Practical Machine Learning

Support Vector Machine

SUPPORT VECTOR MACHINE (SVM)

OUTLINE

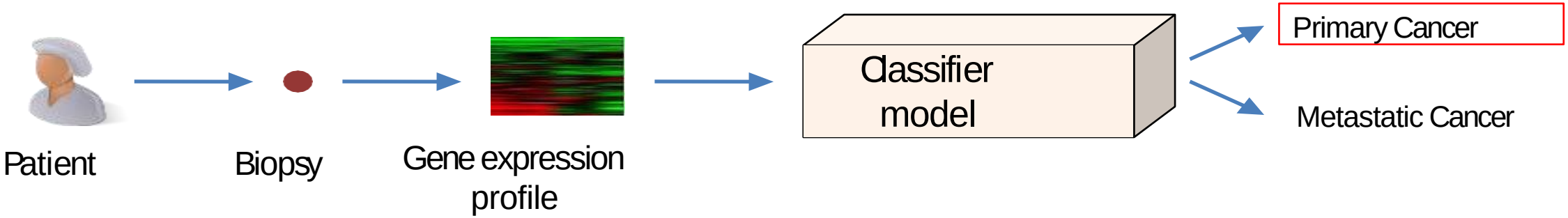
SUPPORT VECTOR MACHINES

- Principles of SVM
- Linear & Non-Linear Classification (Radial and Polynomial)
- Types of SVM
  1. Support Vector Machine for Classification
  2. Support Vector Regression
  3. SVM for outlier detection
  4. SVM for variable selection

Types of Learning Problems

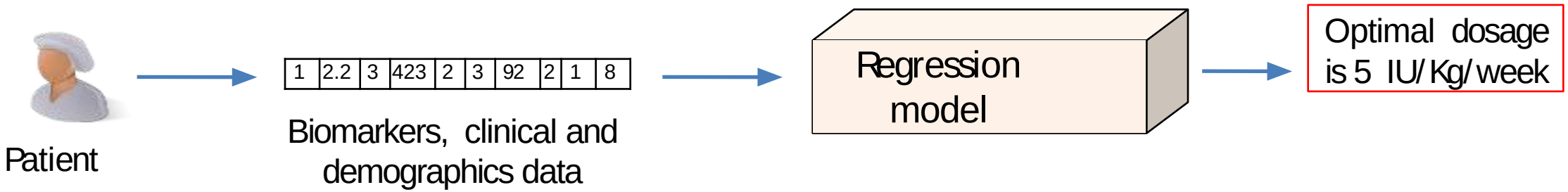
DATA-ANALYSIS PROBLEMS OF INTEREST

1. **Problem1** – **Classification -** Build computational classification models (or "*classifiers*") that assign patients/samples into two or more classes.
   - Classifiers can be used for diagnosis, outcome prediction, and other classification tasks.
   - E.g., build a decision-support system to diagnose primary and metastatic cancers from gene expression profiles of the patients:



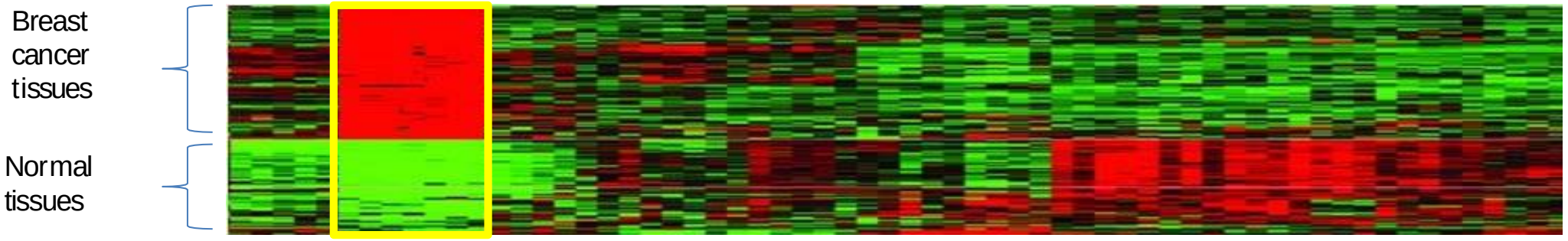
DATA-ANALYSIS PROBLEMS OF INTEREST

2. **Problem 2 – Regression -** Build computational regression models to predict values of some continuous response variable or outcome.
   - Regression models can be used to predict survival, length of stay in the hospital, laboratory test values, etc.
   - E.g., build a decision-support system to predict optimal dosage of the drug to be administered to the patient. This dosage is determined by the values of patient biomarkers, and clinical and demographics data:



DATA-ANALYSIS PROBLEMS OF INTEREST

3. **Problem 3 – Feature Selection** - Out of all measured variables in the dataset, select the smallest subset of variables that is necessary for the most accurate prediction (classification or regression) of some variable of interest (e.g., phenotypic response variable).
   - E.g., find the most compact panel of breast cancer biomarkers from microarray gene expression data for 20,000 genes:

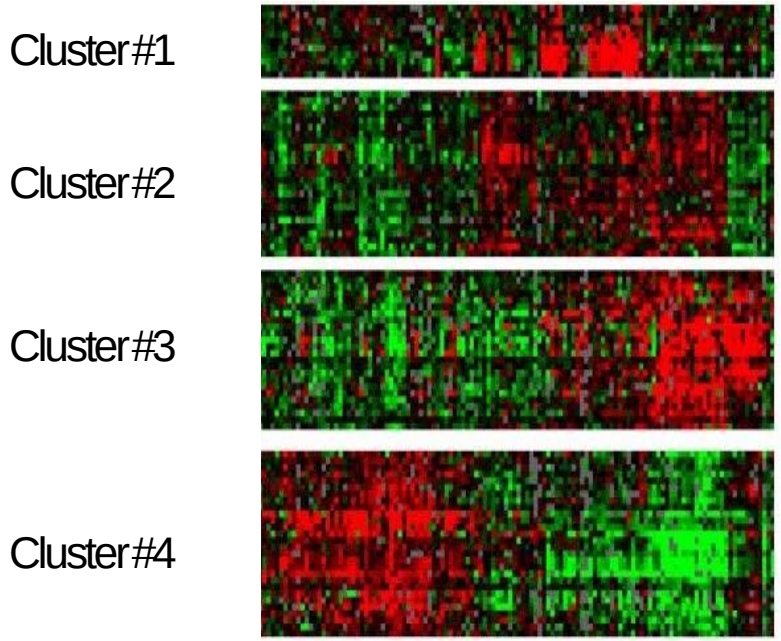


DATA-ANALYSIS PROBLEMS OF INTEREST

4. **Problem 4 – Outlier Detection -** Build a computational model to identify novel or outlier patients/samples.
   - Such models can be used to discover deviations in sample handling protocol when doing quality control of assays, etc.
   - E.g., build a decision-support system to identify aliens.

DATA-ANALYSIS PROBLEMS OF INTEREST

5. **Problem 5 – Clustering -** Group patients/samples into several clusters based on their similarity.
   - These methods can be used to discovery disease sub-types and for other tasks.
   - E.g., consider clustering of brain tumor patients into 4 clusters based on their gene expression profiles. All patients have the same pathological sub-type of the disease, and clustering discovers new disease subtypes that happen to have different characteristics in terms of patient survival and time to recurrence after treatment.



BASIC CONCEPTS OF SVM

BASIC PRINCIPLES OF CLASSIFICATION

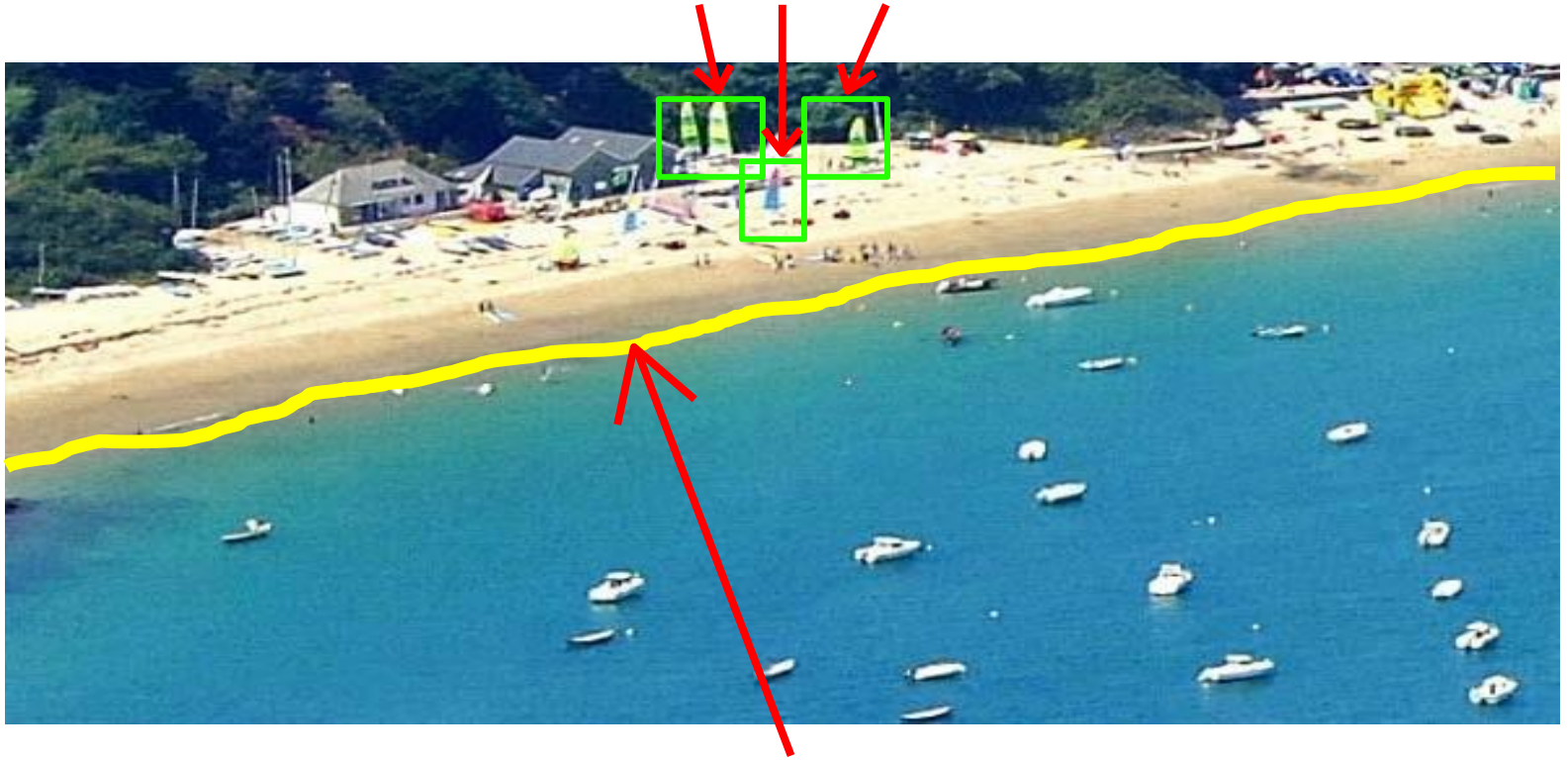
Want to classify objects as boats and houses.

BASIC PRINCIPLES OF CLASSIFICATION

- All objects before the coast line are boats and all objects after the coast line are houses.
- Coast line serves as a *decision surface* that separates two classes.

