Two statistical models

Václav Hlaváč

Czech Technical University in Prague
Czech Institute of Informatics, Robotics and Cybernetics
160 00 Prague 6, Jugoslávských partyzánů 1580/3, Czech Republic
http://people.ciirc.cvut.cz/hlavac, vaclav.hlavac@cvut.cz
also Center for Machine Perception, http://cmp.felk.cvut.cz

Courtesy: M.I. Schlesinger, V. Franc.

Outline of the talk:

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

Two often used statistical models

- Consider class-conditional probabilities (also named likelihoods) $p_{X|Y}\colon X \times Y \to \mathbb{R}$ dealing with observations $x \in X$, under the condition that the object is in a state $y \in Y$.
- This probability $p_{X|Y}$ is the often used statistical model of the recognized objects.
- There are two important simple special cases of the likelihood $p_{X|Y}$:
  - Conditional independence of features .

    (Relates to Naïve Bayes classifier, which more specific as it assumes statistical independence of features).
  - Gaussian probability distribution.

The difference between the probability and likelihood (1)

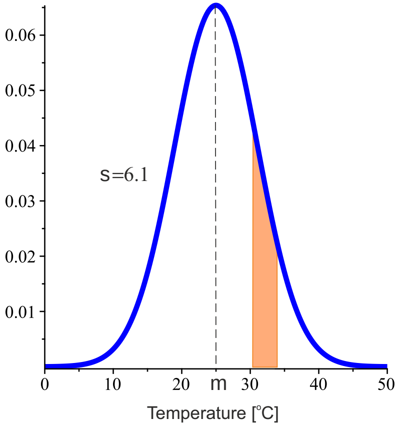
- In statistics, the likelihood function (simply likelihood) is the probability that some fixed outcome was generated by a random distribution with a specific unknown parameter.
- Probability predicts future outcome (events) given a fixed parameter(s) value(s).
- Consider a probability model with parameters Θ. $p(x|\Theta)$ has two interpretations and names.
  - Probability of X given parameters Θ.

    Probability is the area under fixed probability distribution.
  - Likelihood of parameters Θ given that x was observed.

    Likelihood L is the y axis value for fixed data points x with distribution that can be moved to the new position on the x-axis if the particular value x was observed.

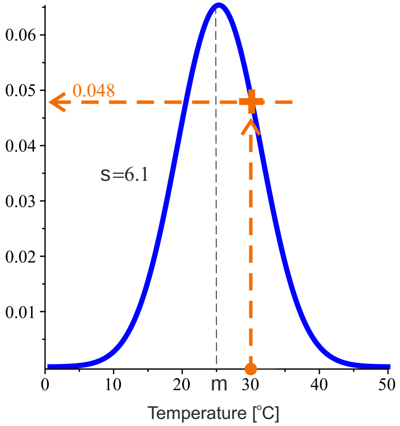
    In the example in the bottom right figure:
    $L = p(\text{Gaussian}, \mu = 25, \sigma = 6.1|\text{temperature} = 30°\text{C}) = 0.048$.

