ISE426 Fall 2015

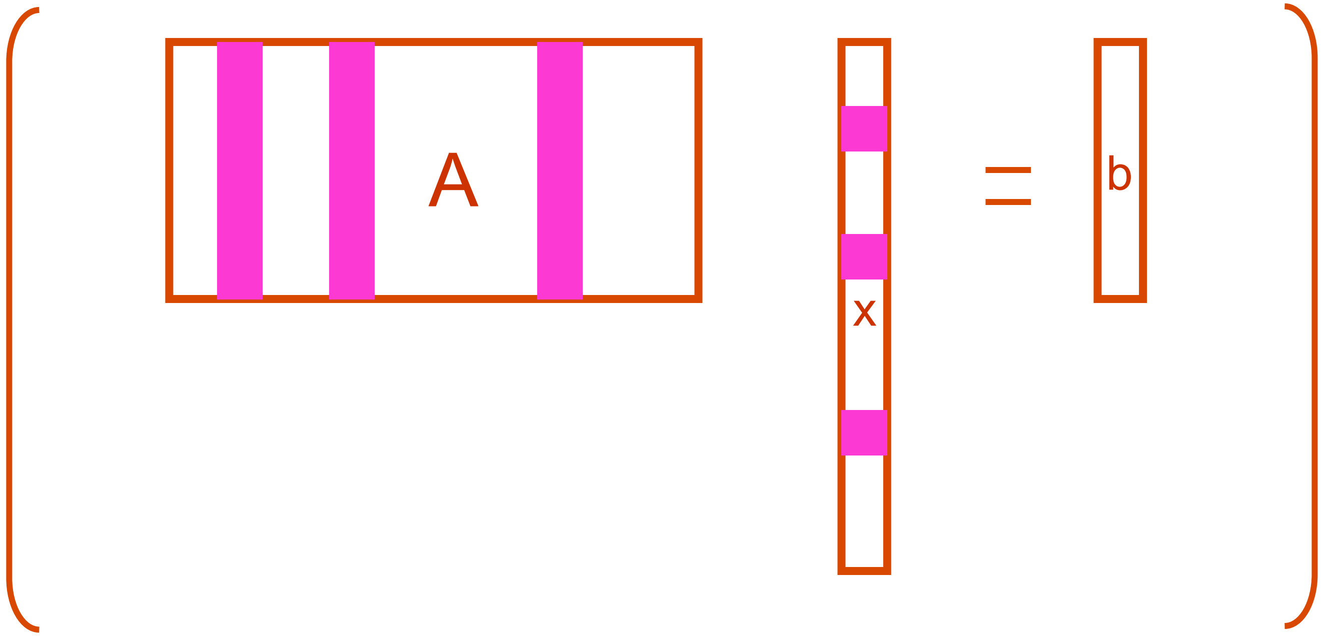
Lecture 23 – November 24, 2015

Convex quadratic programming in Machine Learning:

Sparse Optimization

Support Vector Machines and

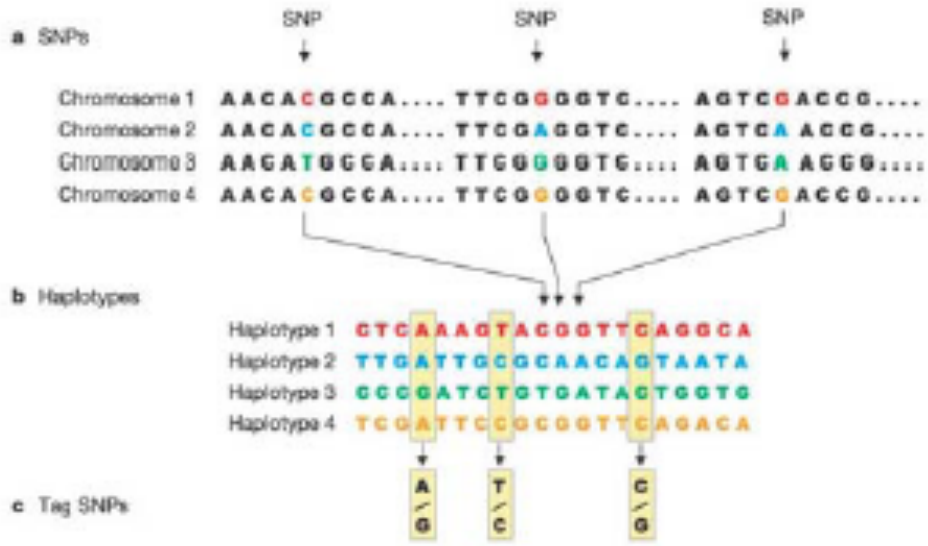
Lasso (sparse LS regression)



$$Ax \approx b$$

x has few nonzero elements: $\|x\|_0$ is small!

Example from gene expression



- Single Nucleotide Polymorphism (SNP) – point sites of variation in traits
- Each SNP associated with two alleles (states)

Classifying state of a disease based on some of the SNPs

Not known which SNPs are important – use feature selection

600,000 SNPs and 5,000 individuals/data points.

Sparse solutions

Sparse signal reconstruction

$$\begin{aligned} \min \quad & ||x||_0 \\ s.t. \quad & Ax = b. \end{aligned}$$

Sparse solution $x \in \mathbf{R}^n$, matrix $A \in \mathbf{R}^{m \times n}, n >> m$

The system is underdetermined, but if card(x)<m, can recover signal.

How do we formulate this as an MILP?

$$\begin{aligned} \min \quad & \sum y_i \\ s.t. \quad & Ax = b. \\ & x_i \leq My_i, \;\; i = 1, \ldots, n \\ & -x_i \leq My_i \;\; i = 1, \ldots, n \\ & y_i \in \{0, 1\} \;\; i = 1, \ldots, n \end{aligned}$$

Sparse solution using l_1-norm

The problem is difficult in general. Typical relaxation,

$$\begin{aligned} \min \quad & \sum y_i \\ s.t. \quad & Ax = b. \\ & x_i \le M y_i,\ i = 1, \dots, n \\ & -x_i \le M y_i\ i = 1, \dots, n \\ & 0 \le y_i \le 1\ i = 1, \dots, n \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \min \quad & \sum \frac{|x_i|}{M} \\ s.t. \quad & Ax = b. \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \min \quad & ||x||_1 \\ s.t. \quad & Ax = b. \end{aligned}$$

Sparse solutions using the l_1-norm

Sparse signal reconstruction

$$\begin{aligned} \min \quad & \|Ax - b\| \\ s.t. \quad & \|x\|_0 \leq k \end{aligned}$$

k-sparse signal $x \in \mathbf{R}^n$, matrix $A \in \mathbf{R}^{m \times n}, n >> m$

The system is underdetermined, but if card(x)<k, can recover signal.

How do we formulate this as an MILP?

$$\begin{aligned} \min \quad & \|Ax - b\|^2 \\ s.t. \quad & \sum y_i \leq k \\ & x_i \leq M y_i, \;\; i = 1, \ldots, n \\ & -x_i \leq M y_i \;\; i = 1, \ldots, n \\ & y_i \in \{0, 1\} \;\; i = 1, \ldots, n \end{aligned}$$

Recovery by using the l_1-norm

The problem is difficult in general. Typical relaxation,

$$\begin{aligned} \min \quad & \|Ax - b\| \\ s.t. \quad & \sum y_i \le k \\ & x_i \le M y_i,\ \ i = 1, \ldots, n \\ & -x_i \le M y_i \ i = 1, \ldots, n \\ & 0 \le y_i \le 1 \ i = 1, \ldots, n \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \min \quad & \|Ax - b\|^2 \\ s.t. \quad & \sum \frac{|x_i|}{M} \le k \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} \min \quad & \|Ax - b\|^2 \\ s.t. \quad & ||x||_1 \le t(= kM?) \end{aligned}$$

Other formulations

Regularized regression or Lasso:

$$\min \quad \frac{1}{2}||Ax-b||^2 + \lambda||x||_1$$

Sparse regressor selection

Noisy signal recovery

$$\begin{aligned} \min \quad & ||Ax-b|| \\ s.t. \quad & ||x||_1 \le t. \end{aligned}$$

$$\begin{aligned} \min \quad & ||x||_1 \\ s.t. \quad & ||Ax-b|| \le \epsilon. \end{aligned}$$

Types of convex problems

$$\min \quad \frac{1}{2}||Ax - b||^2 + \lambda||x||_1$$

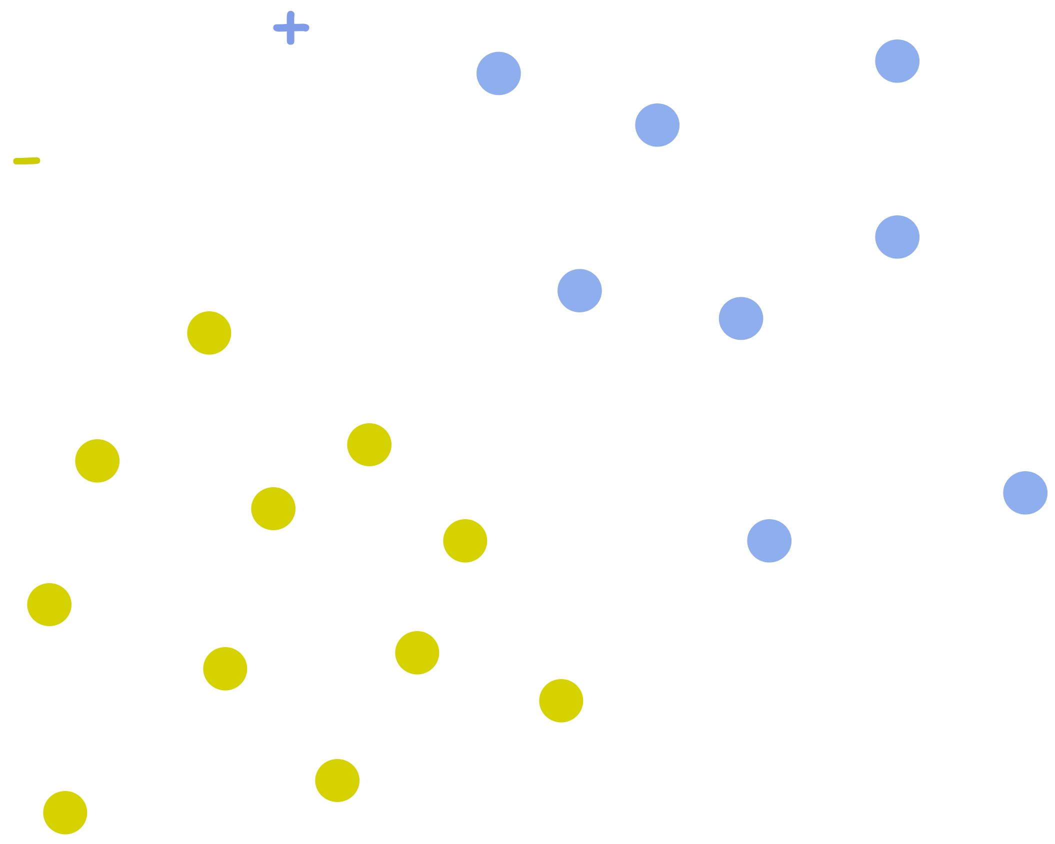
Variable substitution: $x = x' - x'',\ x' \geq 0,\ x'' \geq 0$

$$\begin{aligned} \min \quad & \frac{1}{2}||A(x' - x'') - b||^2 + \lambda(x' + x'') \\ \text{s.t.} \quad & x' \geq 0, x'' \geq 0 \end{aligned}$$

