the dependence on the state of the object. If the state is fixed then the mutual dependence among the features disappears.
- This is the simplest model of the conditional independence.

CONDITIONAL INDEPENDENCE OF FEATURES (2)

However, the conditional independence assumption does not mean that the features are also *a priori* mutually independent.

In general,

$$p_X(x) \neq \prod_{i=1}^{n} p_{X_i}(x_i) \,.$$

THE SIMPLEST CASE
FEATURES ONLY $\{0, 1\}$, TWO STATES ONLY

Simplifying assumptions: Features x_i, $i = 1, \ldots, n$, assume only two values {0,1} and the number of hidden states is 2, $k_1 = 1$ or $k_2 = 2$.

The strategy: solving any Bayesian and non-Bayesian task under our simplest assumptions can be implemented as a decomposition of the set of vertices on an n-dimensional hypercube by a hyperplane.

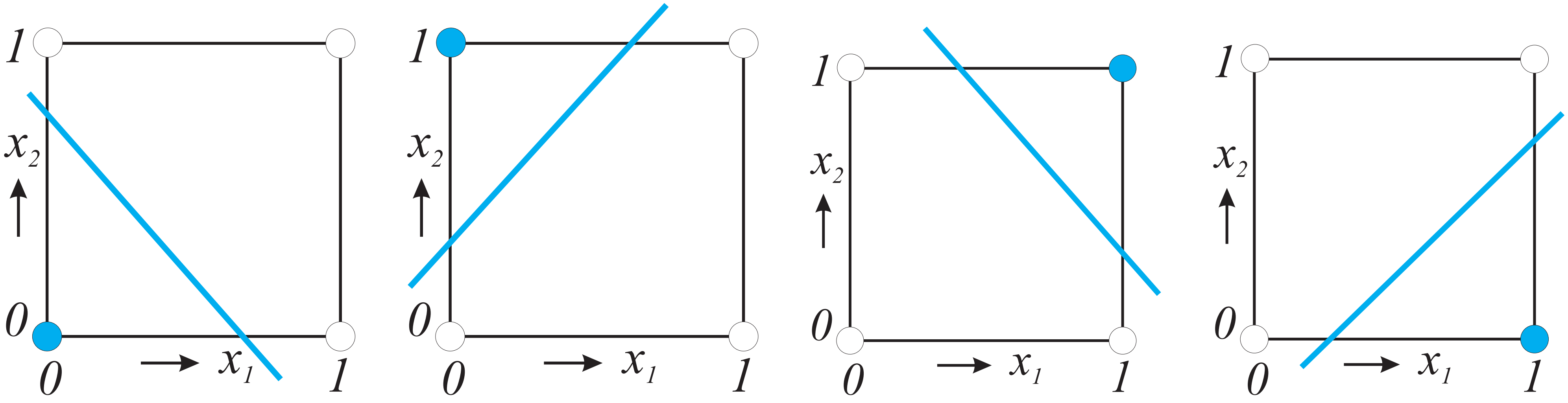


Illustration for n=2.

THE SIMPLEST CASE
FEATURES ONLY $\{0,1\}$, TWO STATES ONLY, cont.