probability



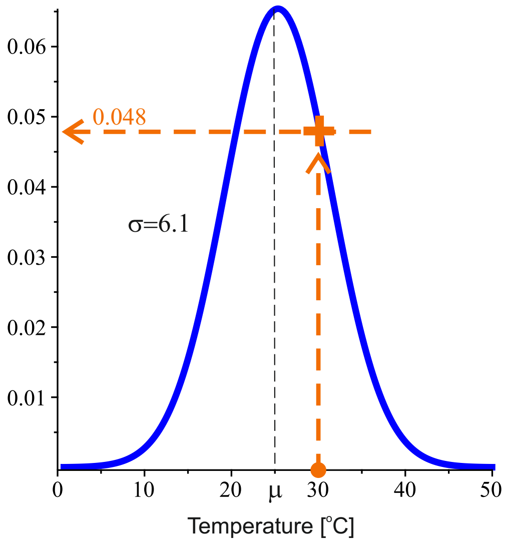
versus

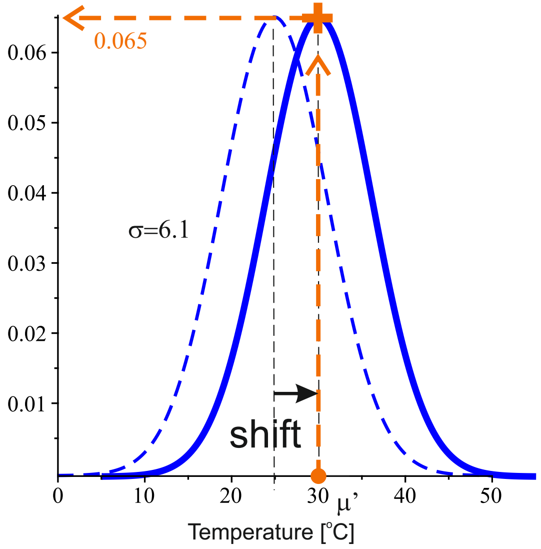
likelihood



The difference between the probability and likelihood (2)

- Let us recall from the previous slide: We assume the Gaussian distribution of errors while measuring temperature using a particular thermometer with $\mu = 25°\mathrm{C}$ and $\sigma = 6.1°\mathrm{C}$.
- We measured the temperature $30°\mathrm{C}$. The corresponding likelihood was estimated as, cf. figure left, $L = p(\text{Gaussian}, \mu = 25, \sigma = 6.1 \,|\, \text{temperature} = 30°\mathrm{C}) = 0.048$.
- If we shifted the distribution over that $\mu' = 30$, cf. figure right, the new likelihood would be 0.065. The value on the right side of the class-conditional probability $p(x|y)$ is fixed.





Class-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 class-conditional probabilities $p_{X|Y}(x \,|\, y)$ have the form

$$p_{X|Y}(x \,|\, y) = \prod_{i=1}^{n} p_{X_i|Y}(x_i \,|\, y)\ . \tag{1}$$

- Features become mutually independent at the fixed state y.
- The object features x_i, $i \in I$, are dependent on each other but all the dependence is realized via the dependence on the state of the object y in the formula (1). If the state is fixed then the mutual dependence among the features disappears.
- This is the simplest model of the class-conditional independence.

Conditional independence of features (2)

- However, the class-conditional independence assumption (1) from slide 5, repeated here,

$$p_{X|Y}(x \mid y) = \prod_{i=1}^{n} p_{X_i|Y}(x_i \mid y) \, .$$

  does not mean that the features are also *a priori* mutually independent.

- In general,

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

Naïve Bayes classifier

- The assumption is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naïve.
- Types of Naïve Bayes classifiers:
  - *Multinomial Naïve Bayes classifiers* - corresponds to already introduced class-conditional independence.
  - *Bernoulli Naïve Bayes classifiers* - This is similar to the multinomial naïve Bayes classifier but the predictors are boolean variables. The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.
  - *Gaussian Naïve Bayes classifiers* - We assume that values are sampled from a Gaussian distribution.