Convex non-smooth objective with linear inequality constraints

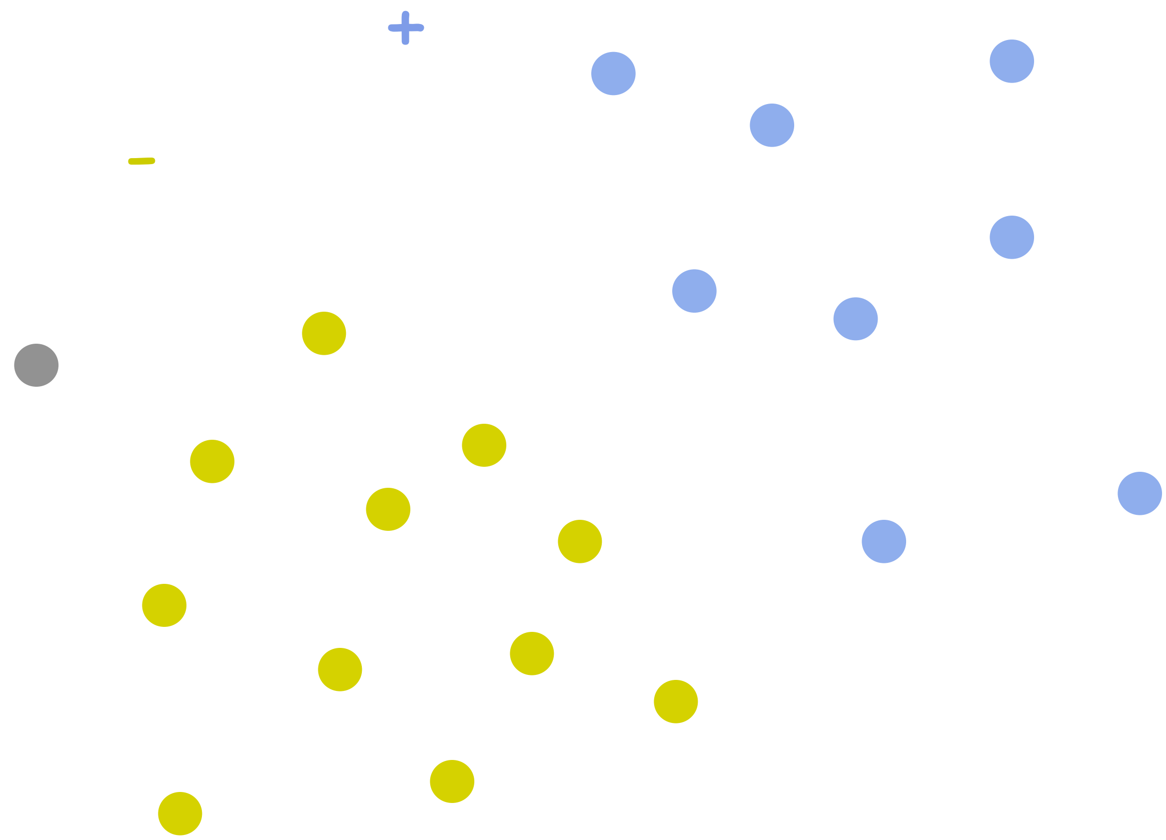
Binary classification problem

Two sets of labeled points



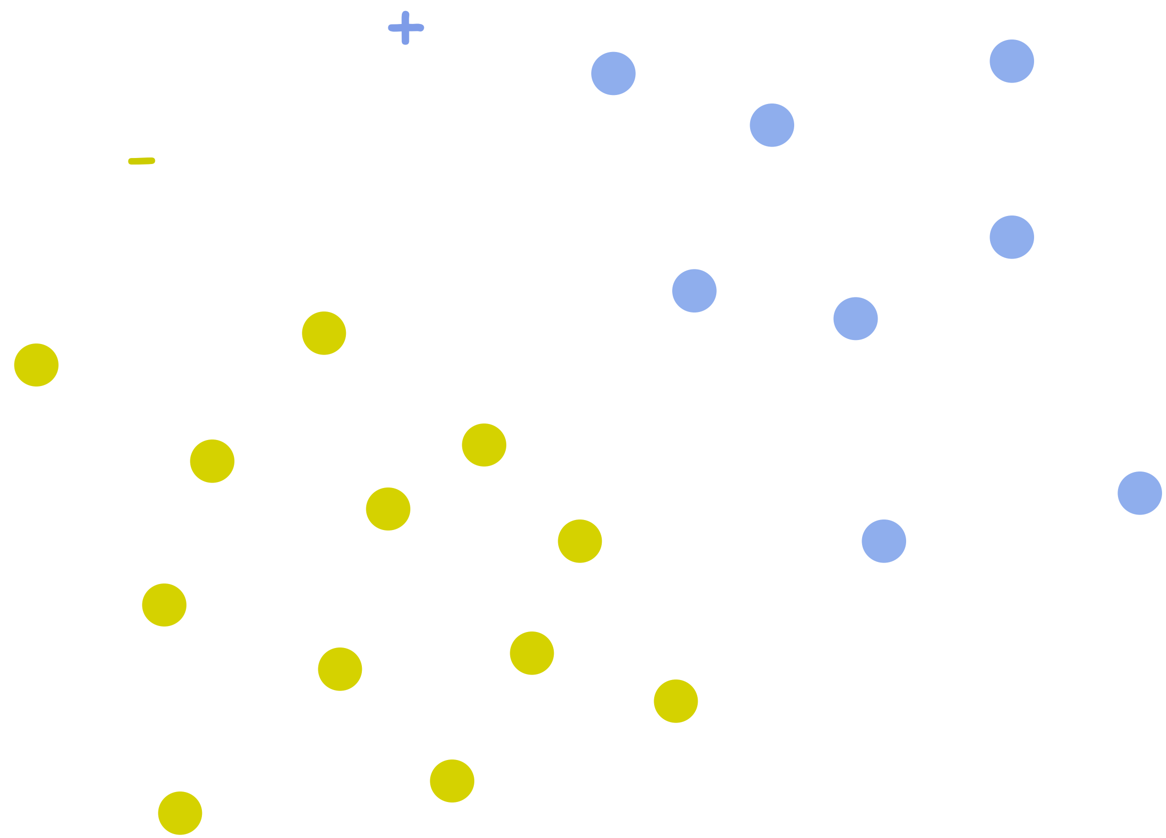
Binary classification problem

How to label
this new point?