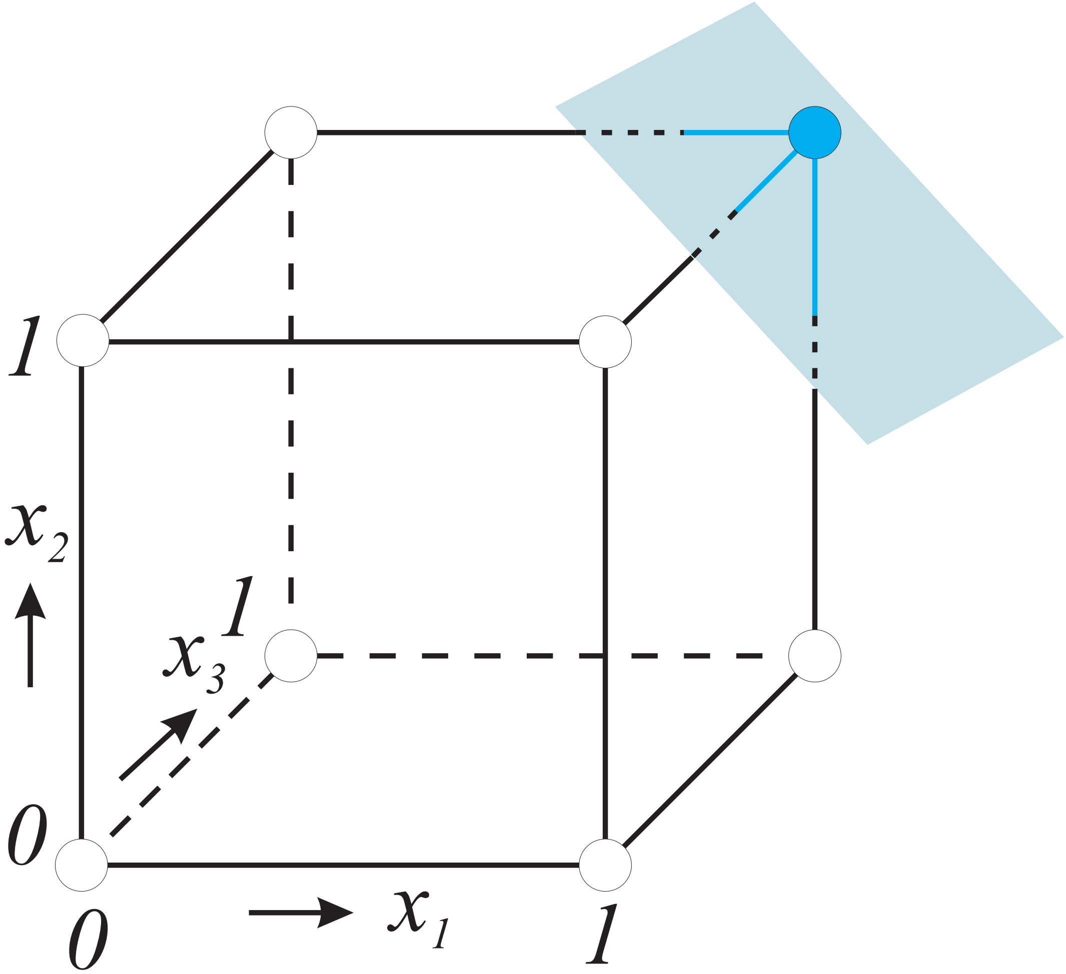


Illustration for n=3, only one of eight possible cases shown.

THE SIMPLEST CASE, THE STRATEGY

- The strategy decomposes the set of vertices on an n-dimensional hypercube by a hyperplane.
- An interval of values of likelihood ratio corresponds to each decision d, i.e., the decision d is taken for which

$$\theta_{\min} < \frac{p_{X|K}(x \,|\, k=1)}{p_{X|K}(x \,|\, k=2)} \leq \theta_{\max} \,,$$

where $\theta_{\min}$ and $\theta_{\max}$ are threshold values.

- Inequality is not changed if monotonic function as $\log$ is applied,

$$\theta'_{\min} < \log \frac{p_{X|K}(x \,|\, k=1)}{p_{X|K}(x \,|\, k=2)} \leq \theta'_{\max}$$

MORE THAN TWO FEATURES

Let us assume n features, $i = 1 \ldots n$. Then

$$\begin{aligned} & \log \frac{p_{X|K}(x \,|\, k=1)}{p_{X|K}(x \,|\, k=2)} = \\ = \quad & \sum_{i=1}^{n} \log \frac{p_{X_i|K}(x_i \,|\, k=1)}{p_{X_i|K}(x_i \,|\, k=2)} = \\ = \quad & \sum_{i=1}^{n} x_i \log \frac{p_{X_i|K}(1 \,|\, k=1)\; p_{X_i|K}(0 \,|\, k=2)}{p_{X_i|K}(1 \,|\, k=2)\; p_{X_i|K}(0 \,|\, k=1)} \\ & + \sum_{i=1}^{n} \log \frac{p_{X_i|K}(0 \,|\, k=1)}{p_{X_i|K}(0 \,|\, k=2)} \,. \end{aligned}$$

The logarithm of the likelihood ratio is a linear function of features x_i.