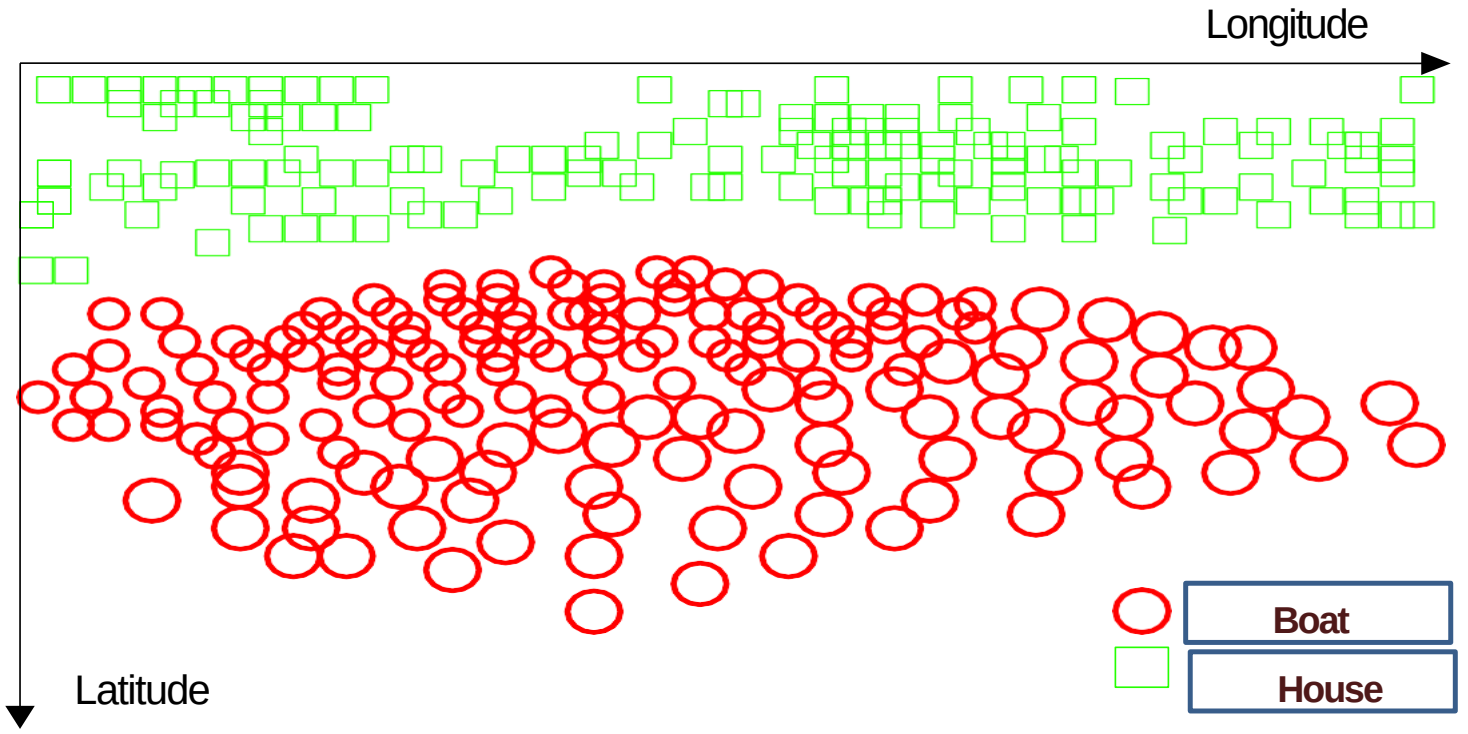
Basic principles of classification

These boats will be misclassified as houses



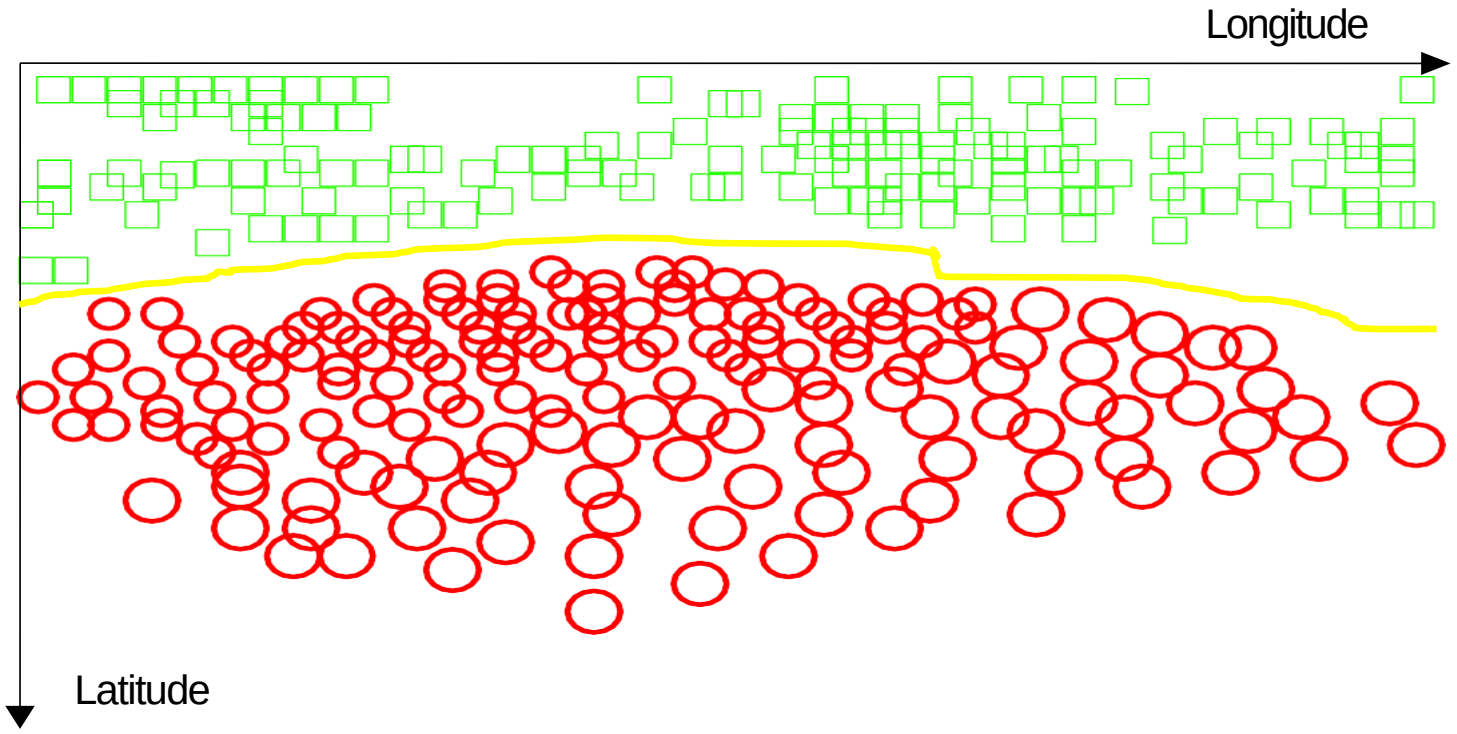
Let me draw a decision boundary…(Yellow line)

BASIC PRINCIPLES OF CLASSIFICATION



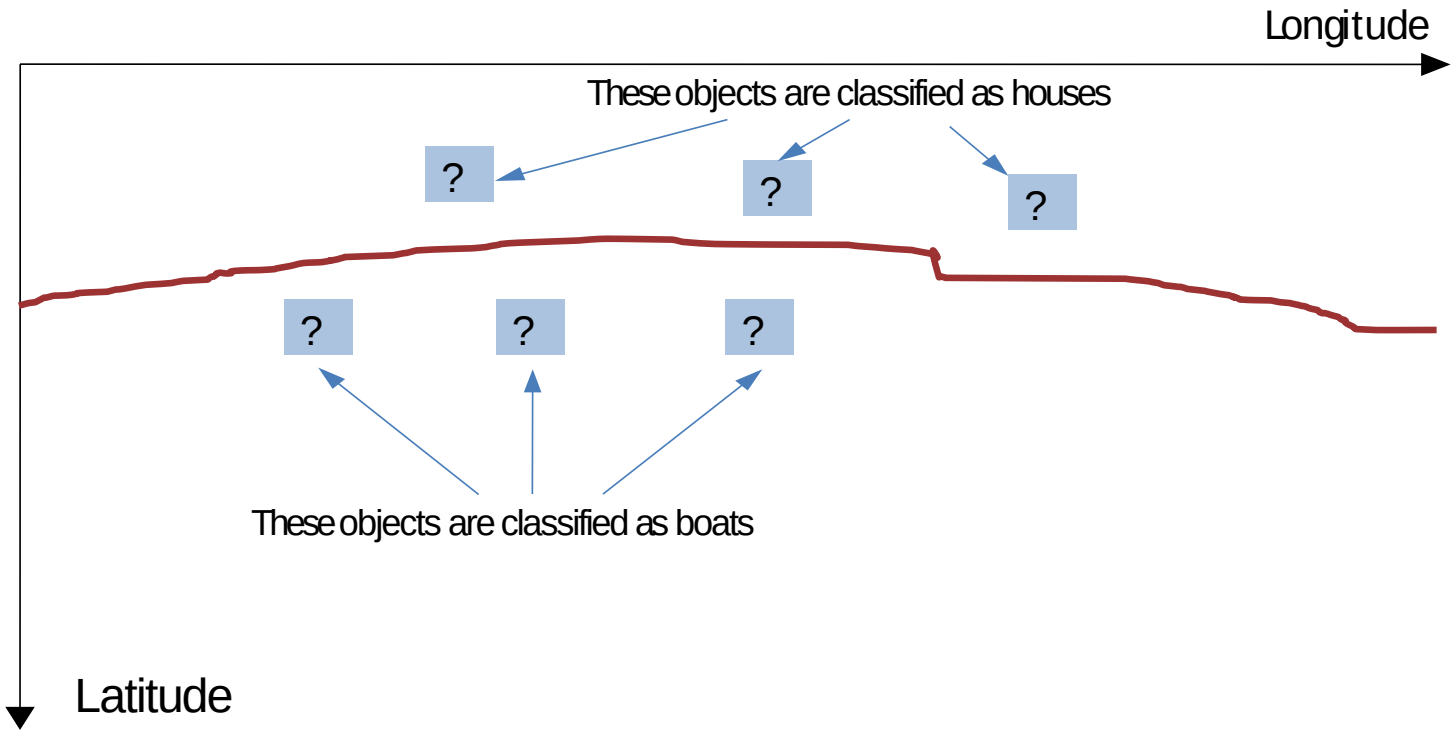
- The methods that build classification models (i.e., "*classification algorithms*") operate very similarly to the previous example.
- First all objects are represented geometrically.

BASIC PRINCIPLES OF CLASSIFICATION



Then the algorithm seeks to find a decision surface that separates classes of objects

BASIC PRINCIPLES OF CLASSIFICATION



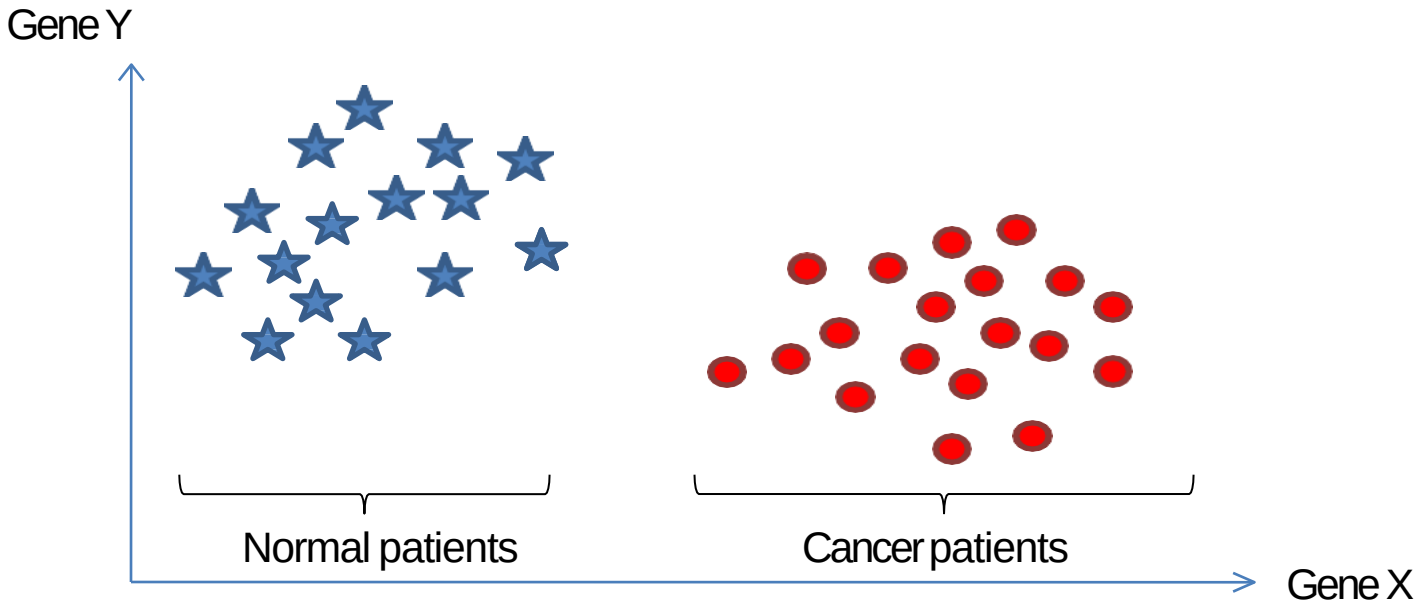
Unseen (new) objects are classified as “boats” if they fall below the decision surface and as “houses” if the fall above it