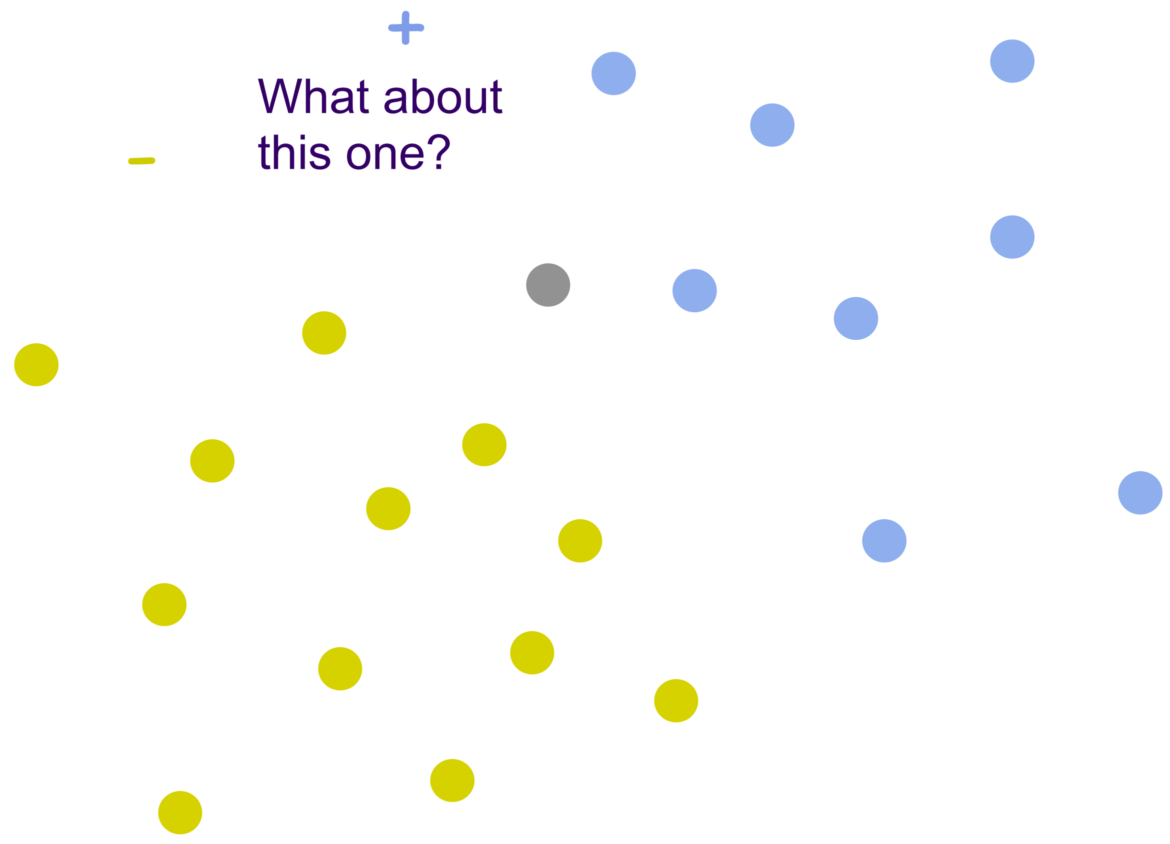


Binary classification problem

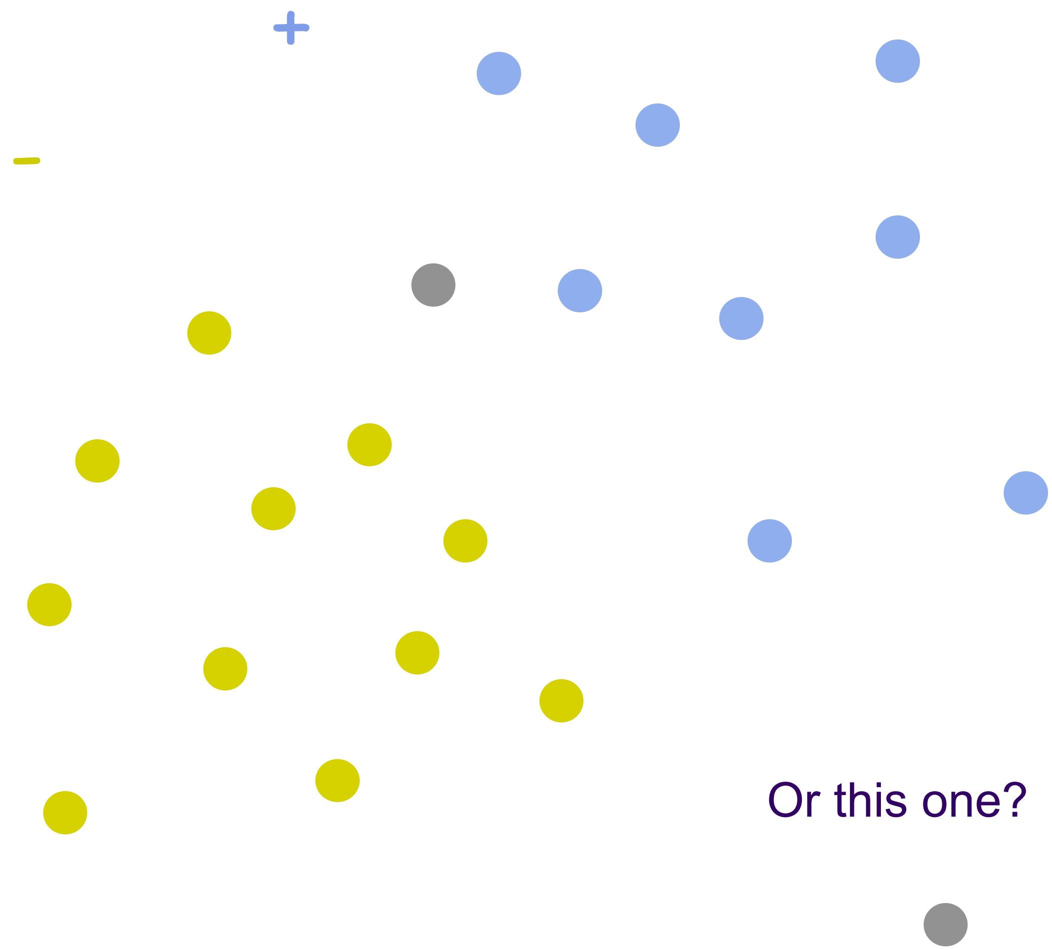
Probably green



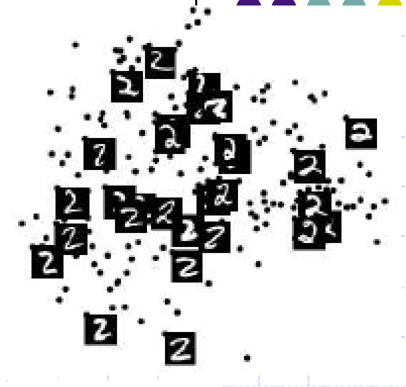
Binary classification problem



Binary classification problem



Examples from image classification



- Optical character recognition
  - Automatically read digits in zip code
    - 256 dim vector of pixels, 10 classes,
    - classification or clustering task
- Face recognition and detection
  - much larger dimension, nonlinear representation,
  - Non-euclidean similarity measures

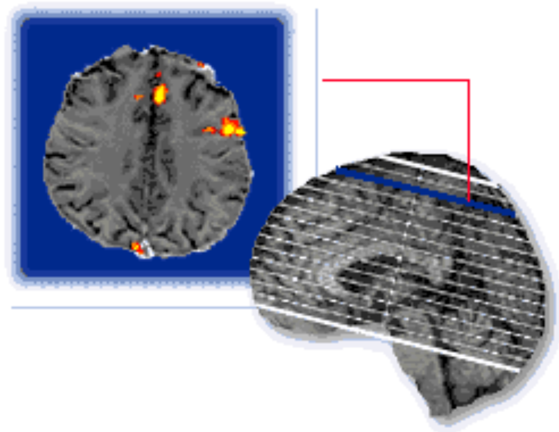
Examples from text and internet

- Text categorization
  - detect spam/nonspam emails
    - Many possible features
    - False positives are very bad, false negatives are OK.
    - Online setting possible, huge data sets.
  - choose articles of interest to individualize news sites
    - Large dimension – size of dictionary, small training set, possibly online setting
    - Only few words are important.
- Ranking
  - Predict a page rank for a given a search query
    - How to do it? Predict relative ranks of each pair of pages?

Examples from Medicine

- Functional Magnetic resonance imaging
  - Uses a standard MRI scanner to acquire functionally meaningful brain activity
  - Measures changes in blood oxygenation
  - Non-invasive, no ionizing radiation
  - Good combination of spatial / temporal resolution
    - Voxel sizes ~4mm
    - Time of Repetition (TR) ~1s

  About 30000 voxels are active and measured.
  - Only a few (probably) contribute to what the subject is “feeling” during the experiment (anger, frustration, boredom..)



- Breast cancer risk patients
  - Take several measurements of a patient and some basic characteristics an predict if the patient is at high risk
  - Low dimensional, but very different attributes. Large scale data.
  - May involve “active learning” – additional labels obtained by involving more tests or a professional.
  - KDD 2008 cup challenge

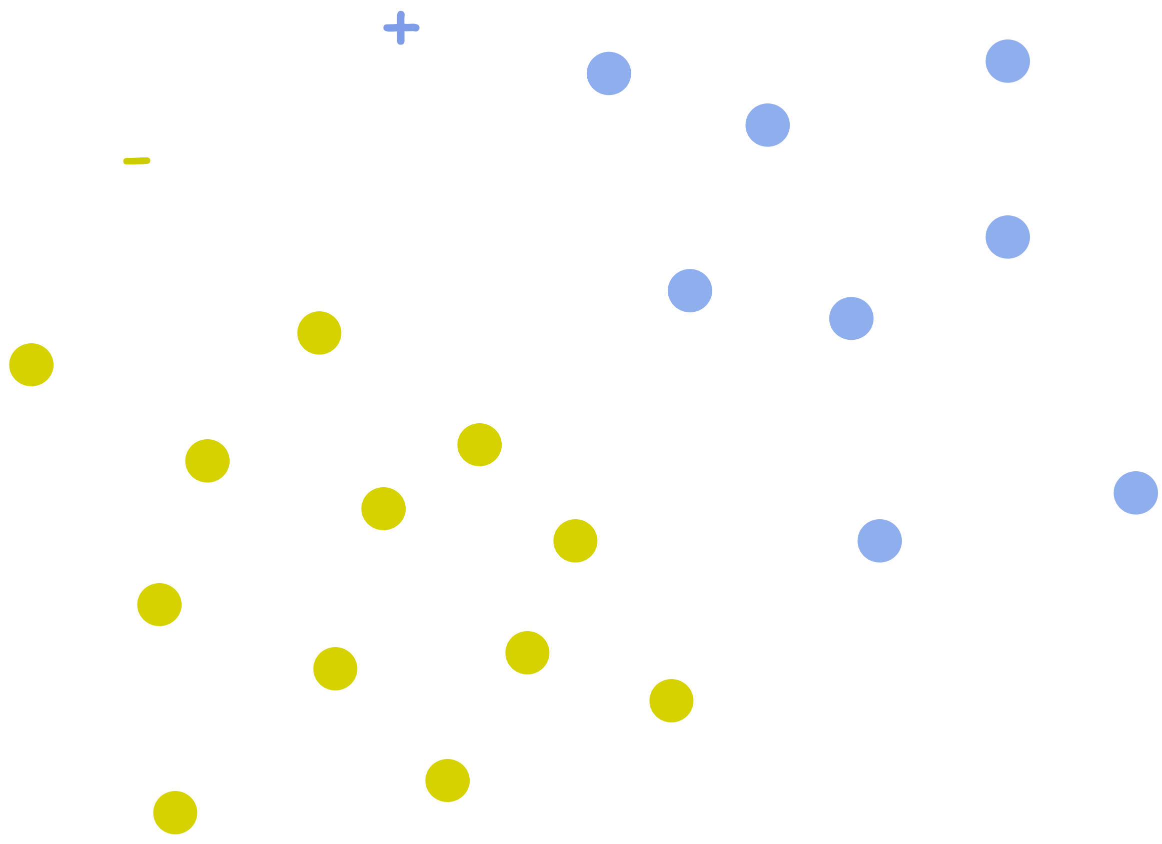
fMRI image courtesy of fMRI Research Center @ Columbia Unoversity

The binary classification problem

- The universe of data-label pairs (x, y),
- $y \in \{+1, -1\}$ for all $x \in \mathbf{R}^m$.
- Given a set $X \subset \mathbf{R}^m$ of n vectors.
- For each $x_i \in X$ the label y_i is known.
- Find a function $f(x) \approx y$