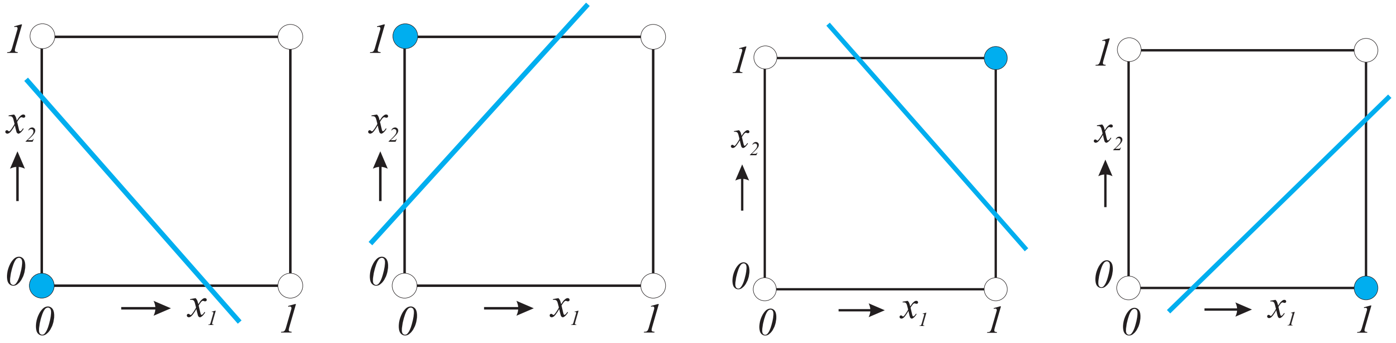
Courtesy: Rohith Gandhi

The simplest case: $\{0, 1\}$ features only, 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, $y_1 = 1$ or $y_2 = 2$.

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: $\{0, 1\}$ features only, two states only (2)

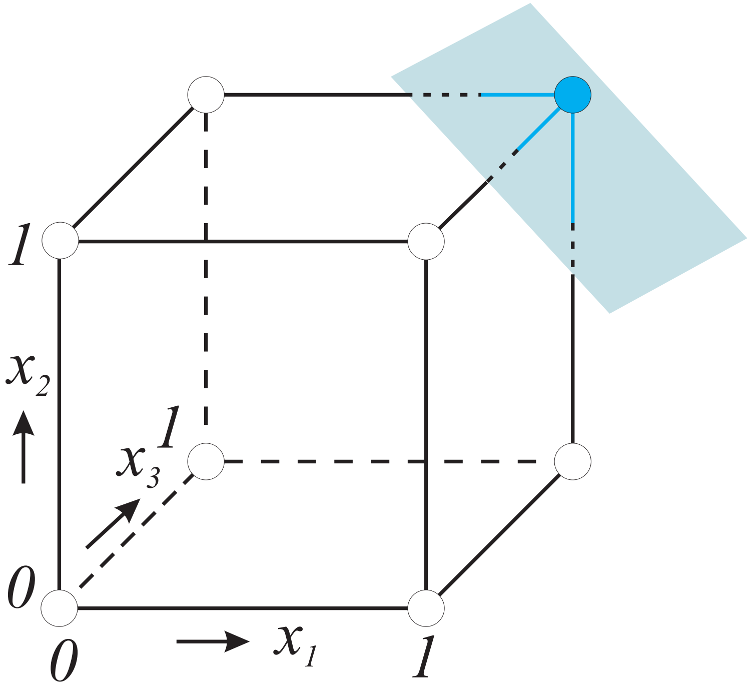


Illustration for n=3. We show only one of eight possible cases.

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 the likelihood ratio corresponds to each decision d, i.e., the decision d is taken for which

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

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

- Inequality does not change if a monotonic function as $\log$ is applied,

$$\theta'_{\min} < \log \frac{p_{X|Y}(x \,|\, y=1)}{p_{X|Y}(x \,|\, y=2)} \leq \theta'_{\max} \,, \;\; \theta'_{\min} = \log \theta_{\min} \,, \;\; \theta'_{\max} = \log \theta_{\max}$$

More than two features

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

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

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

Linearity

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

We can rewrite the above expression to

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

- If the tasks are expressed by a firmly chosen function $p_{X|Y}$ then various strategies differ only by a threshold value θ.
- In addition, if the function $p_{X|Y}$ 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

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

- 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 continuous (infinite) case.
- It is sufficient to mention that the number $p_{X|Y}(x \mid y)$ does not mean a probability but a probability density.