SVM APPROACH

- Basically, Support vector machines (SVMs) is a binary classification algorithm that offers a solution to **problem #1**.
- Extensions of the basic SVM algorithm can be applied to solve problems #2 to #5.
- SVMs are important because of
  - (a) theoretical reasons:
    - Robust to very large number of variables and small samples
    - Can learn both simple and highly complex classification models
    - Employ sophisticated mathematical principles to avoid overfitting

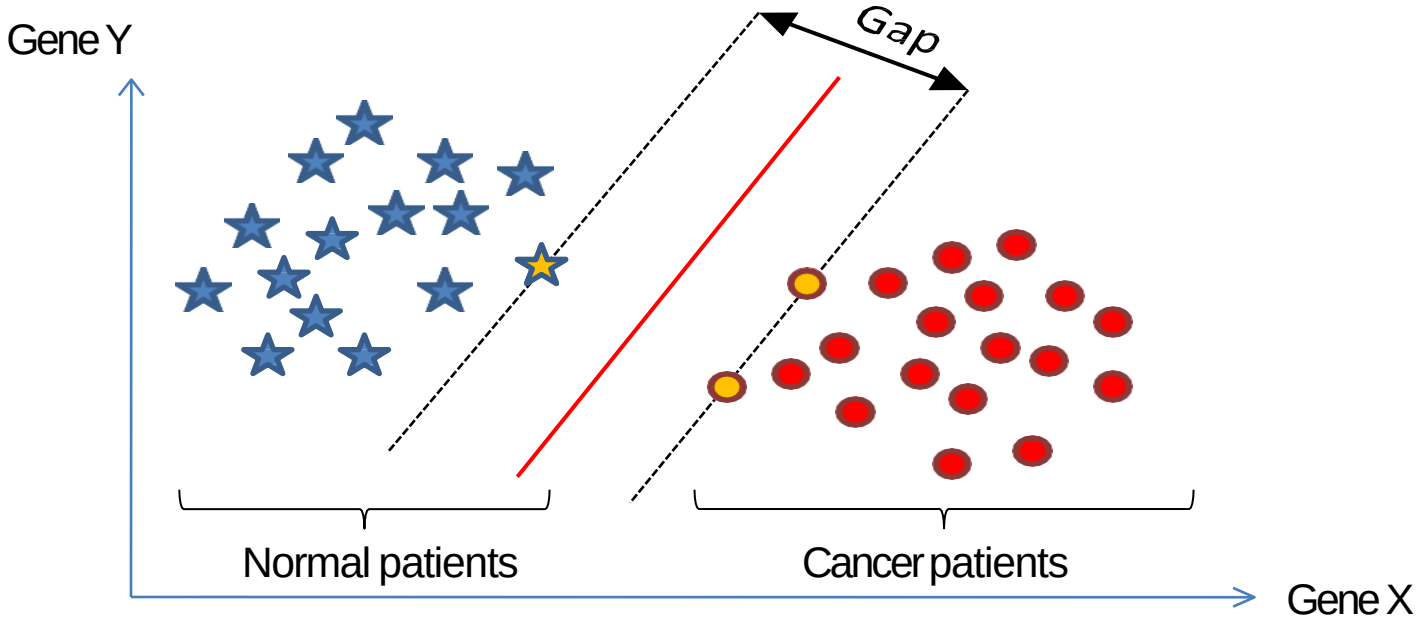
  and (b) superior empirical results.

MAIN IDEAS OF SVMS



- Consider example dataset described by 2 genes, gene X and gene Y
- Represent patients geometrically (by “vectors”)

MAIN IDEAS OF SVMS



Find a linear **decision surface** ("hyperplane") that can separate patient classes

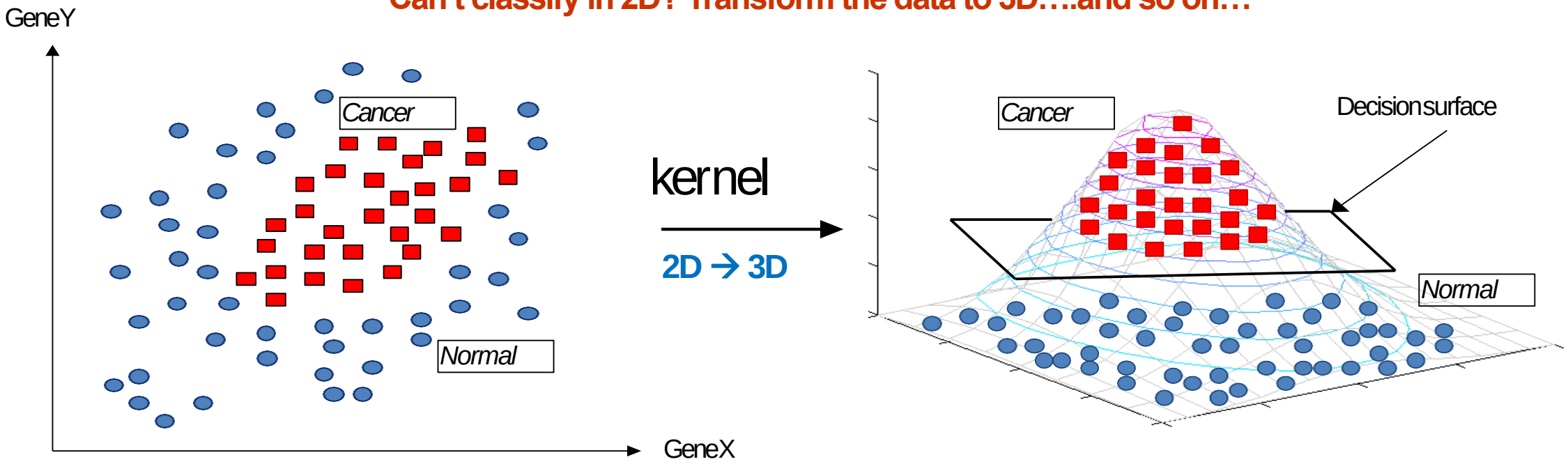
and

has the largest distance (i.e., **largest "gap"** or "margin") between

border-line patients (i.e., **"support vectors"**);

MAIN IDEAS OF SVMS

Can't classify in 2D? Transform the data to 3D….and so on…



- If such linear decision surface does not exist, the data is mapped into a much **higher dimensional space** ("feature space") where the separating decision surface is found;

- The feature space is constructed via very clever mathematical projection → **"kernel trick"**

MATHEMATICAL CONCEPTS

GEOMETRIC REPRESENTATION -VECTORS IN N-DIMENSIONAL SPACE (R^N)

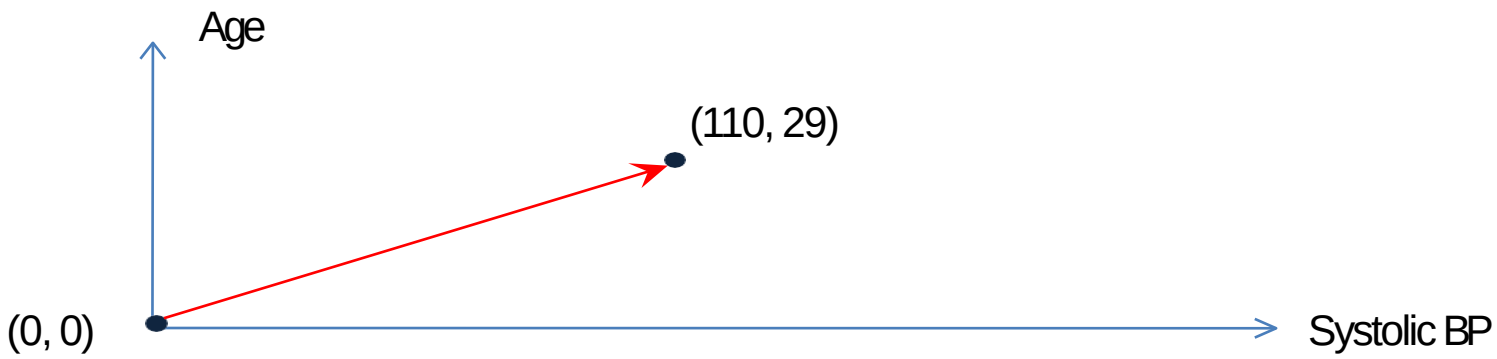
- Assume that a sample/patient is described by *n* characteristics ("features" or "variables")

- **Representation:** Every sample/patient is a vector in R^n with tail at point with 0 coordinates and arrow-head at point with the feature values.

- **Example:** Consider a patient described by 2 features:

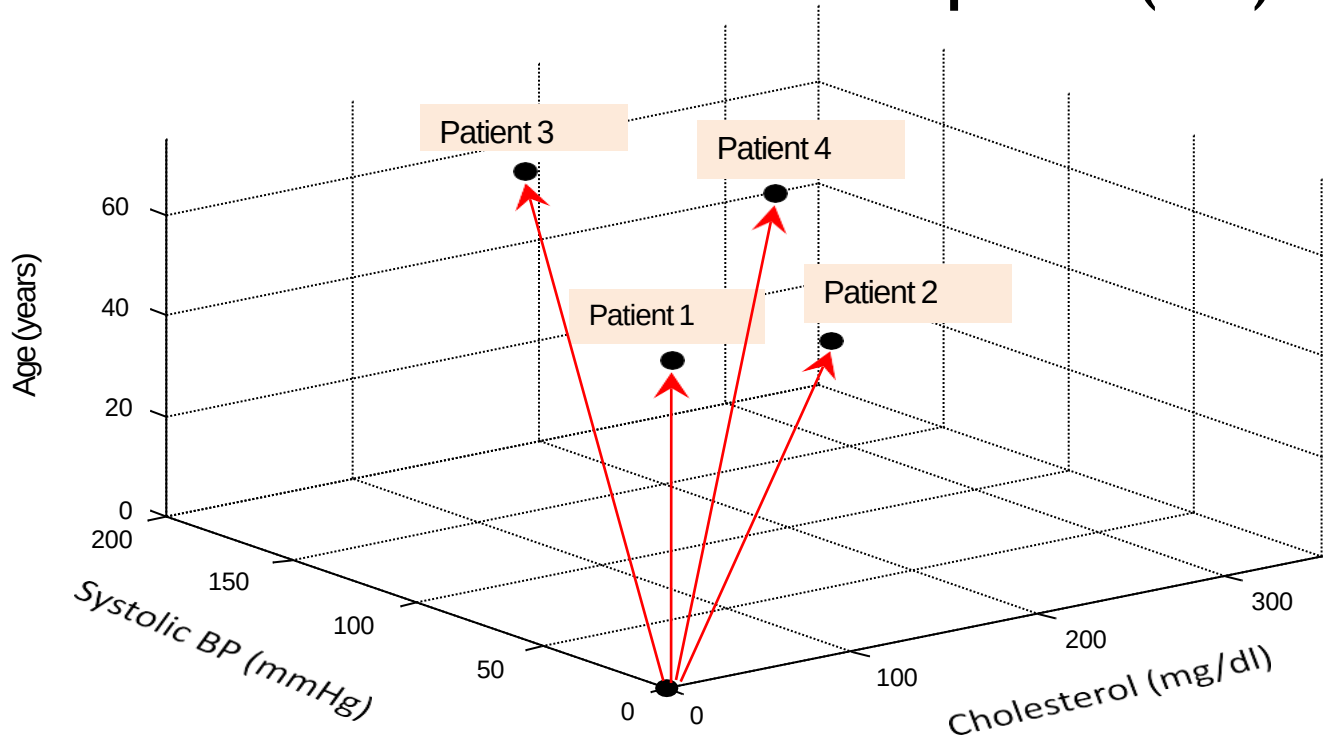
  Systolic BP = 110 and *Age* = 29.