Linear classifier

Idea: separate a space into two half-spaces

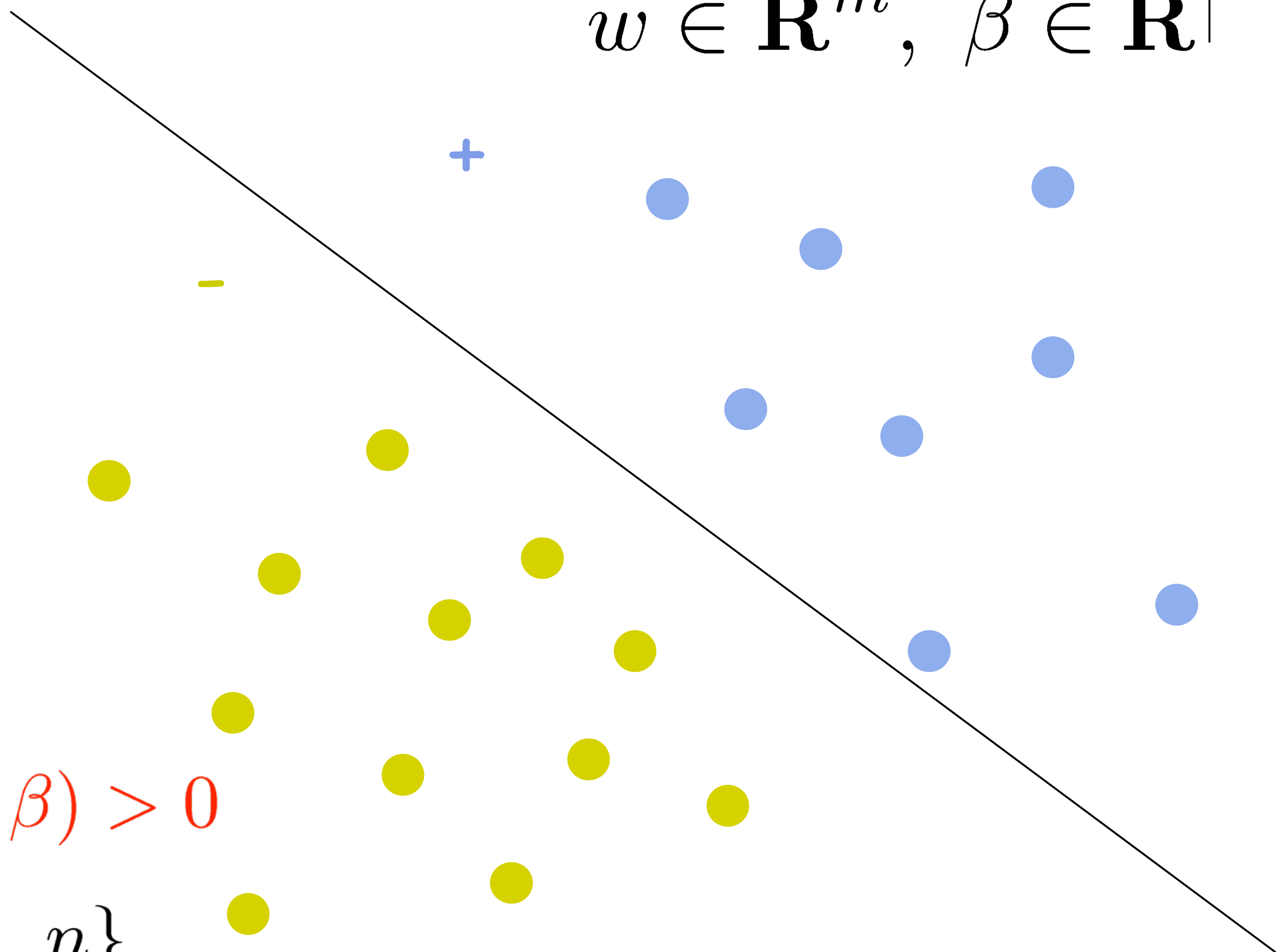


Linear classifier

$$w^\top x + \beta = 0$$

$$w \in \mathbf{R}^m, \;\; \beta \in \mathbf{R}$$

Like this:



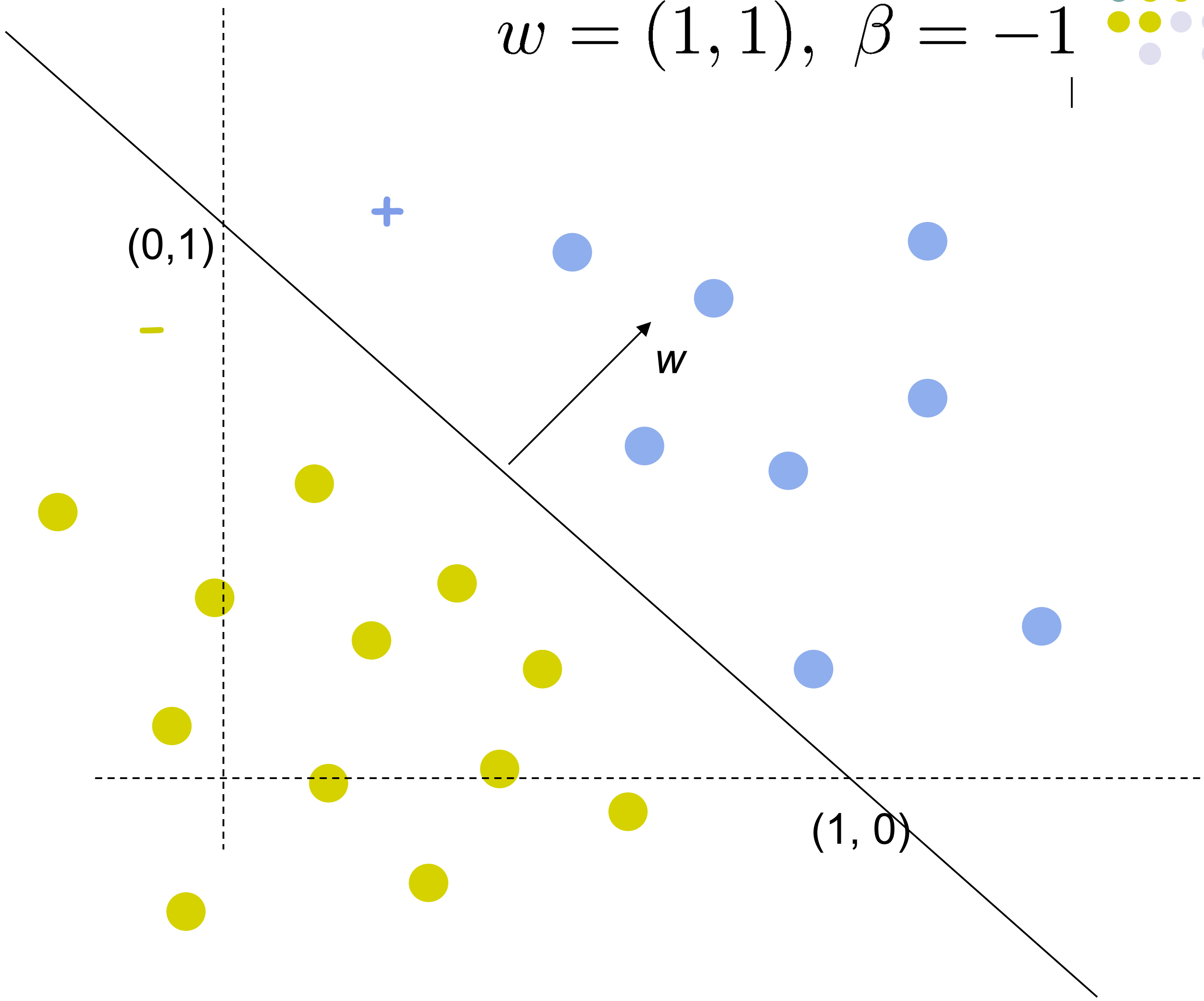
$$y_i(w^\top x_i + \beta) > 0$$

$$\forall i \in \{1..n\}$$

Linear classifier

$$x_1 + x_2 - 1 = 0$$

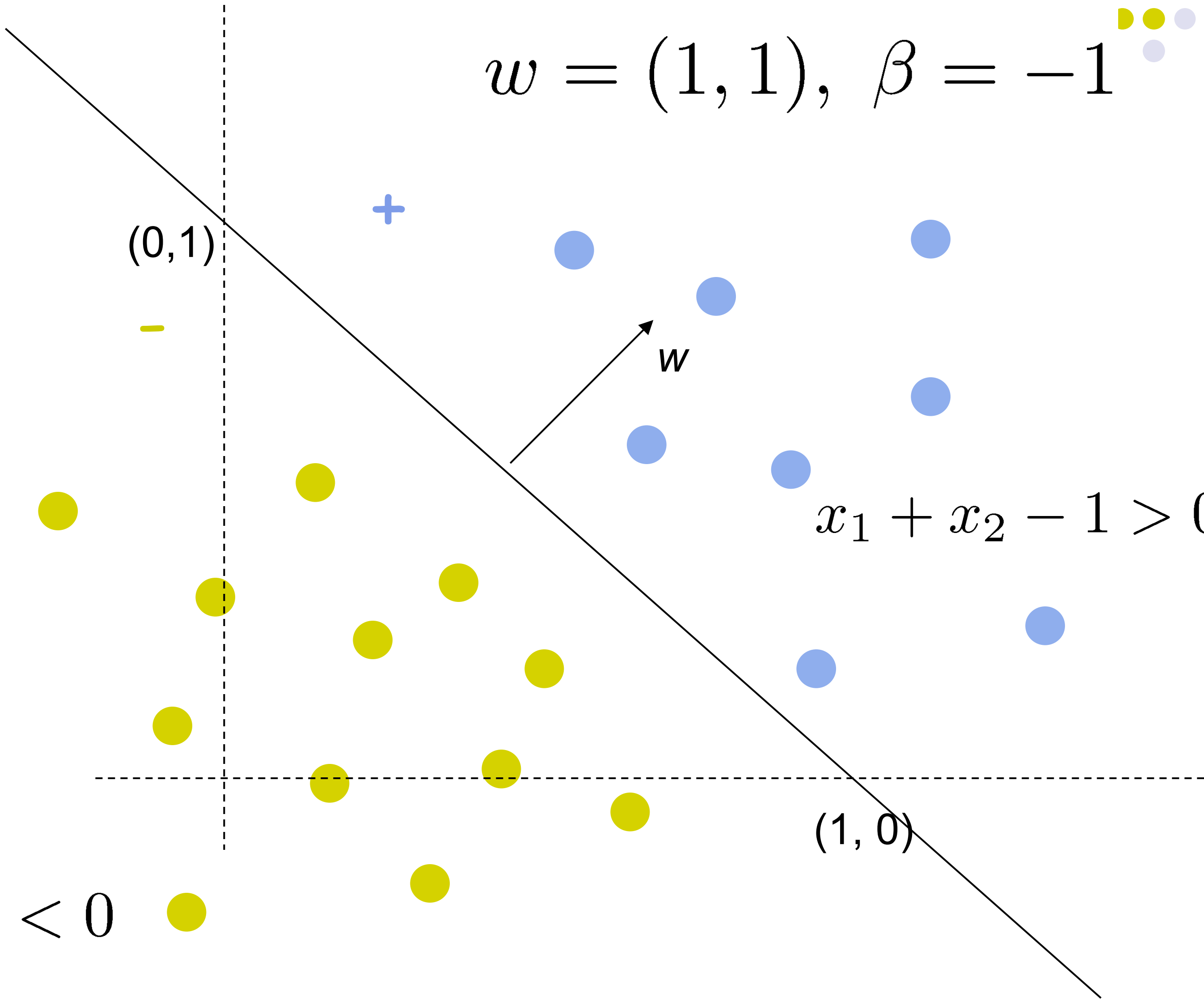
$$w = (1, 1),\ \beta = -1$$



Linear classifier

$$x_1 + x_2 - 1 = 0$$

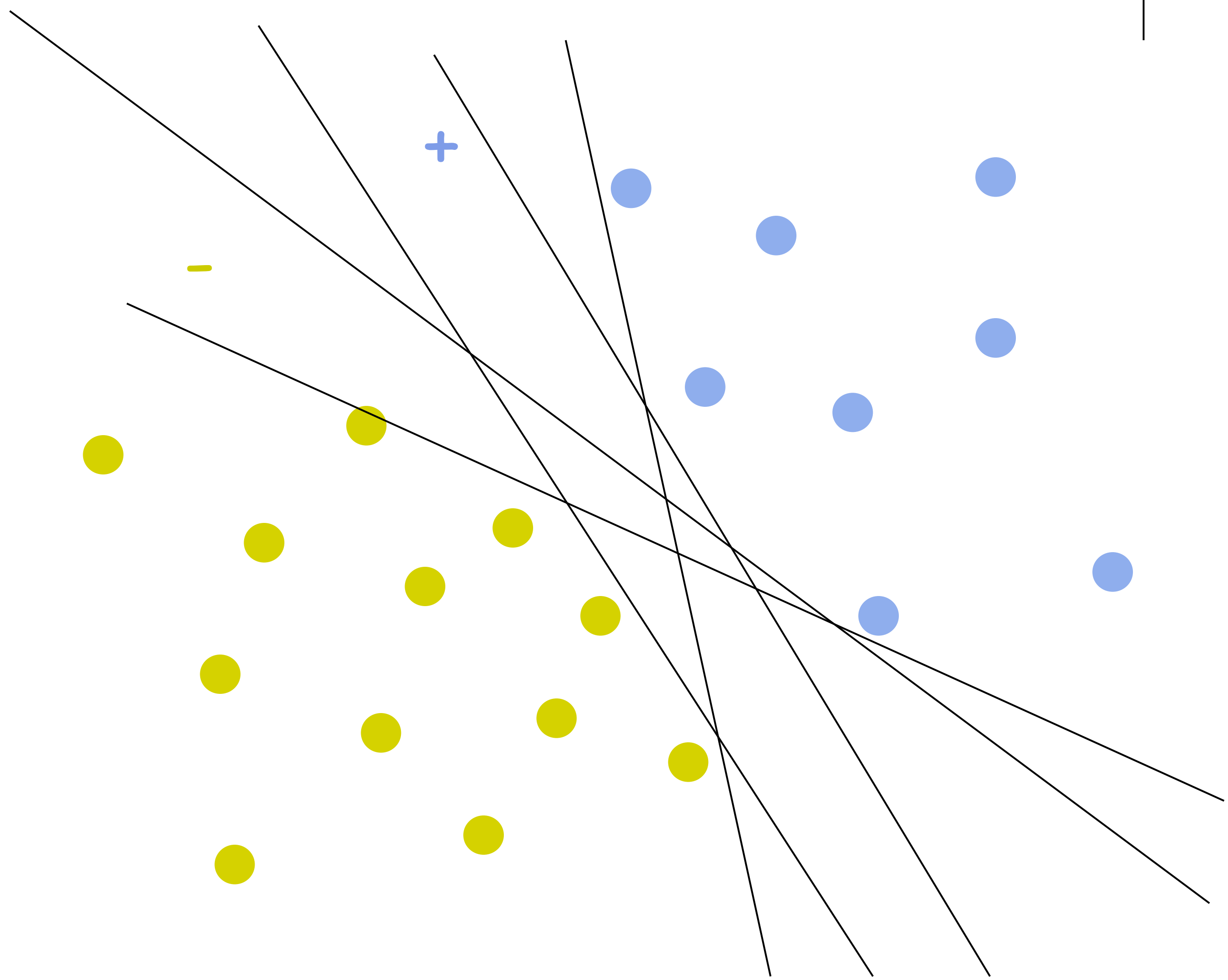
$$w = (1, 1), \ \beta = -1$$



$$x_1 + x_2 - 1 > 0$$

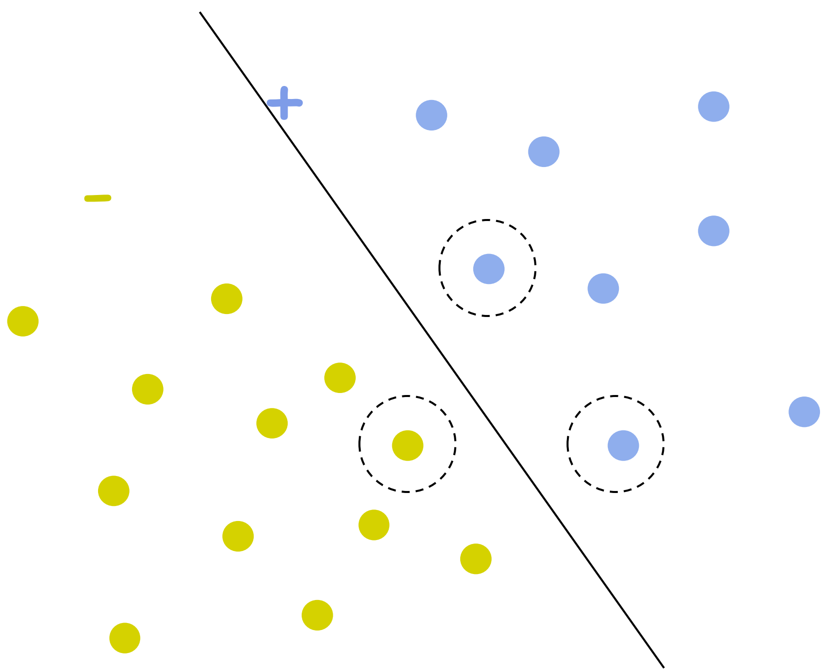
$$x_1 + x_2 - 1 < 0$$

Linear classifier



Support vector machines

Assume each x_i is not known exactly,
but $z_i \in B(x_i, r)$



$$\min_{z_i \in B_i} y_i(w^\top z_i + \beta) \geq 0, \ \forall i \in \{1..n\}$$

$$\Downarrow$$

$$y_i(w^\top x_i + \beta) - \frac{r}{\|w\|} w^\top w \geq 0, \ \forall i \in \{1..n\}$$

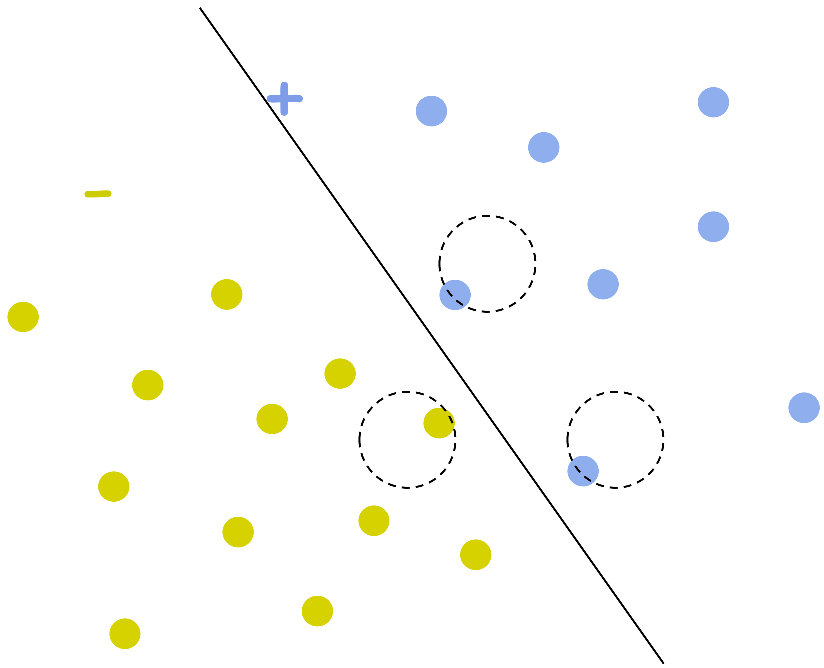
$$\Downarrow$$

$$y_i(w^\top x_i + \beta) - ||w||r \geq 0, \ \forall i \in \{1..n\}$$

Find the largest r or the smallest $||w||$

Support vector machines

Assume each x_i is not known exactly, but $z_i \in B(x_i, r)$



$$\min_{z_i \in B_i} y_i(w^\top z_i + \beta) \geq 0, \ \forall i \in \{1..n\}$$

$$\Downarrow$$

$$y_i(w^\top x_i + \beta) - \frac{r}{\|w\|} w^\top w \geq 0, \ \forall i \in \{1..n\}$$