LINEARITY

$$\theta'_{\min} < \log \frac{p_{X|K}(x \,|\, k=1)}{p_{X|K}(x \,|\, k=2)} \le \theta'_{\max}$$

Can be rewritten to

$$\theta'_{\min} < \sum_{i=1}^{n} \alpha_i \, x_i \le \theta'_{\max} .$$

- If the tasks are expressed by a firmly chosen function $p_{X|K}$ then various strategies differ only by a threshold value.
- If, in addition, the function $p_{X|K}$ varies then also the coefficients α_i start varying.
- At all these changes, it remains valid that all decision regions are regions, where values of a linear function belong to a contiguous interval.

CASE OF TWO POSSIBLE DECISIONS ONLY

- The observation set X is to be divided into two subsets X_1 and X_2.
- The decision function assumes the form

$$x \in \begin{cases} X_1, & \text{if} \quad \sum_{i=1}^{n} \alpha_i\, x_i \leq \theta\,, \\ X_2, & \text{if} \quad \sum_{i=1}^{n} \alpha_i\, x_i > \theta\,. \end{cases}$$

- This means that for objects characterized by binary and conditionally independent features, the search for the needed strategy is equal to searching for coefficients α_i and the threshold value θ.
- Linear classifiers deal with how to tune these coefficients and thresholds properly.

GAUSSIAN PROBABILITY DISTRIBUTION

- Let a set of observations X be an n-dimensional linear space.
- So far, it has been assumed that X is a finite set. Nevertheless, the results derived earlier can be used in most situations even in this case.
- It is sufficient to mention that the number $p_{X|K}(x \,|\, k)$ does not mean a probability but a probability density.

MULTI-DIMENSIONAL GAUSSIAN RANDOM VARIABLE

We will assume $p_{X|K} \colon X \times K \to \mathbb{R}$ of the form

$$p_{X|K}(x \,|\, k) = C(A^k) \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^k \left(x_i - \mu_i^k\right) \left(x_j - \mu_j^k\right) \right) , \text{ where}$$

- k is a superscript index and not a power.
- x_i is a value of the i-th feature of the object.
- μ_i^k is the conditional mathematical expectation of the i-th feature under the condition that the object is in the state k.
- A^k is the inverse covariance matrix, $A^k = (B^k)^{-1}$. The element b_{ij}^k in the matrix B^k corresponds to the covariance between the i-th and the j-th features, i.e., the conditional mathematical expectation of the product $(x_i - \mu_i^k)(x_j - \mu_j^k)$ under the condition that the object is in the state k.
- $C(A^k)$ is a normalization coefficient (the integral over the whole domain of the function $= 1$).

SPECIAL CASE (1):
2 HIDDEN STATES, 2 DECISIONS

The optimal decision strategy is a quadratic decision function

$$x \in \begin{cases} X_1, & \text{if} \quad \sum\limits_i \sum\limits_j \alpha_{ij}\, x_i\, x_j + \sum\limits_i \beta_i\, x_i \le \gamma\,, \\ X_2, & \text{if} \quad \sum\limits_i \sum\limits_j \alpha_{ij}\, x_i\, x_j + \sum\limits_i \beta_i\, x_i > \gamma\,. \end{cases}$$

Coefficients α_{ij}, β_i, $i, j = 1, 2, \ldots, m$, and the threshold value γ depend on a statistical model of the object, i.e., on matrices A^1, A^2, vectors μ^1, μ^2,

and also on the fact which Bayesian or non-Bayesian decision task is to be solved.