  This patient can be represented as a vector in R^2:



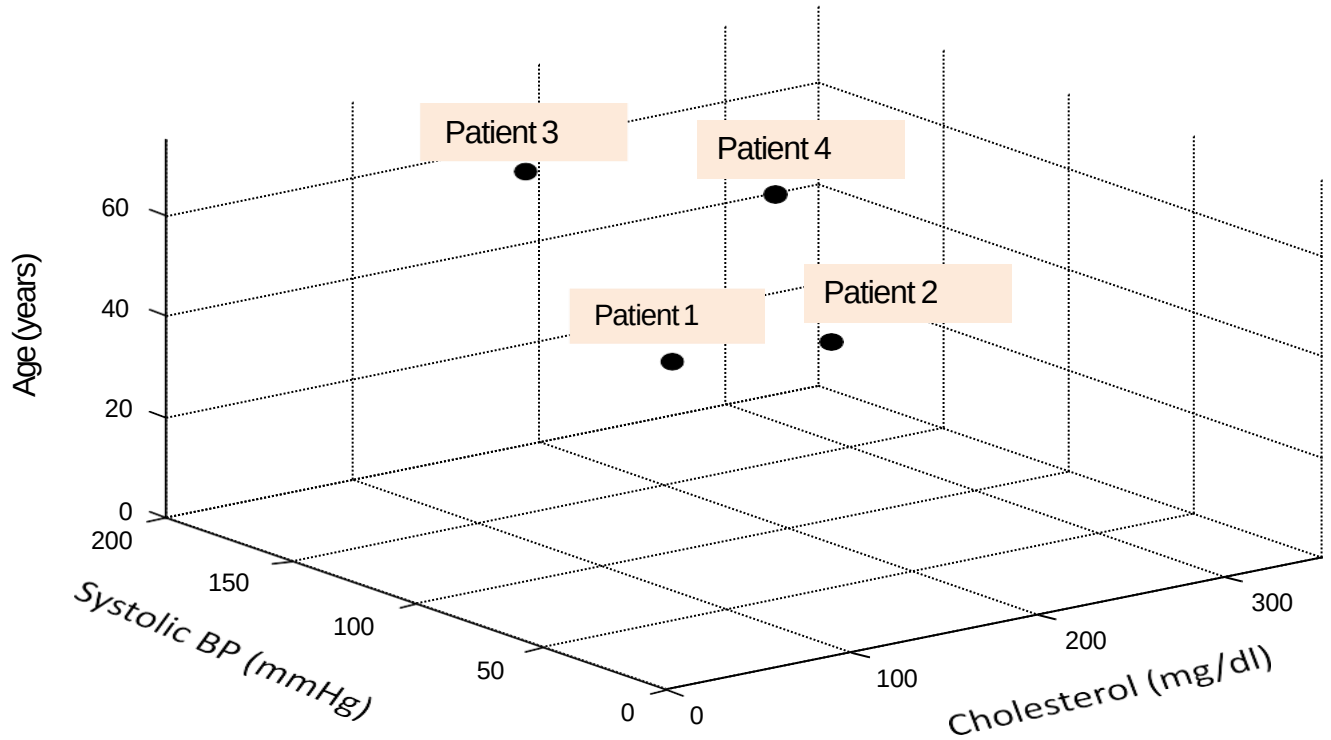
HOW TO REPRESENT SAMPLES GEOMETRICALLY?

Vectors in n-dimensional space (R^n)



Patient id	Cholesterol (mg/dl)	Systolic BP (mmHg)	Age (years)	Tail of the vector	Arrow-head of the vector
1	150	110	35	(0,0,0)	(150, 110, 35)
2	250	120	30	(0,0,0)	(250, 120, 30)
3	140	160	65	(0,0,0)	(140, 160, 65)
4	300	180	45	(0,0,0)	(300, 180, 45)

VECTORS IN N-DIMENSIONAL SPACE ($\mathbb{R}^N$)

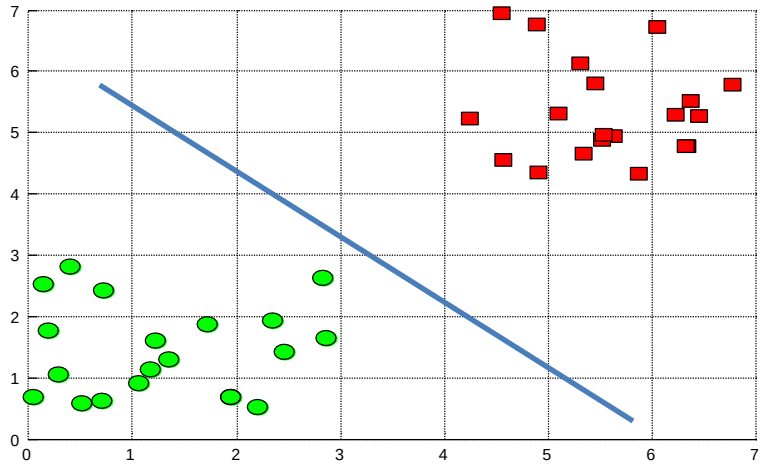


Since we assume that the tail of each vector is at point with 0 coordinates, we will also depict vectors as points (where the arrow-head is pointing).

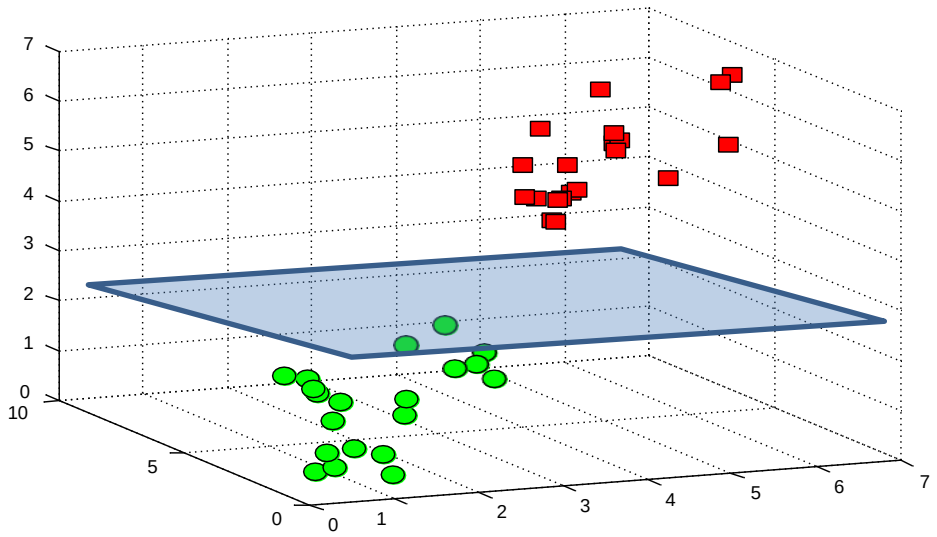
PURPOSE OF VECTOR REPRESENTATION

- Having represented each sample/patient as a vector allows now to geometrically represent the decision surface that separates two groups of samples/patients.

A decision surface in R^2



A decision surface in R^3

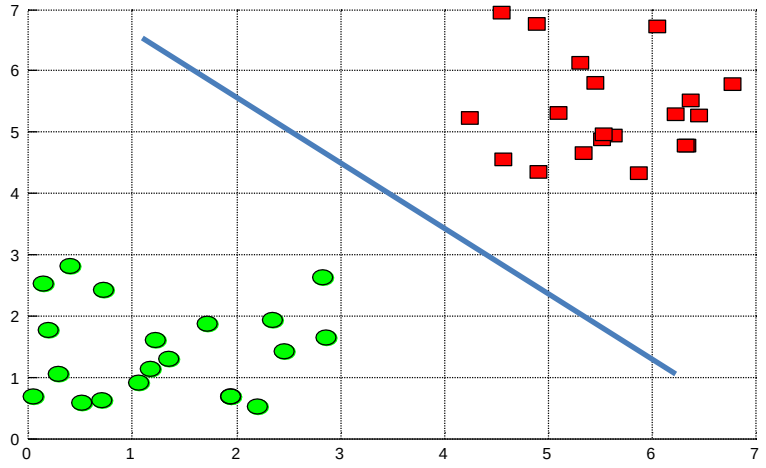


- In order to define the decision surface, we need to introduce some basic math elements…

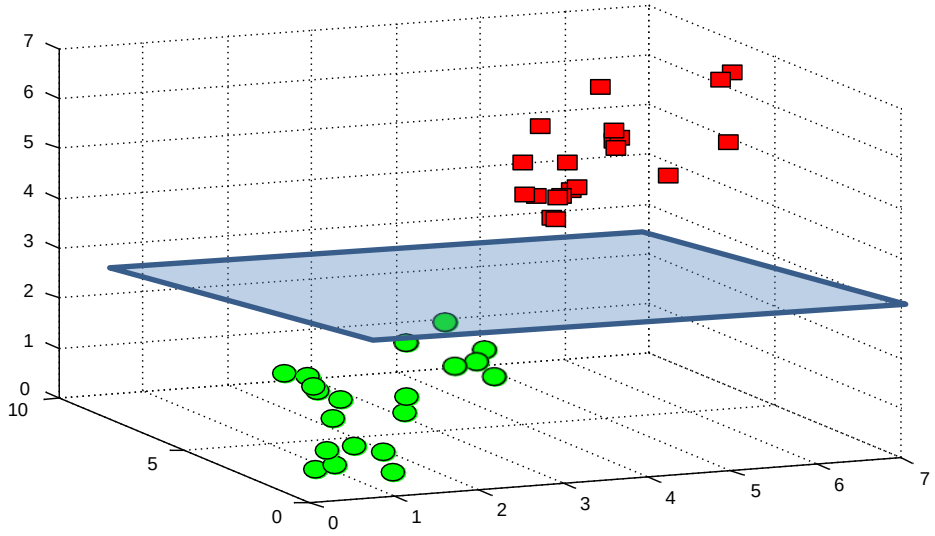
HYPERPLANES AS DECISION SURFACES

- A hyperplane is a linear decision surface that splits the space into two parts;
- It is obvious that a hyperplane is a binary classifier.

A hyperplane in $\mathbb{R}^2$ is a line



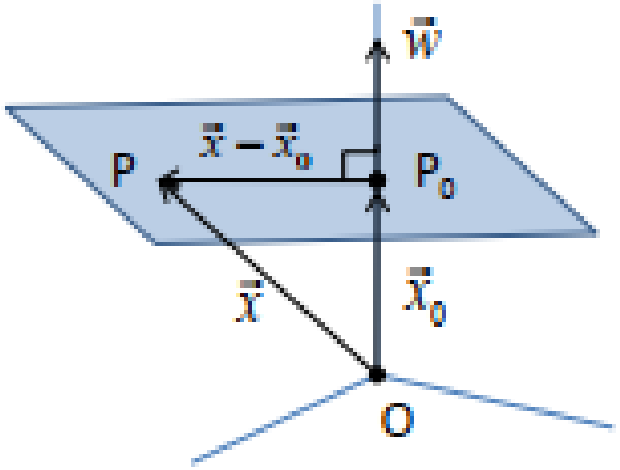
A hyperplane in $\mathbb{R}^3$ is a plane



A hyperplane in $\mathbb{R}^n$ is an *n*-1 dimensional subspace

EQUATION OF A HYPERPLANE

Consider the case of $\mathbb{R}^3$:



An equation of a hyperplane is defined by a point (P_0) and a perpendicular vector to the plane ($\vec{w}$) at that point.

Define vectors: $\vec{x}_0 = \overrightarrow{OP_0}$ and $\vec{x} = \overrightarrow{OP}$, where *P* is an arbitrary point on a hyperplane.

A condition for *P* to be on the plane is that the vector $\vec{x} - \vec{x}_0$ is perpendicular to $\vec{w}$:

$$\vec{w} \cdot (\vec{x} - \vec{x}_0) = 0 \quad \text{or}$$

$$\vec{w} \cdot \vec{x} - \vec{w} \cdot \vec{x}_0 = 0 \quad \text{define } b = -\vec{w} \cdot \vec{x}_0$$

$$\boxed{\vec{w} \cdot \vec{x} + b = 0}$$

The above equations also hold for $\mathbb{R}^n$ when n>3.

EQUATION OF A HYPERPLANE

Example

$\vec{w} = (4,-1,6)$

$P_0 = (0,1,-7)$

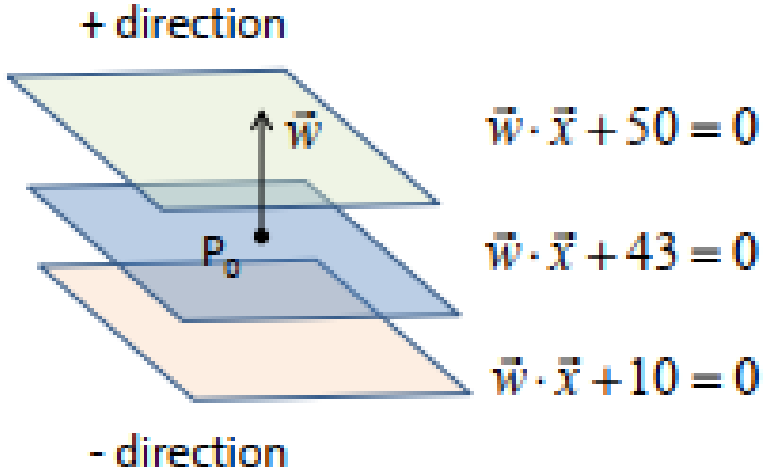
$b = -\vec{w} \cdot \vec{x}_0 = -(0-1-42) = 43$

$\Rightarrow \vec{w} \cdot \vec{x} + 43 = 0$

$\Rightarrow (4,-1,6) \cdot \vec{x} + 43 = 0$

$\Rightarrow (4,-1,6) \cdot (x_{(1)}, x_{(2)}, x_{(3)}) + 43 = 0$

$\Rightarrow 4x_{(1)} - x_{(2)} + 6x_{(3)} + 43 = 0$



What happens if the *b* coefficient changes?
The hyperplane moves along the direction of $\vec{w}$.
We obtain "parallel hyperplanes".