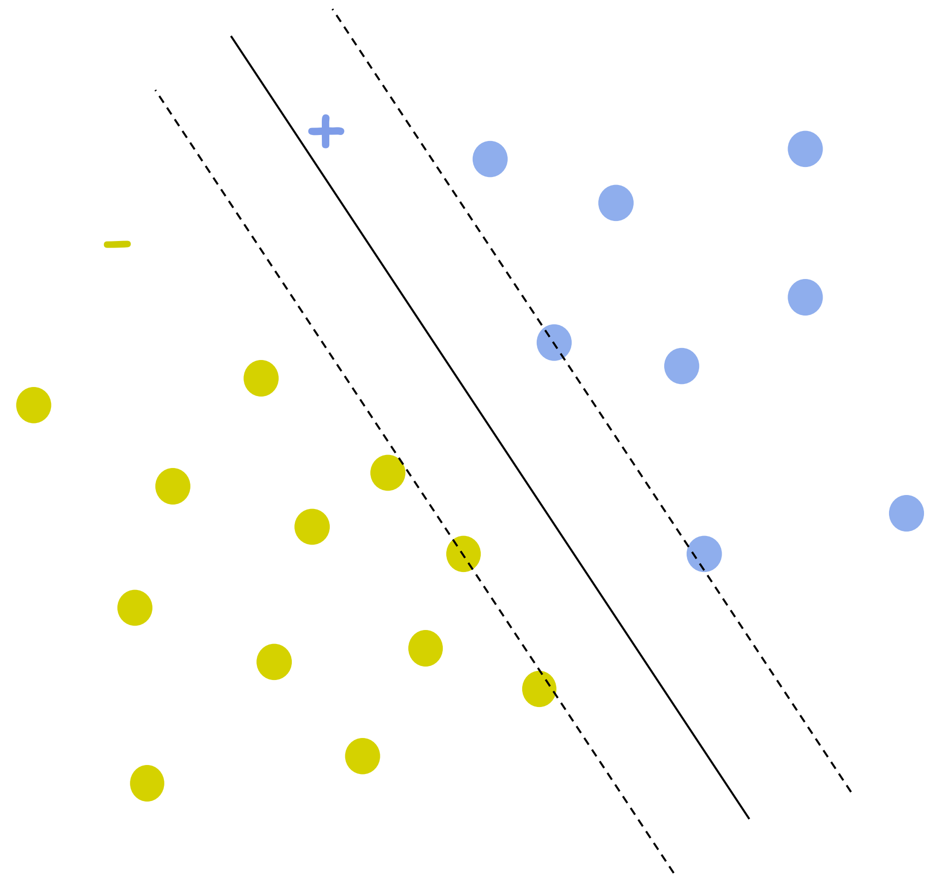
$$\Downarrow$$

$$y_i(w^\top x_i + \beta) - \|w\| r \geq 0, \ \forall i \in \{1..n\}$$

Find the largest r or the smallest $||w||$

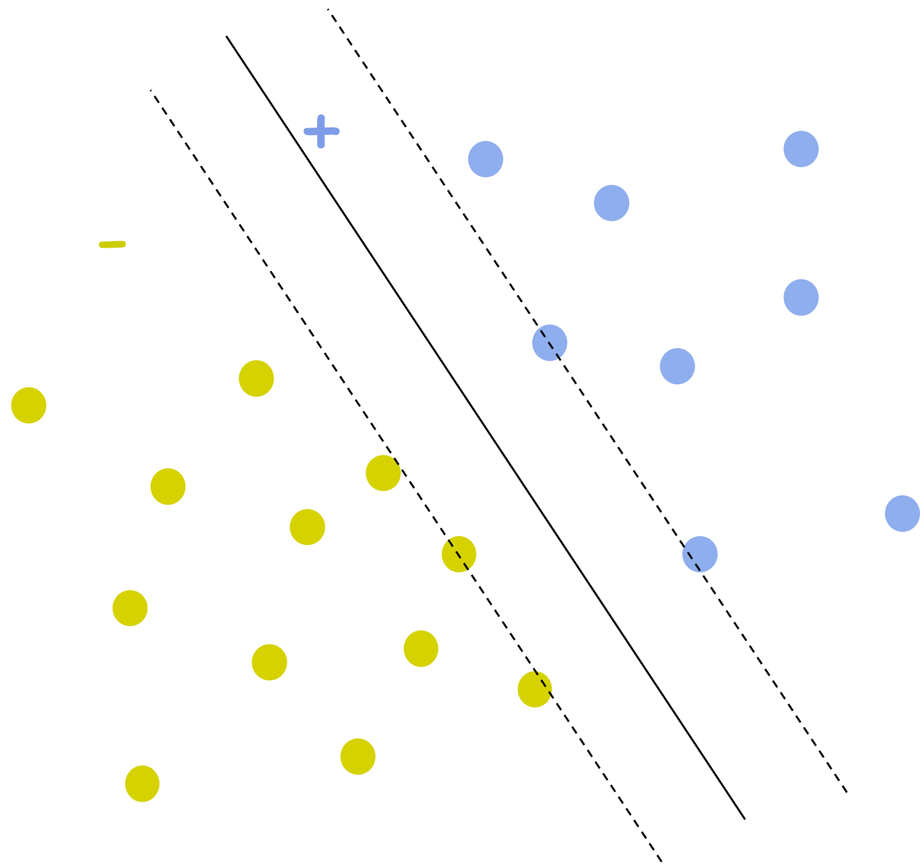
Support vector machines

$$\min_{w,\beta} ||w||, \text{ s.t. } y_i(w^\top x_i + \beta) - 1 \geq 0, \ \forall i \in \{1..n\}$$



Support vector machines

$$\min_{w,\beta} \frac{1}{2}||w||^2, \text{ s.t. } y_i(w^\top x_i + \beta) - 1 \geq 0, \ \forall i \in \{1..n\}$$



Optimization Problem

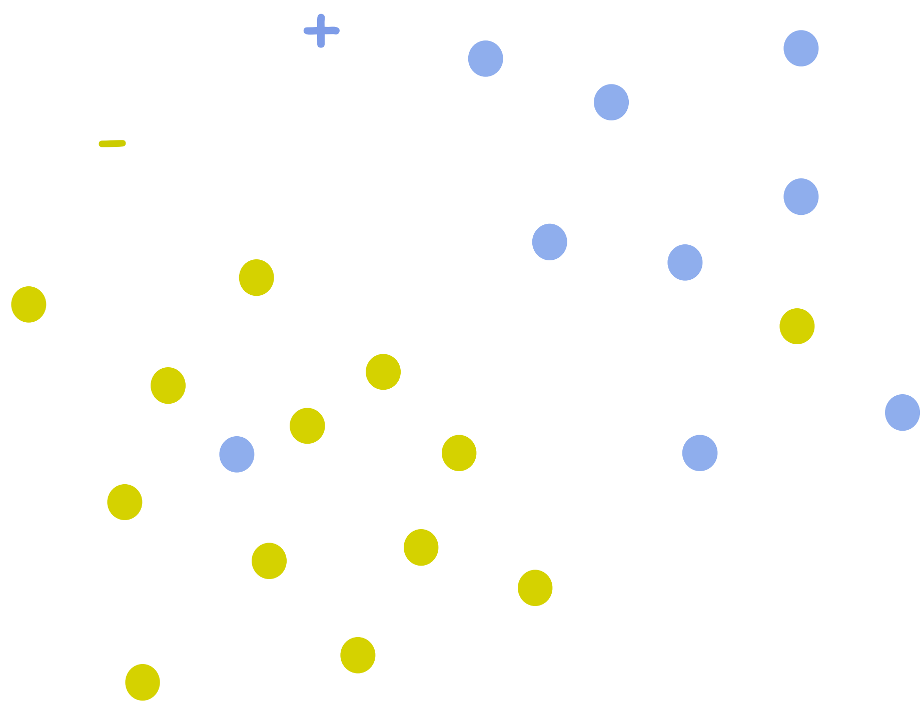
Total number of data points: n

$$\begin{aligned} \min_{w \in \mathbf{R}^m, \beta \in \mathbf{R}} \quad & \frac{1}{2} w^\top w \\ \text{s.t.} \quad & y_i(w^\top x_i + \beta) \geq 1, \quad i = 1, \ldots, n \end{aligned}$$

How many variables? Constraints? What can go wrong?

Support vector machines

$$y_i(w^\top x_i - b) - 1 \geq 0,\ \forall i \in \{1..n\}\ \ -\ \ \text{no such } w!$$



Soft margin SVM

Total number of data points: n

$$\begin{array}{rl} \min_{\xi,w,\beta} & \frac{1}{2}w^\top w \\ \text{s.t.} & y_i(w^\top x_i + \beta) \geq 1 - \xi_i, \quad i = 1, \ldots, n \end{array}$$

What's wrong with this formulation?

Soft margin SVM

Total number of data points: n

$$\begin{aligned} \min_{\xi,w,\beta} \quad & \frac{1}{2}w^\top w + c\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(w^\top x_i + \beta) \geq 1 - \xi_i, \quad i = 1,\ldots,n \\ & \xi \geq 0, \qquad i = 1,\ldots,n. \end{aligned}$$

How many variables? Constraints?

Soft margin SVM

Total number of data points: n

$$\begin{aligned} \min_{\xi,w,\beta} \quad & \frac{1}{2}w^\top w + c\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(w^\top x_i + \beta) \geq 1 - \xi_i, \quad i = 1,\ldots,n \\ & \xi \geq 0, \qquad i = 1,\ldots,n. \end{aligned}$$

How many variables? Constraints?

What if n is very large? What if m is very large?