Multi-dimensional Gaussian random variable

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

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

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

Special case (1): two hidden states, two decisions

The optimal decision strategy is a quadratic decision function

$$x \in \begin{cases} X_1, & \text{if} \quad \sum_i \sum_j \alpha_{ij}\, x_i\, x_j + \sum_i \beta_i\, x_i \leq \gamma\,, \\ X_2, & \text{if} \quad \sum_i \sum_j \alpha_{ij}\, x_i\, x_j + \sum_i \beta_i\, x_i > \gamma\,. \end{cases}$$

- 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): two hidden states, two decisions

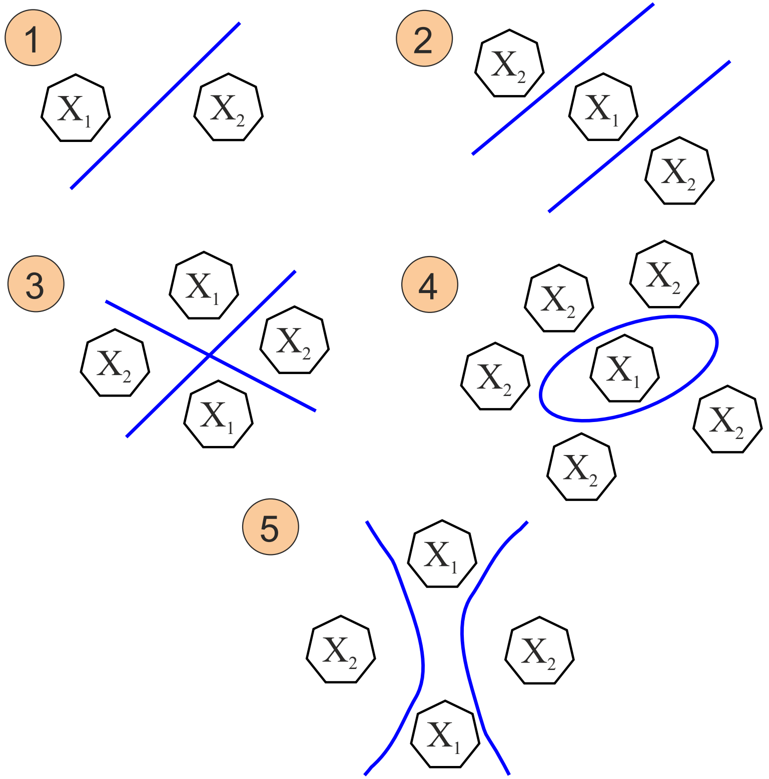
Coefficients α_{ij}, β_i, $i, j = 1, 2, \ldots, m$, the threshold γ depend

- on a statistical model of the object, i.e., on the matrices A^1, A^2 and on vectors μ^1, μ^2
- on the Bayesian or non-Bayesian decision problem to be solved.

Special case (3): two hidden states, two 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.



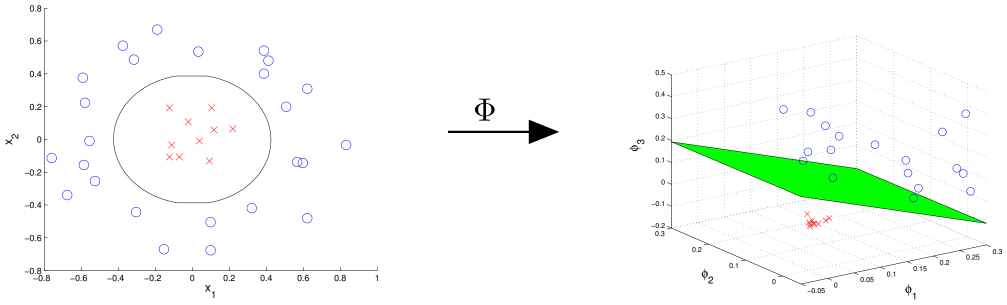
Straightening of the feature space (1)