SPECIAL CASE (2):
2 HIDDEN STATES, 2 DECISIONS

Coefficients α_{ij}, β_i, $i, j = 1, 2, \ldots, m$, the threshold γ depend

◆ on a statistical model of the object, i.e., on matrices A^1, A^2, vectors μ^1, μ^2

◆ on the Bayesian or non-Bayesian decision problem to be solved.

SPECIAL CASE (3):
2 HIDDEN STATES, 2 DECISIONS

Even in the two-dimensional case, the variability of geometrical forms, which the sets X_1 and X_2 assume, is quite rich.

The border between the sets X_1 and X_2 can be

1. A single straight line in between.
2. A pair of parallel lines located in the way that X_1 is positioned between the lines and X_2 is the rest.
3. A pair of intersecting straight lines $\Longrightarrow$ four sectors. Two sectors represent the set X_1 and the other two X_2.
4. An ellipse, X_1 lies inside and X_2 outside.
5. The border can be created by hyperbolae, X_1 is between the hyperbolae, X_2 is expressed as two convex sets. Both sets are marked off by one of the continuous hyperbolae.

STRAIGHTENING OF THE FEATURE SPACE (1)

Goal: to express a given nonlinear decision problem as a linear problem.

This allows to use very well developed linear classifiers,

$$q(x, w, b) = w^\top x + b = \sum_{i=1}^{n} w_i x_i + b,$$

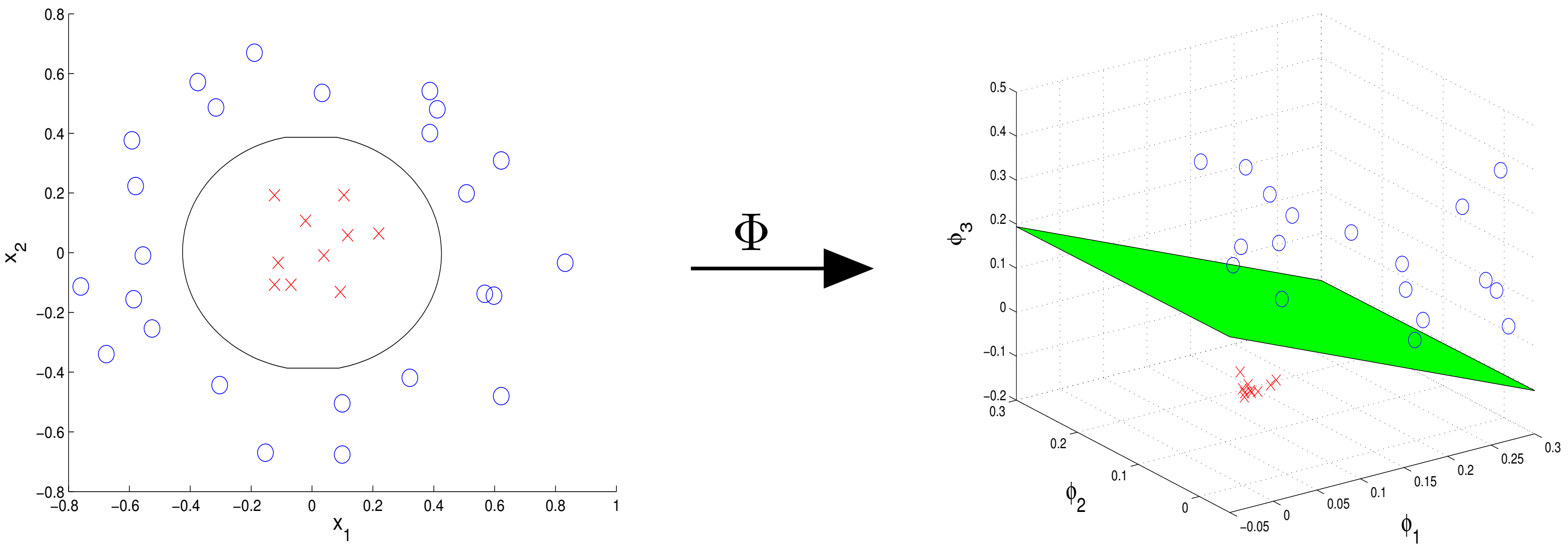
where i is the index of vectors w, $x \in \mathcal{X}$.