Optimization Problem

$$
\begin{aligned}
\min_{\xi,w,\beta} \quad & \frac{1}{2} w^\top w + c \sum_{i=1}^{n} \xi_i \\
\text{s.t.} \quad & y_i(w^\top x_i + \beta) \geq 1 - \xi_i, \quad i = 1, \ldots, n \\
& \xi \geq 0, \qquad i = 1, \ldots, n.
\end{aligned}
$$

Every optimization problem has:

1. optimality conditions and 2. dual problem

Optimization Problem

$$\text{At optimality } w^* = \textstyle\sum_{i=1}^{n} \alpha_i y_i x_i, \quad 0 \leq \alpha_i \leq c$$

$$||w^*||^2 = (\sum_{i=1}^{n} \alpha_i y_i x_i)^\top (\sum_{i=1}^{n} \alpha_i y_i x_i) = \sum_{i,j=1}^{n} y_i y_j {x_i}^\top x_j \alpha_i \alpha_j$$

$$\begin{aligned} \min_{\alpha,\beta,\xi} \quad & \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j {x_i}^\top x_j \alpha_i \alpha_j + c \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & \sum_{j} y_i y_j {x_i}^\top x_j \alpha_j + y\beta + \xi_i \geq 1, \quad i = 1, \ldots, n \\ & \xi_i \geq 0, \ \ 0 \leq \alpha_i \leq c, \qquad i = 1, \ldots, n, \end{aligned}$$

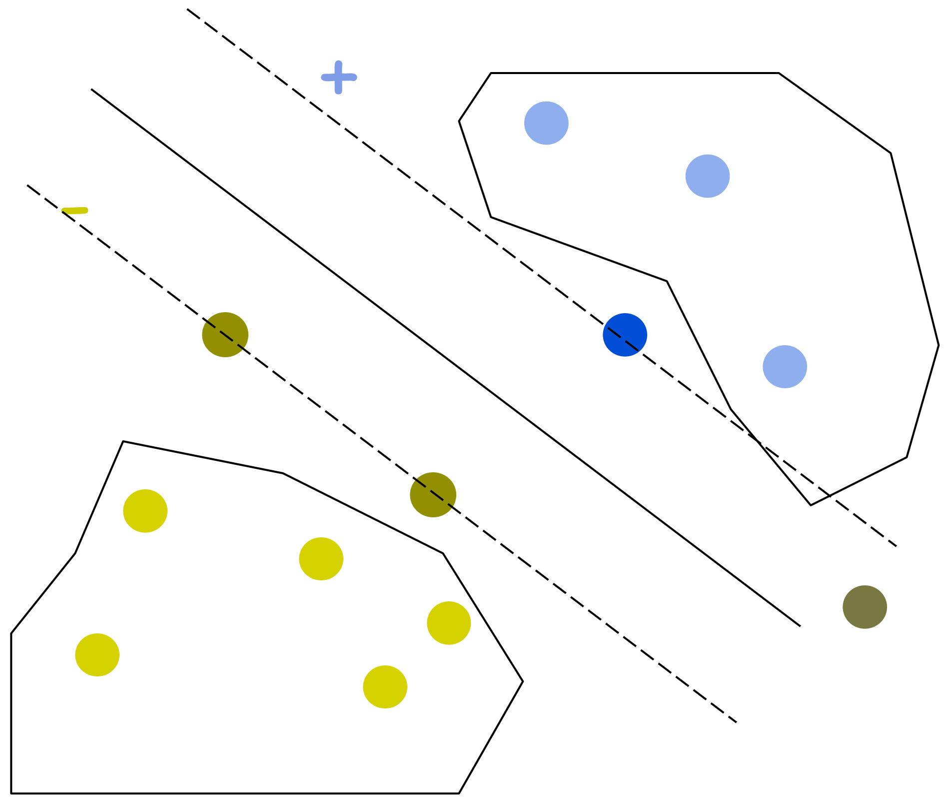
How many variables? Constraints?