Distance between two parallel hyperplanes $\vec{w} \cdot \vec{x} + b_1 = 0$ and $\vec{w} \cdot \vec{x} + b_2 = 0$ is equal to $D = |b_1 - b_2| / \|\vec{w}\|$.

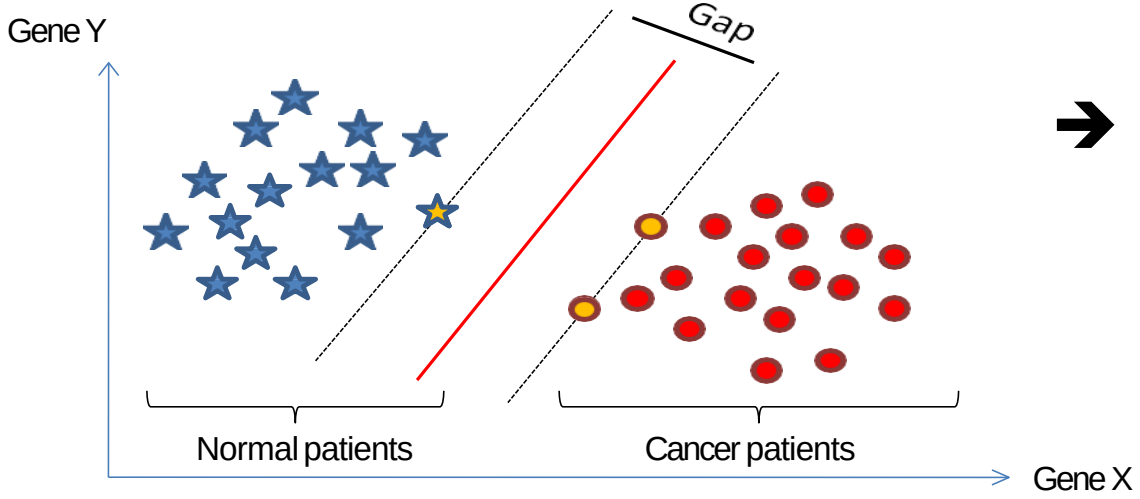
RECAP

We know…

- How to represent patients (as "vectors")
- How to define a linear decision surface ("hyperplane")

We need to know…

- How to efficiently compute the hyperplane that separates two classes with the largest "gap"?

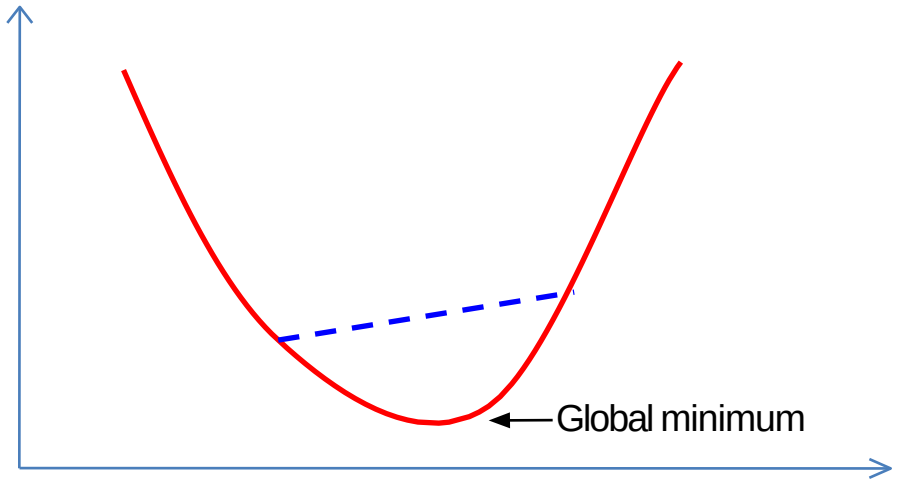


➔ Need to introduce basics of relevant optimization theory

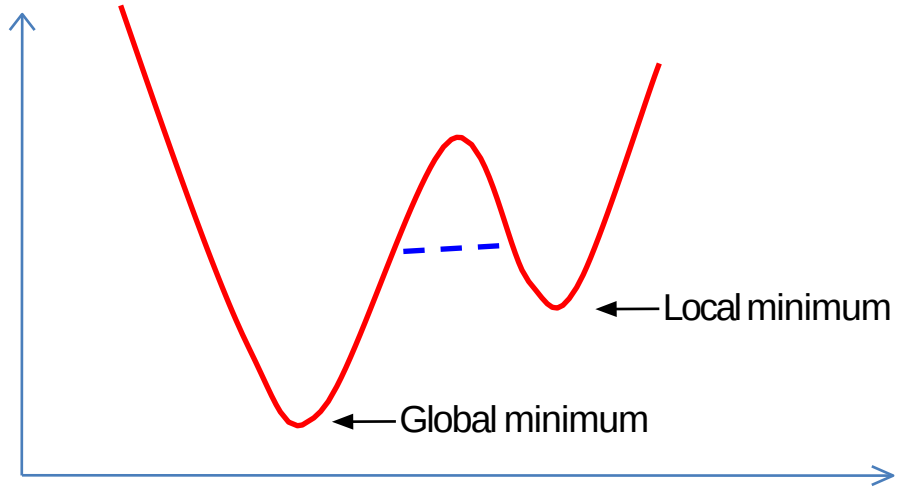
BASICS OF OPTIMIZATION: CONVEX FUNCTIONS

- **Convex Function:** A function is called *convex* if the function lies below the straight line segment connecting two points, for any two points in the interval.

- **Property:** Any local minimum is a global minimum!



Convex function



Non-convex function

BASICS OF OPTIMIZATION: EXAMPLE QP PROBLEM

Consider $\vec{x} = (x_1, x_2)$

Minimize $\underbrace{\frac{1}{2}\| \vec{x} \|_2^2}$ subject to $\underbrace{x_1 + x_2 - 1 \geq 0}$

quadratic objective *linear constraints*

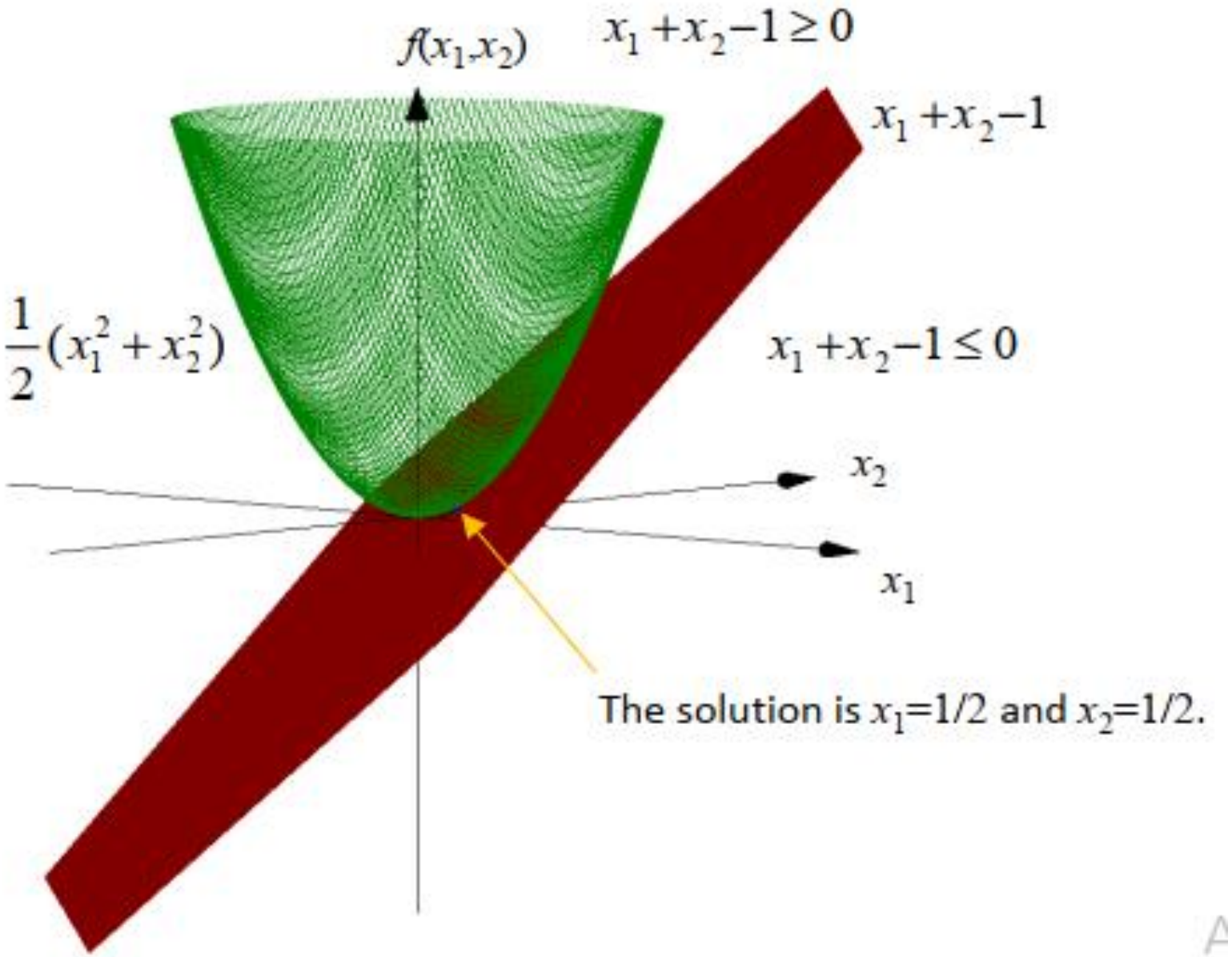
This is QP problem, and it is a convex QP as we will see later

We can rewrite it as:

Minimize $\underbrace{\frac{1}{2}(x_1^2 + x_2^2)}$ subject to $\underbrace{x_1 + x_2 - 1 \geq 0}$