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 $\Phi\colon \mathcal{X} \to \mathcal{F}$, 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.

$$x \in \begin{cases} X_1, & \text{if} \quad \sum\limits_i \alpha_i \, x_i \le \gamma \, , \\[2ex] X_2, & \text{if} \quad \sum\limits_i \alpha_i \, x_i > \gamma \, . \end{cases}$$

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 $x = (x_1, x_2, \dots, x_i, \dots, x_n)$ are transformed to

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

New linear decision rule $$x' \in \begin{cases} X'_1, & \text{if} \quad \sum_i \alpha_i\, x'_i \leq \gamma\,, \\ X'_2, & \text{if} \quad \sum_i \alpha_i\, x'_i > \gamma\,. \end{cases}$$

Straightening, a 1D example

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

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

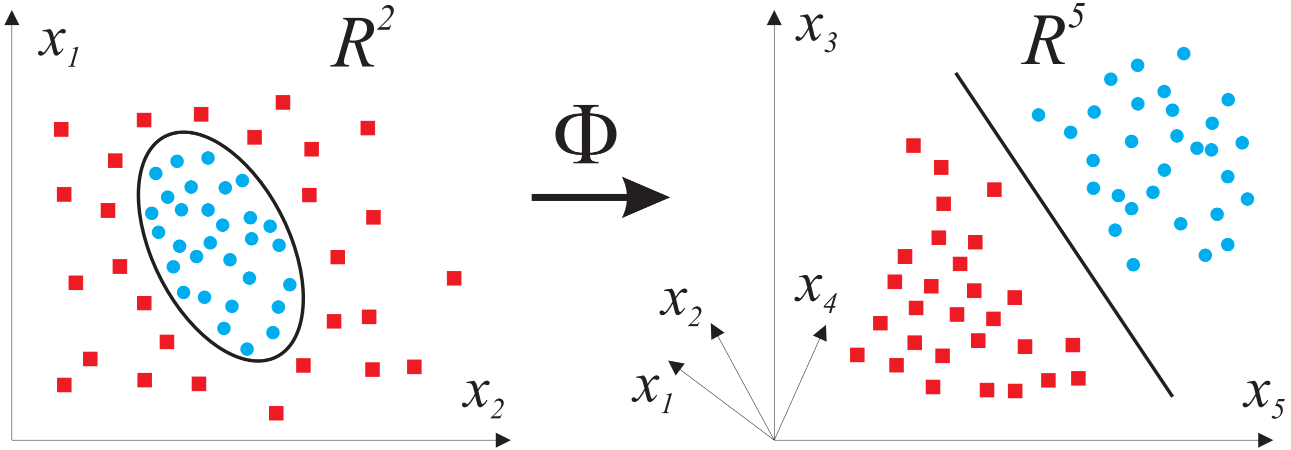
Straightening $\quad x'_1 = x^2, \quad x'_2 = x.$

$$x \in \begin{cases} X_1, & \text{if} \quad \alpha_1\, x'_1 + \alpha_2\, x'_2 > \theta \,, \\ X_2, & \text{if} \quad \alpha_1\, x'_1 + \alpha_2\, x'_2 \leq \theta \,, \end{cases}$$

where $x'_1 = x^2$, $x'_2 = x$, $\alpha_1 = -1$, $\alpha_2 = 2\, x_0$, $\theta = x_0^2 - \delta$.

Straightening, a 2D example

- Assume 2D feature space and a quadratic decision strategy, i.e., $q(x)$ is a polynomial of degree two.
- Mapping functions are: $\Phi_1 = x_1$, $\Phi_2 = x_2$, $\Phi_3 = x_1 x_2$, $\Phi_4 = x_1 x_1$, $\Phi_5 = x_2 x_2$.
- $q(x) = w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2 = w^\top \Phi_i(x)$.


x_1
R^2
x_2
Φ
x_3
R^5
x_2
x_4
x_1
x_5