Support Vectors

$0<\alpha<c$,
$\xi=0$

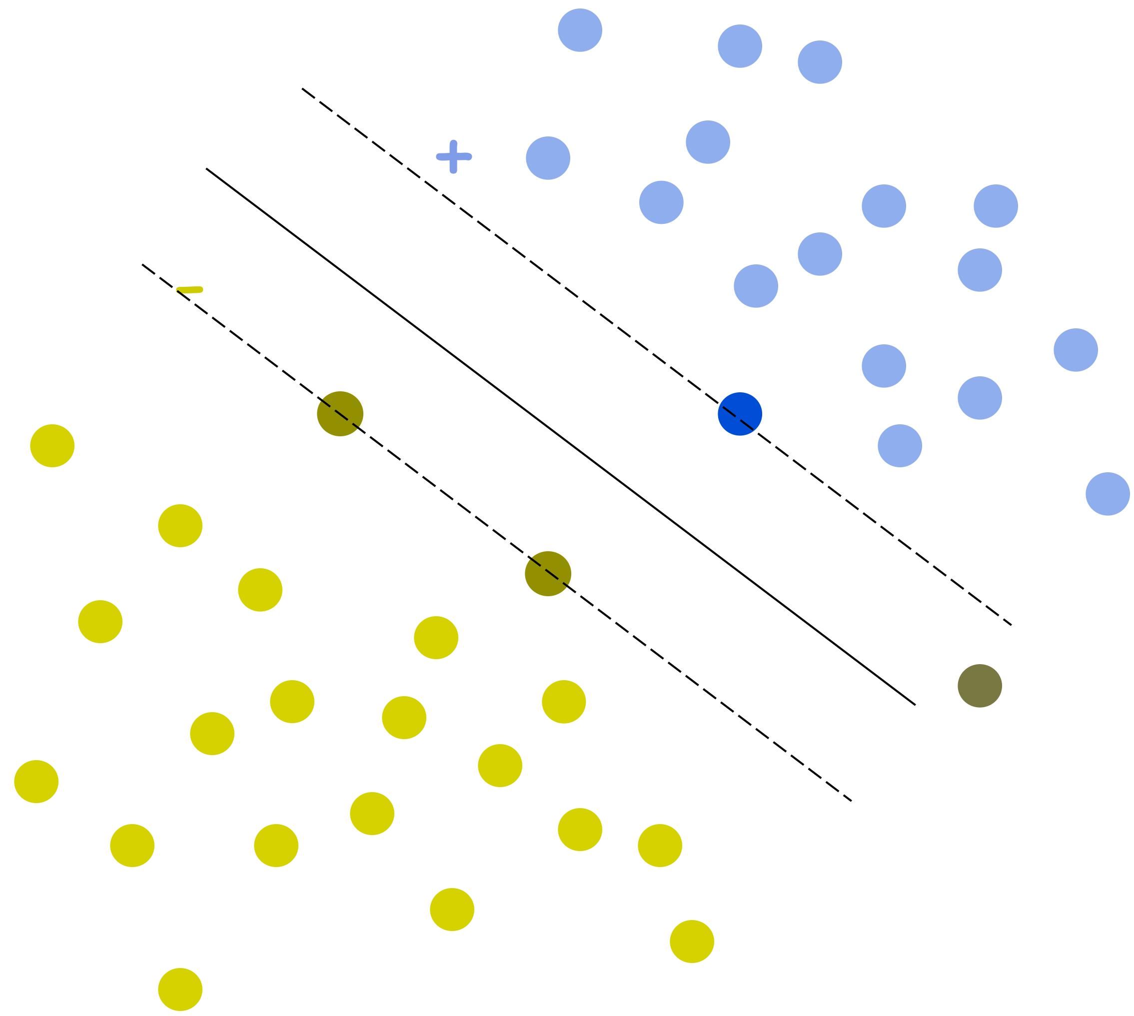
$\alpha=0$,
$\xi=0$

$\alpha=c$,
$\xi>0$

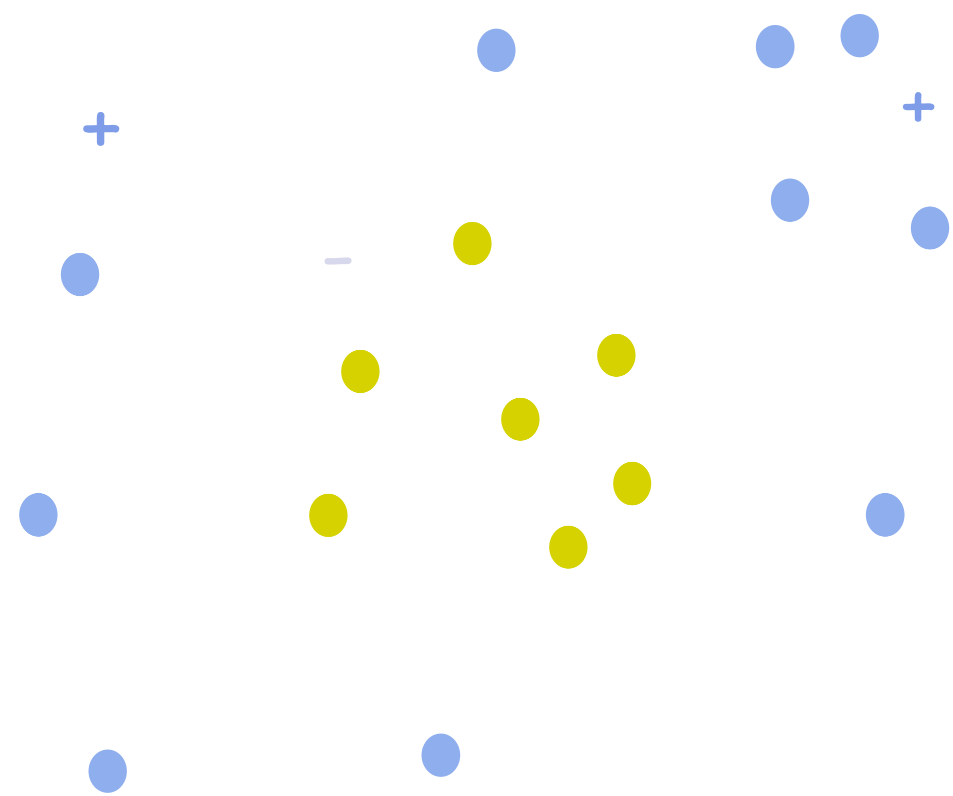


Support Vectors

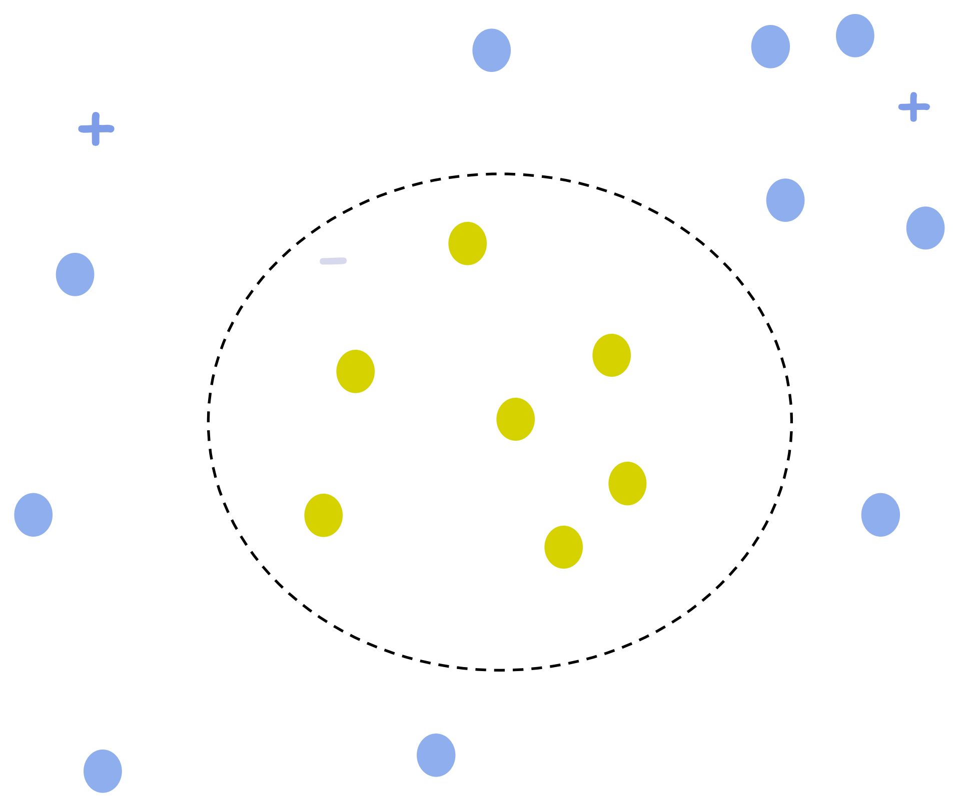
- $0<\alpha<c$, $\xi=0$
- $\alpha=0$, $\xi=0$
- $\alpha=c$, $\xi>0$



Oh, no! What do we do now?



Kernel SVM

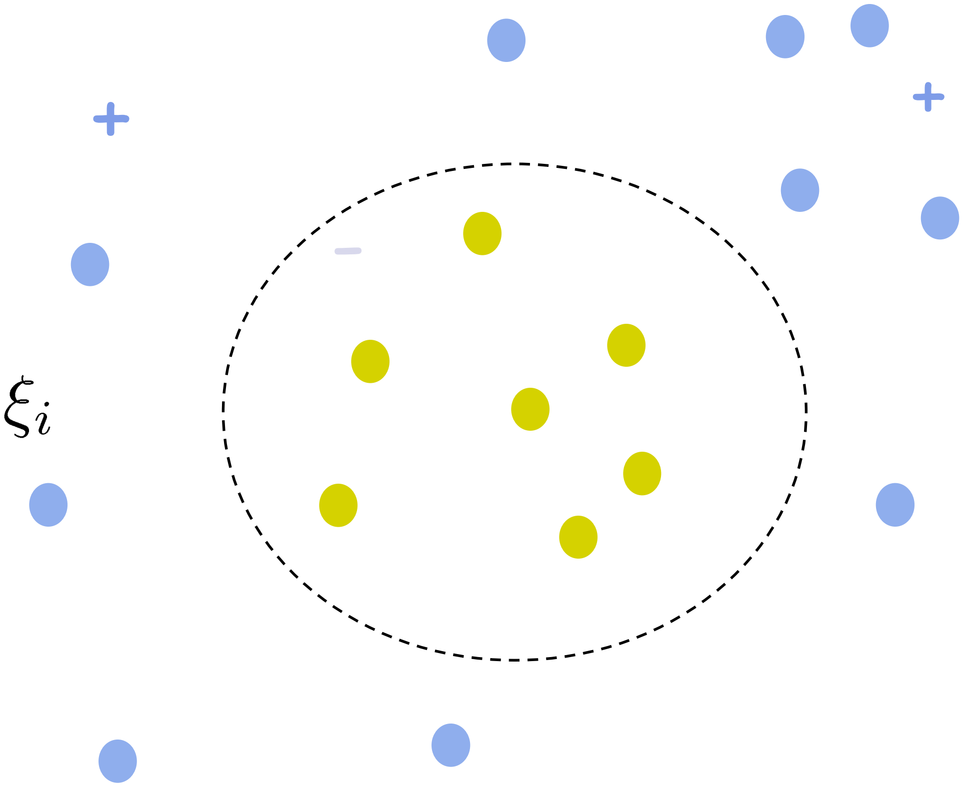


Kernel SVM

$$w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_1 x_2 + w_5 x_2^2 + \beta$$

$$w^\top \phi(x) + \beta, \;\; \phi(x) = (x_1, x_2, x_1^2, x_1 x_2, x_2^2) \in \mathbf{R}^5$$

$$y_i(w^\top \phi(x_i) + \beta) \geq 1 - \xi_i$$



Optimization Problem

At optimality $w^* = \sum_{i=1}^n \alpha_i y_i x_i, \quad 0 \le \alpha_i \le c$

$$||w||^2 = (\sum_{i=1}^n \alpha_i y_i x_i)^\top (\sum_{i=1}^n \alpha_i y_i x_i) = \sum_{i,j=1}^n y_i y_j {x_i}^\top x_j \alpha_i \alpha_j$$

$$\begin{aligned}
\min_{\alpha,\beta,\xi} \quad & \frac{1}{2} \sum_{i,j=1}^n y_i y_j {x_i}^\top x_j \alpha_i \alpha_j + c \sum_{i=1}^n \xi_i \\
\text{s.t.} \quad & \sum_j y_i y_j {x_i}^\top x_j \alpha_j + y\beta + \xi_i \ge 1, \quad i = 1, \dots, n \\
& \xi_i \ge 0, \; 0 \le \alpha_i \le c, \qquad i = 1, \dots, n,
\end{aligned}$$

Optimization Problem

At optimality $w^* = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i), \quad 0 \le \alpha_i \le c$

$$||w||^2 = (\sum_{i=1}^{n} \alpha_i y_i \phi(x_i))^\top (\sum_{i=1}^{n} \alpha_i y_i \phi(x_i)) = \sum_{i,j=1}^{n} y_i y_j \phi(x_i)^\top \phi(x_j) \alpha_i \alpha_j$$

$$\begin{aligned} \min_{\alpha,\beta,\xi} \quad & \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \phi(x_i)^\top \phi(x_j) \alpha_i \alpha_j + c \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & \sum_{j} y_i y_j \phi(x_i)^\top \phi(x_j) \alpha_j + y\beta + \xi_i \ge 1, \quad i = 1, \ldots, n \\ & \xi_i \ge 0, \; 0 \le \alpha_i \le c, \qquad i = 1, \ldots, n, \end{aligned}$$

How many variables? Constraints?

Kernel SVM

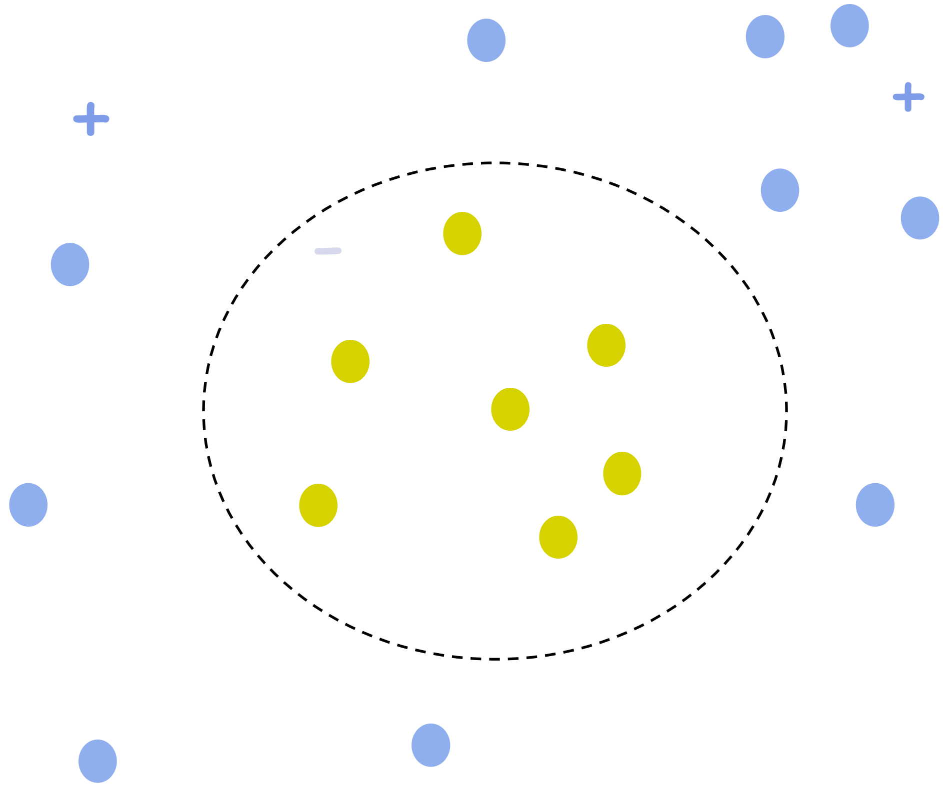
$$\phi(x) = (x_1, x_2, \tfrac{1}{\sqrt{2}}x_1^2, x_1x_2, \tfrac{1}{\sqrt{2}}x_2^2)$$

$$\phi(x)^\top\phi(z) = (x_1z_1 + x_2z_2 + \tfrac{1}{2}x_1^2z_1^2 + x_1x_2z_1z_2 + \tfrac{1}{2}x_2^2z_2^2)$$

$O(m^2)$

$$\phi(x)^\top\phi(z) = \tfrac{1}{2}(x_1z_1 + x_2z_2 + 1)^2 - 1 = \tfrac{1}{2}(x^\top z)^2 - 1$$

$O(m)$



Kernel SVM

$Q_{ij} = y_i y_j {x_i}^\top x_j \;\rightarrow\; Q_{ij} = y_i y_j \phi(x_i)^\top \phi(x_j) = y_i y_j K(x_i, x_j)$

Kernel operation: $K(x_i, x_j) = \phi(x_i)^\top \phi(x_j)$

Examples:

- $K(x_i, x_j) = ({x_i}^\top x_j / a_1 + a_2)^d$
- $K(x_i, x_j) = \exp^{-||x_i - x_j||^2 / 2\sigma^2}$

$$\Downarrow$$

$$\phi(x) \in R^\infty$$