quadratic objective *linear constraints*

BASICS OF OPTIMIZATION: EXAMPLE QP



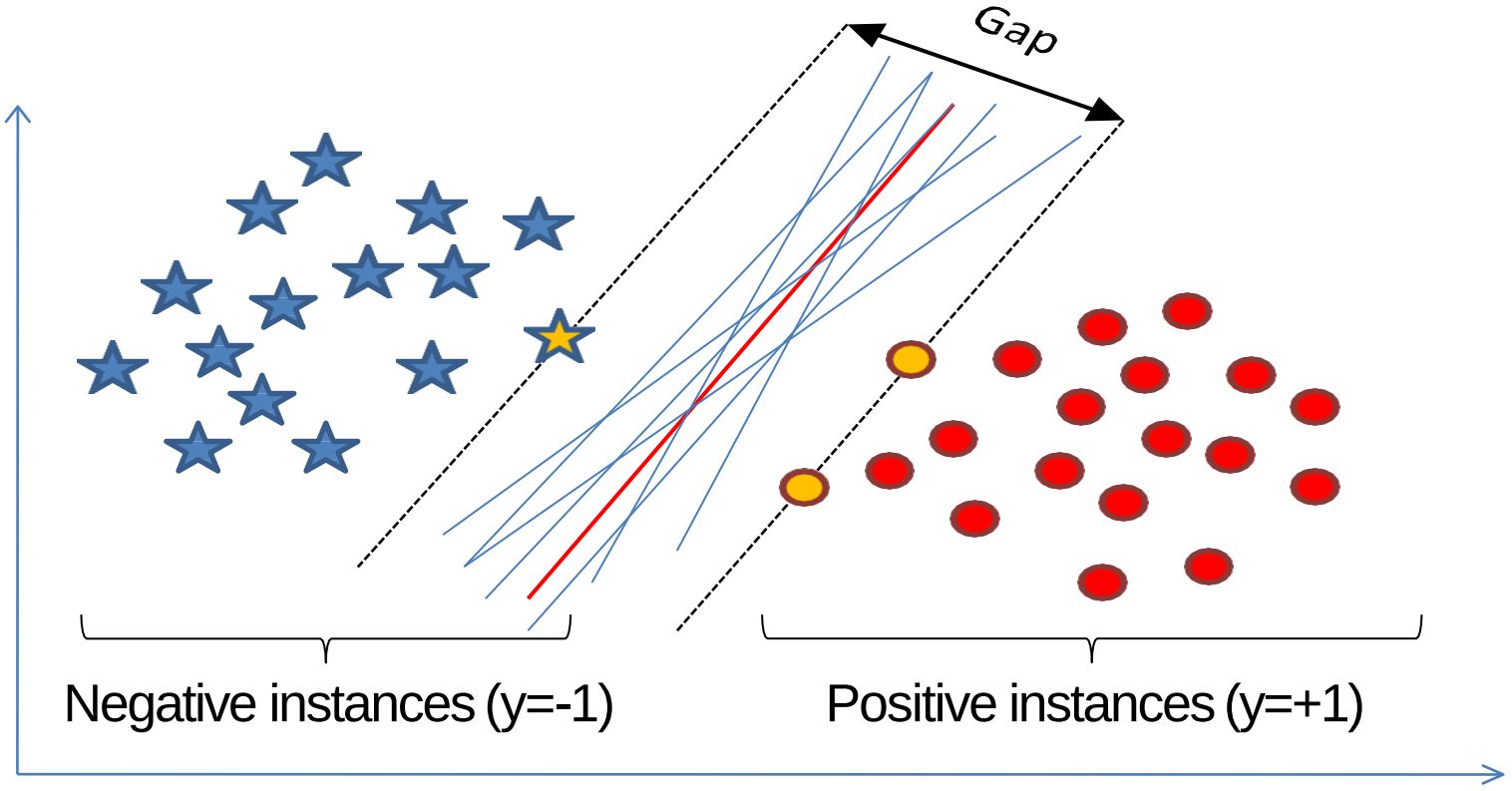
SVM FOR BINARY CLASSIFICATION

CASE 1: LINEARLY SEPARABLE DATA; "HARD-MARGIN" LINEAR SVM

$$\vec{x}_1, \vec{x}_2, ..., \vec{x}_N \in R^n$$

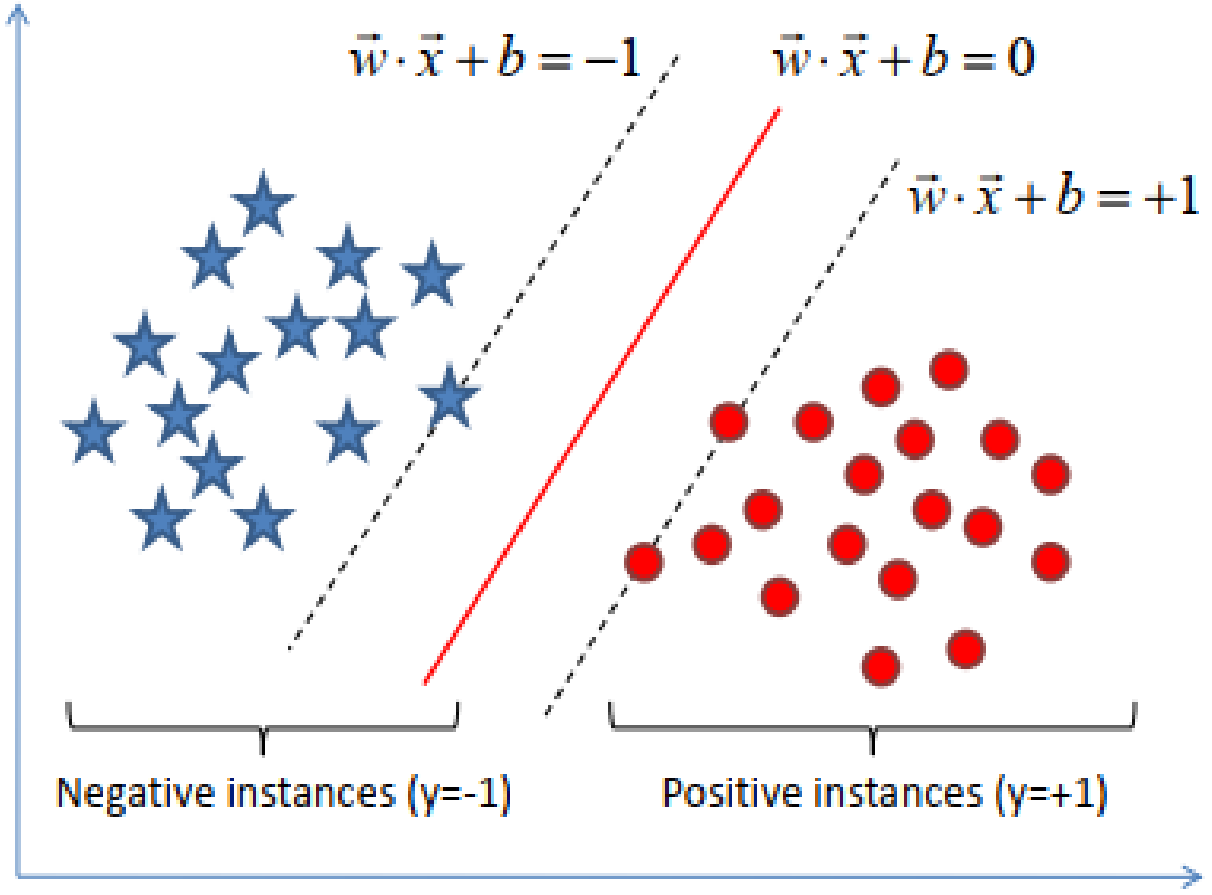
Given training data:

$$y_1, y_2, ..., y_N \in \{-1, +1\}$$



- Want to find a classifier (hyperplane) to separate negative instances from the positive ones.
- An infinite number of such hyperplanes exist.
- SVMs finds the hyperplane that maximizes the gap between data points on the boundaries (so-called "support vectors").
- If the points on the boundaries are not informative (e.g., due to noise), SVMs will not do well.

STATEMENT OF LINEAR SVM CLASSIFIER



The gap is distance between parallel hyperplanes:

$\vec{w}\cdot\vec{x}+b=-1$ and
$\vec{w}\cdot\vec{x}+b=+1$

Or equivalently:

$\vec{w}\cdot\vec{x}+(b+1)=0$
$\vec{w}\cdot\vec{x}+(b-1)=0$

We know that

$D=|b_1-b_2|/\|\vec{w}\|$

Therefore:

$D=2/\|\vec{w}\|$

b1 = b+1
b2 = b-1

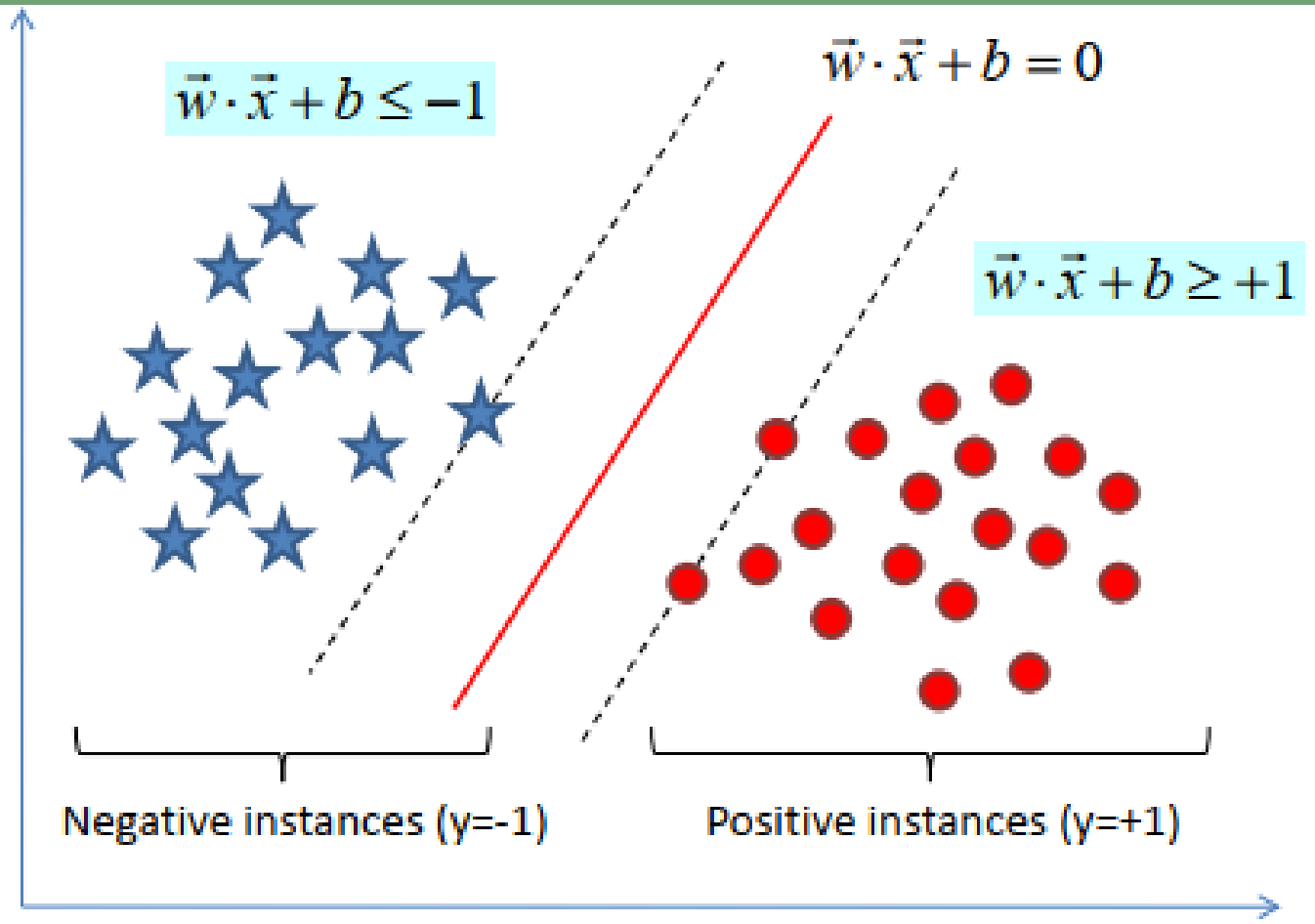
b1-b2
= (b+1) – (b-1)
= 2

Since we want to maximize the gap,

we need to minimize $\|\vec{w}\|$

or equivalently minimize $\frac{1}{2}\|\vec{w}\|^2$ ($\frac{1}{2}$ is convenient for taking derivative later on)

STATEMENT OF LINEAR SVM CLASSIFIER…



In addition we need to impose constraints that all instances are correctly classified. In our case:

$\vec{w}\cdot\vec{x}_i+b\leq -1$ if $y_i=-1$

$\vec{w}\cdot\vec{x}_i+b\geq +1$ if $y_i=+1$

Equivalently:

$$y_i(\vec{w}\cdot\vec{x}_i+b)\geq 1$$

In summary:

Want to minimize $\frac{1}{2}\|\vec{w}\|^2$ subject to $y_i(\vec{w}\cdot\vec{x}_i+b)\geq 1$ for $i=1,\ldots,N$

Then given a new instance x, the classifier is $f(\vec{x})=sign(\vec{w}\cdot\vec{x}+b)$