Notice that $f(x) = 0$ expresses the hyperplane in $\mathbb{R}^n$.

Solution - a nonlinear mapping: vectors of $\mathcal{X}$ are represented in a new space $\mathcal{F}$ using mapping function $\Phi\colon \mathcal{X} \rightarrow \mathcal{F}$, such as

$$q'(x, w, b) = w^\top \Phi(x) + b = \sum_{i=1}^{n} w_i \Phi_i(x_i) + b$$

Input space $\mathcal{X}$ **Feature space $\mathcal{F}$**



STRAIGHTENING OF THE FEATURE SPACE (2)

All variety of geometric forms can be summarized into a single form, in which the border between classes is constituted only by a hyperplane.

$$x \in \begin{cases} X_1, & \text{if} \quad \sum_i \alpha_i \, x_i \le \gamma \,, \\ X_2, & \text{if} \quad \sum_i \alpha_i \, x_i > \gamma \,. \end{cases}$$

The original n dimensional space is transformed into the $\left(n + \frac{1}{2} n(n+1)\right)$-dimensional feature space.

Old dimension	1	2	3	4	5	6	10	20
New dimension	2	5	9	14	20	27	65	230

STRAIGHTENING THE FEATURE SPACE, HOW?

Original features $x = (x_1, x_2, \ldots, x_i, \ldots, x_n)$ are transformed to

$$\begin{array}{rrrrrrrrr}
y = (& x_1\,, & x_2\,, & \ldots, & x_i\,, & \ldots, & x_{n-1}\,, & x_n\,, & \\
& x_1x_1\,, & x_1x_2\,, & \ldots, & x_1x_i\,, & \ldots, & x_1x_{n-1}\,, & x_1x_n\,, & \\
& & x_2x_2\,, & \ldots, & x_2x_i\,, & \ldots, & x_2x_{n-1}\,, & x_2x_n\,, & \\
& & & & & & \vdots & & \\
& & & & x_ix_i\,, & \ldots, & x_ix_{n-1}\,, & x_i\,x_n\,, & \\
& & & & & & \vdots & & \\
& & & & & & x_{n-1}x_{n-1}\,, & x_{n-1}x_n\,, & \\
& & & & & & & x_nx_n\,) & .
\end{array}$$

New linear decision rule $$y \in \begin{cases} Y_1, & \text{if} \quad \sum\limits_i \alpha_i\, y_i \le \gamma\,, \\ Y_2, & \text{if} \quad \sum\limits_i \alpha_i\, y_i > \gamma\,. \end{cases}$$

STRAIGHTENING, A 1D EXAMPLE

Assume x is a 1D random variable with the Gaussian distribution. Original strategy for two classes X_1, X_2

$$x \in \begin{cases} X_1, & \text{if} \quad (x - x_0)^2 < \delta\,, \\ X_2, & \text{if} \quad (x - x_0)^2 \geq \delta\,, \end{cases}$$

Straightening $y_1 = x^2$, $y_2 = x$.

$$x \in \begin{cases} X_1, & \text{if} \quad \alpha_1\, y_1 + \alpha_2\, y_2 > \theta\,, \\ X_2, & \text{if} \quad \alpha_1\, y_1 + \alpha_2\, y_2 \leq \theta\,, \end{cases}$$

where $y_1 = x^2$, $y_2 = x$, $\alpha_1 = -1$, $\alpha_2 = 2\, x_0$, $\theta = x_0^2 - \delta$.

STRAIGHTENING, A 2D EXAMPLE

- Assume 2D feature space and a quadratic decision strategy, i.e., $q(x)$ is a polynomial of degree 2.
- Mapping functions are: $\Phi_1 = x_1$, $\Phi_2 = x_2$, $\Phi_3 = x_1 x_2$, $\Phi_4 = x_1 x_1$, $\Phi_5 = x_2 x_2$.

$$q(x) = w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2 = w^\top \Phi_i(x) \; .$$