SVM OPTIMIZATION PROBLEM: PRIMAL FORMULATION

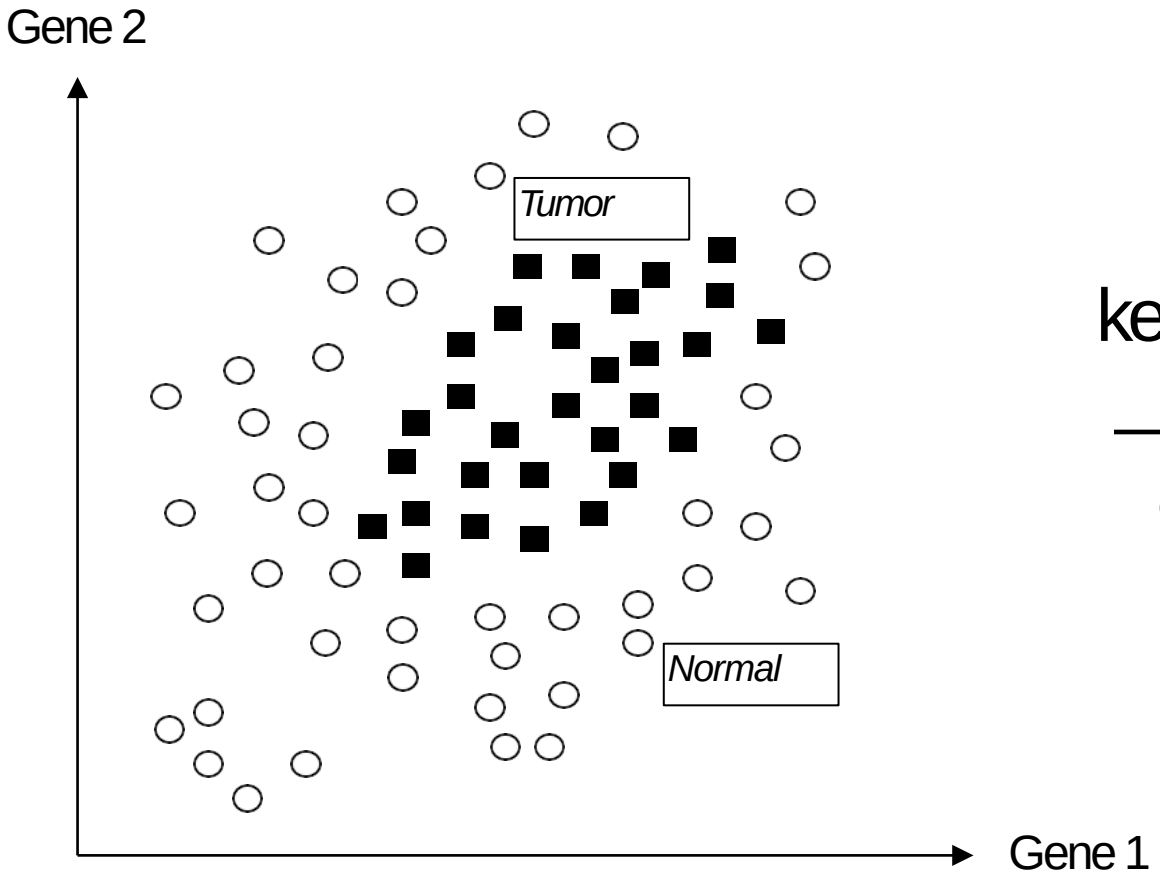
Minimize $\frac{1}{2}\sum_{i=1}^{n} w_i^2$ subject to $y_i(\vec{w}\cdot\vec{x}_i + b) - 1 \geq 0$ for $i = 1,\ldots,N$

Objective function

Constraints

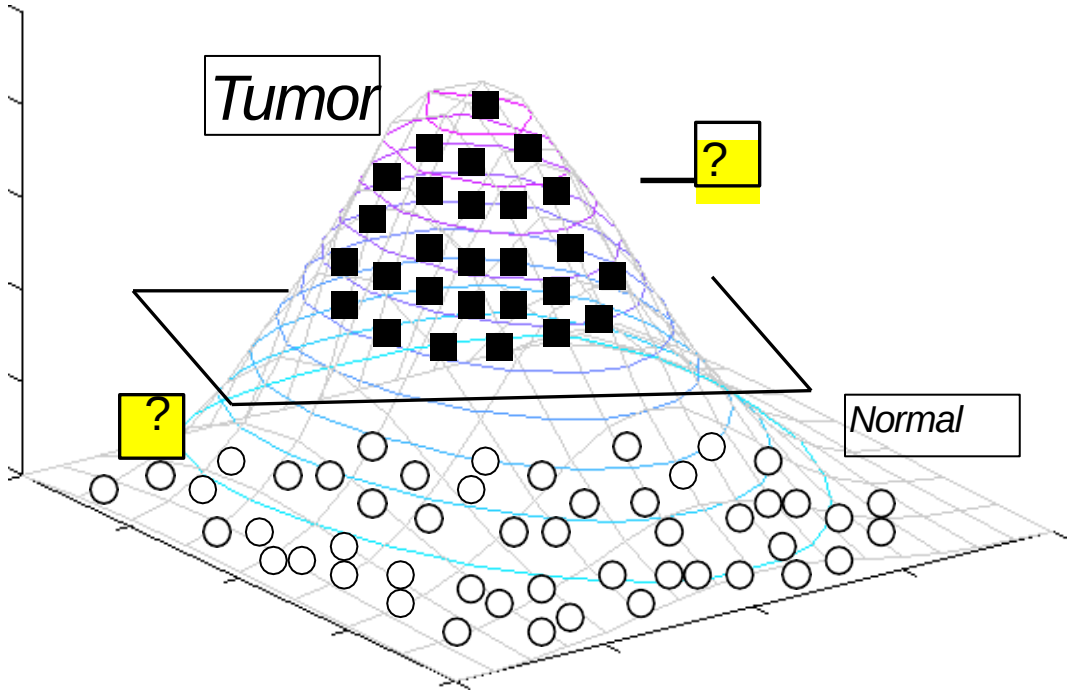
- This is called "*primal formulation of linear SVMs*".
- It is a convex quadratic programming (QP) optimization problem with n variables (w_i, $i = 1,\ldots,n$), where n is the number of features in the dataset.

CASE 2: NOT LINEARLY SEPARABLE DATA; KERNEL TRICK



kernel

Φ



Data is not linearly separable in the input space

Data is linearly separable in the feature space obtained by a kernel

$$\Phi : \mathbf{R}^N \rightarrow \mathbf{H}$$

POPULAR KERNELS

A kernel is a dot product in *some* feature space:

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i) \cdot \Phi(\vec{x}_j)$$

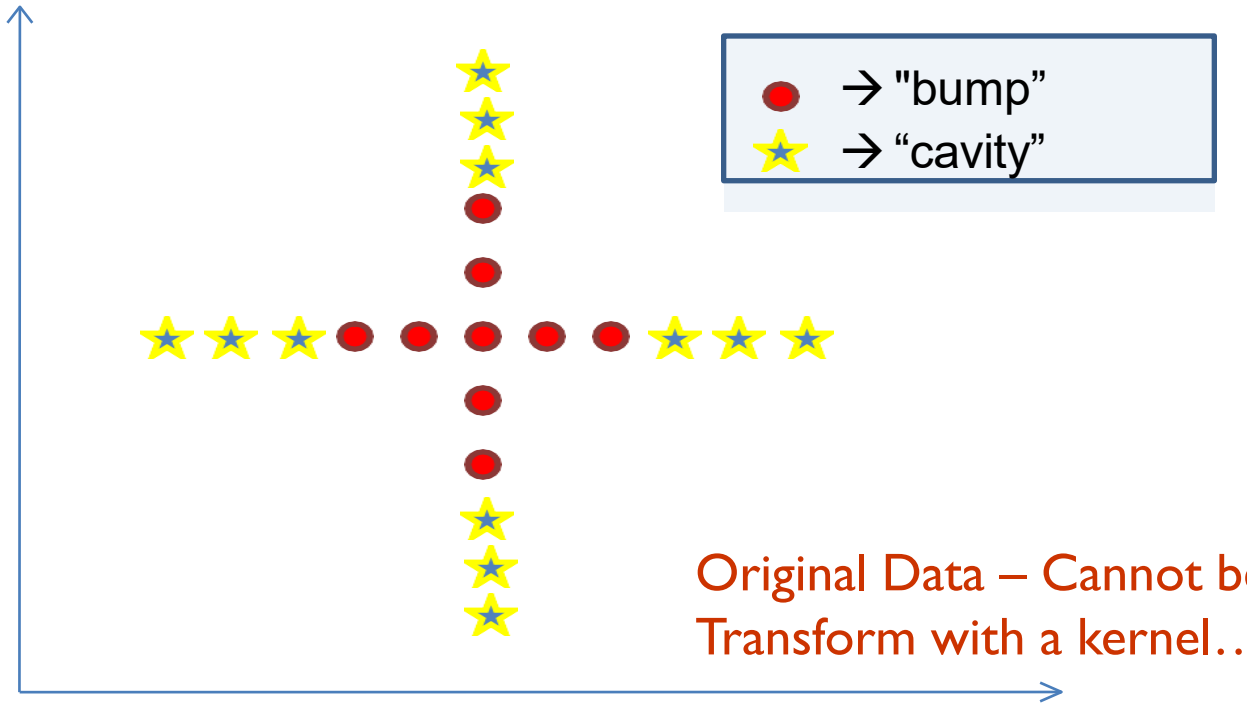
Examples:

$K(\vec{x}_i, \vec{x}_j) = \vec{x}_i \cdot \vec{x}_j$	Linear kernel
$K(\vec{x}_i, \vec{x}_j) = \exp(-\gamma \left\| \vec{x}_i - \vec{x}_j \right\|^2)$	Gaussian kernel
$K(\vec{x}_i, \vec{x}_j) = \exp(-\gamma \left\| \vec{x}_i - \vec{x}_j \right\|)$	Exponential kernel
$K(\vec{x}_i, \vec{x}_j) = (p + \vec{x}_i \cdot \vec{x}_j)^q$	Polynomial kernel
$K(\vec{x}_i, \vec{x}_j) = (p + \vec{x}_i \cdot \vec{x}_j)^q \exp(-\gamma \left\| \vec{x}_i - \vec{x}_j \right\|^2)$	Hybrid kernel
$K(\vec{x}_i, \vec{x}_j) = \tanh(k\vec{x}_i \cdot \vec{x}_j - \delta)$	Sigmoidal

UNDERSTANDING THE GAUSSIAN KERNEL

Consider Gaussian kernel: $K(\vec{x}, \vec{x}_j) = \exp(-\gamma \left\| \vec{x} - \vec{x}_j \right\|^2)$

Geometrically, this is a "bump" or "cavity" centered at the training data point $\vec{x}_j$:

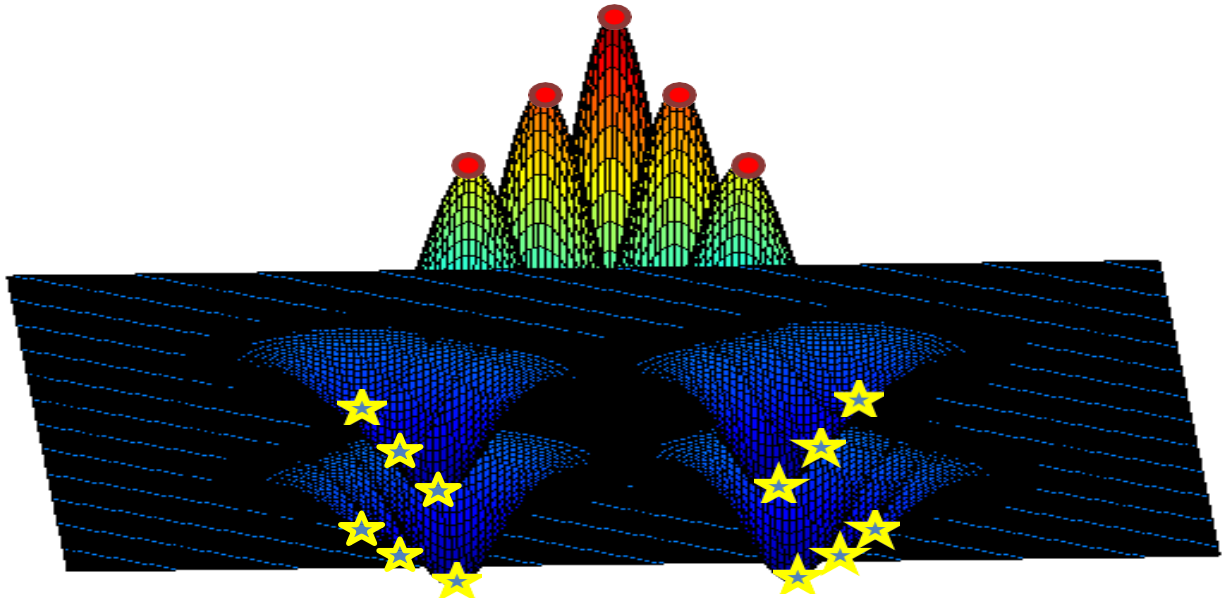


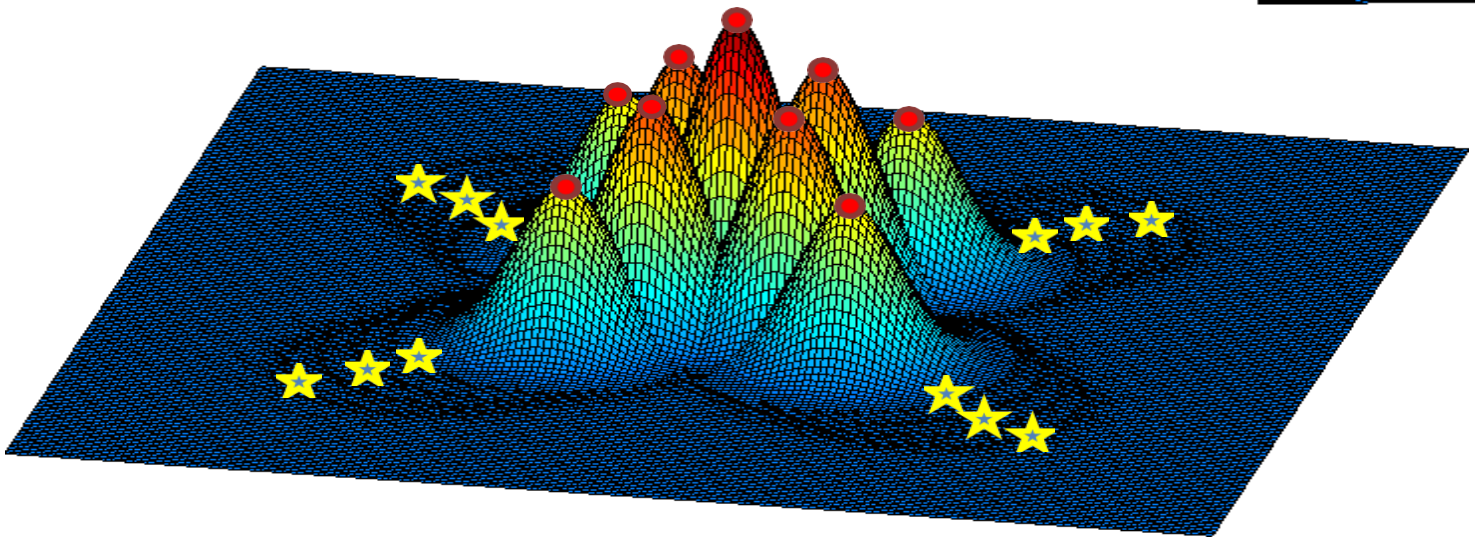
The resulting mapping function is a **combination** of bumps and cavities.

Original Data – Cannot be classified as it if.
Transform with a kernel…

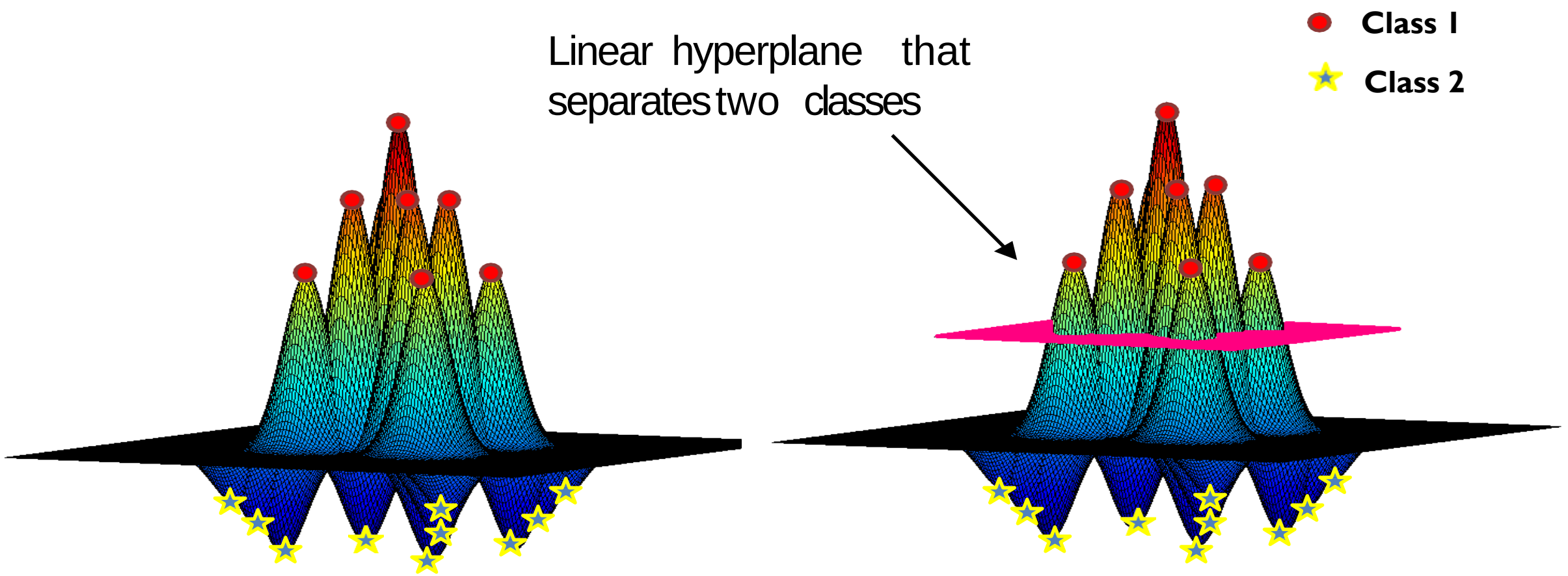
UNDERSTANDING THE GAUSSIAN KERNEL

Several more views of the data is mapped to the feature space by Gaussian kernel





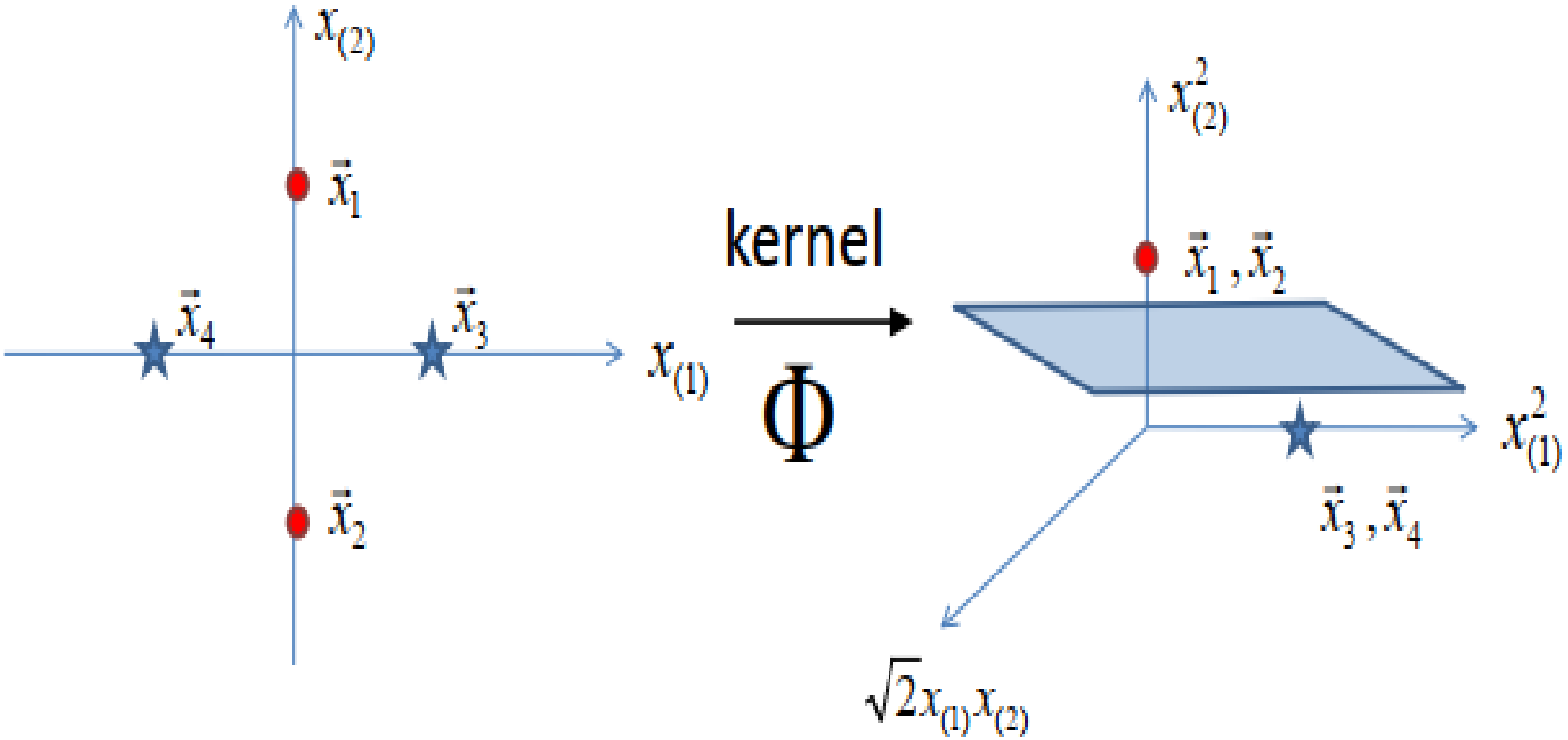
UNDERSTANDING THE GAUSSIAN KERNEL



Now, a hyper plane (pink color) will be able to classify – yellow color stars & Red color circles

EXAMPLE OF BENEFITS OF USING A KERNEL

Therefore, the explicit mapping is $\Phi(\vec{x}) = \begin{pmatrix} x_{(1)}^2 \\ \sqrt{2}x_{(1)}x_{(2)} \\ x_{(2)}^2 \end{pmatrix}$



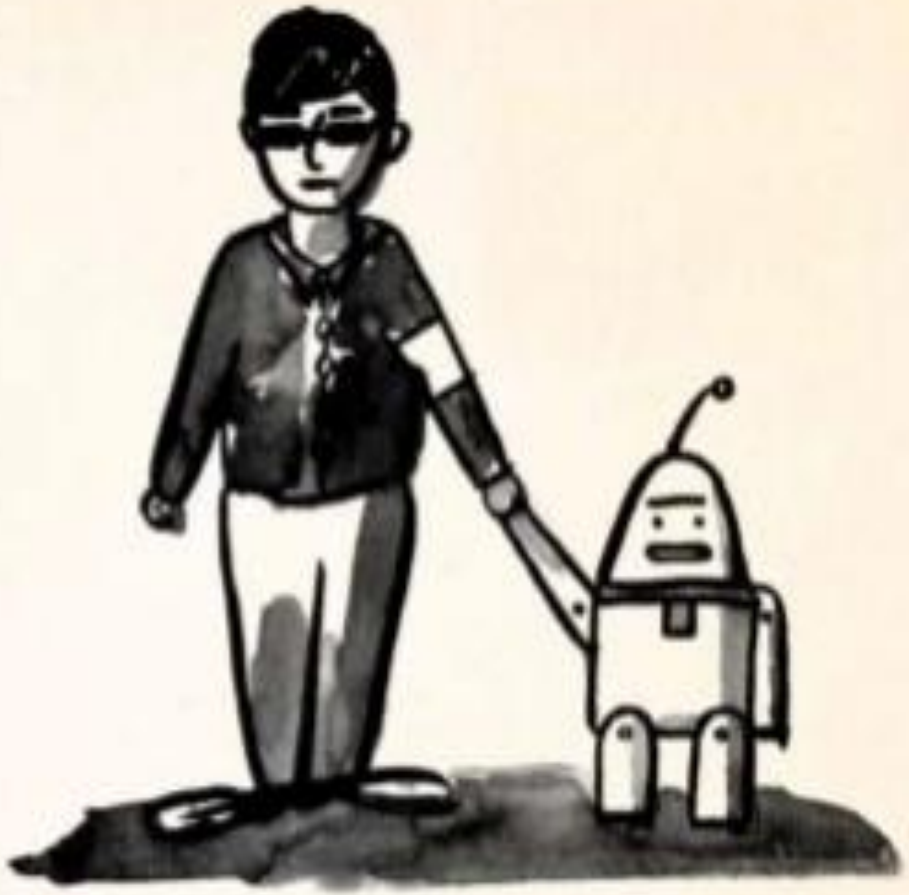
STRONG POINTS OF SVM-BASED LEARNING METHODS

- Empirically achieve excellent results in high-dimensional data with very few samples
- Internal capacity control to avoid overfitting
- Can learn both simple linear and very complex nonlinear functions by using "kernel trick"
- Robust to outliers and noise (use "slack variables")
- Convex QP optimization problem (thus, it has global minimum and can be solved efficiently)
- Solution is defined only by a small subset of training points ("support vectors")
- Number of free parameters is bounded by the number of support vectors and not by the number of variables
- Do not require direct access to data, work only with dot-products of data-points.

WEAK POINTS OF SVM-BASED LEARNING METHODS

- Measures of uncertainty of parameters are not currently well-developed
- Interpretation is less straightforward than classical statistics
- Lack of parametric statistical significance tests
- Power size analysis and research design considerations are less developed than for classical statistics

CASE STUDIES



Thank You!

QUESTIONS?