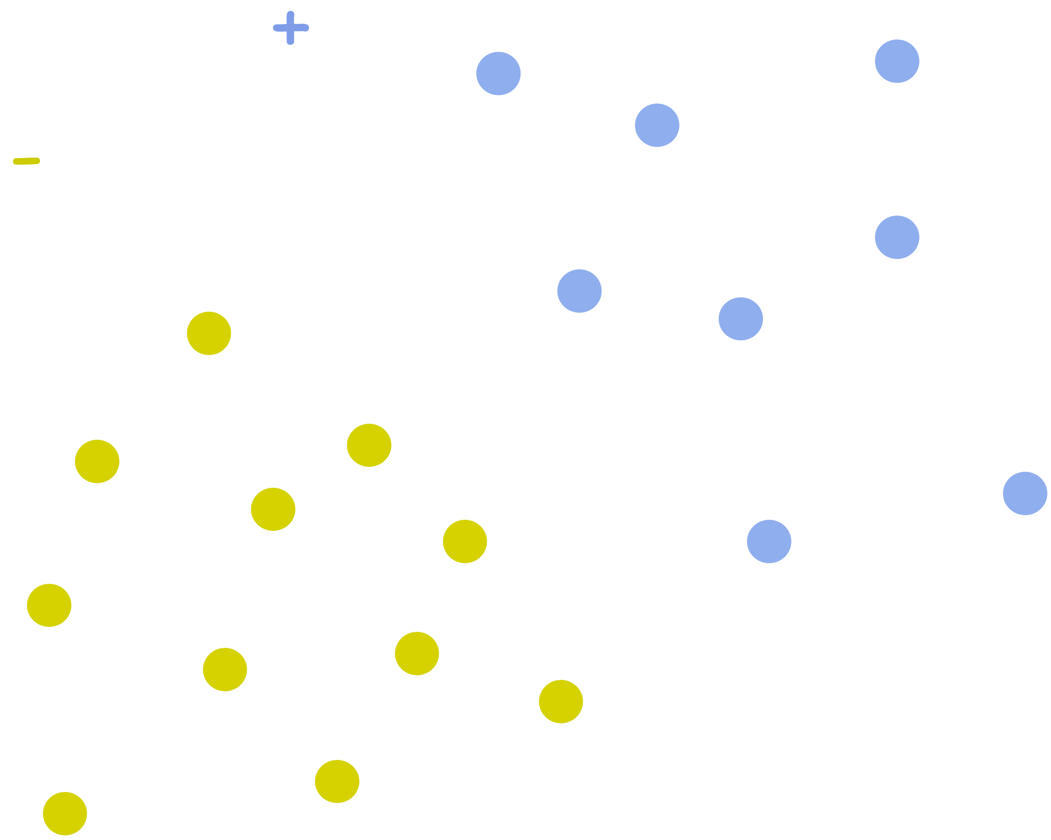
Optimization Problems in Machine Learning

Katya Scheinberg
Lehigh University

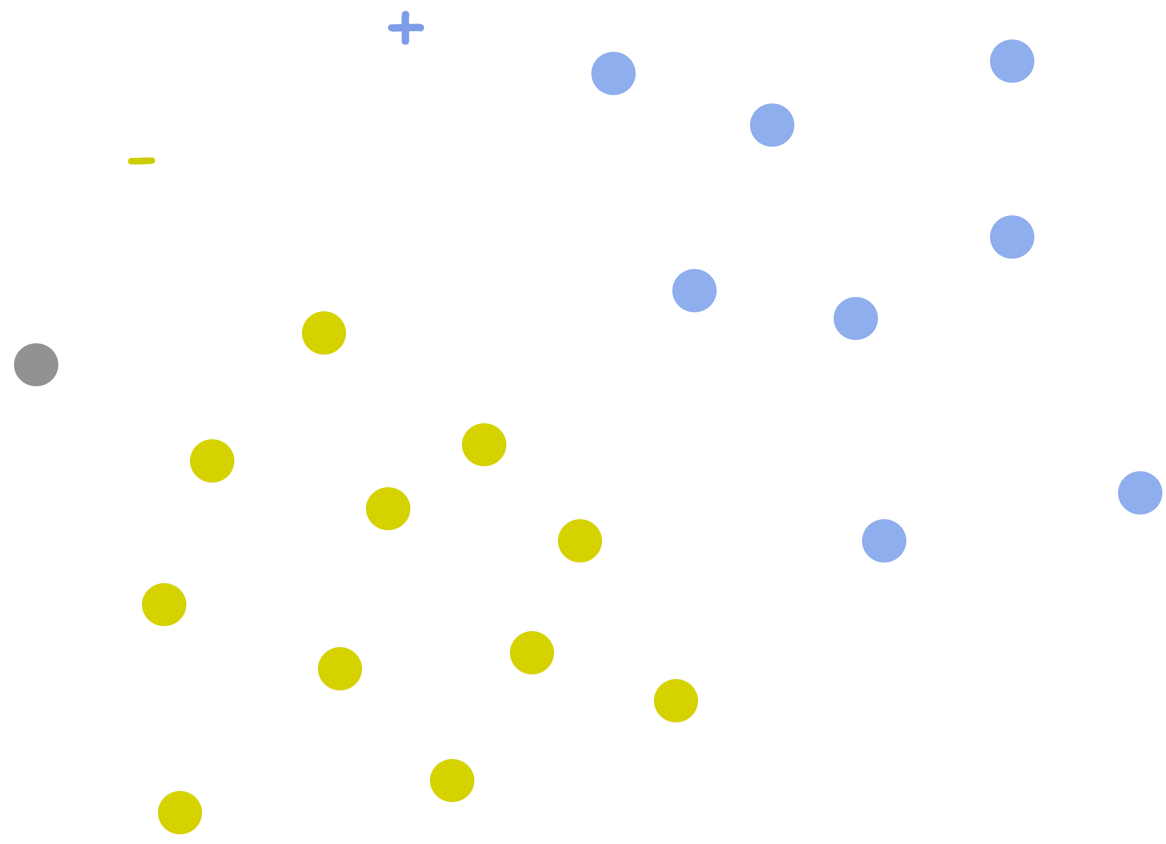
Binary classification problem

Two sets of labeled points



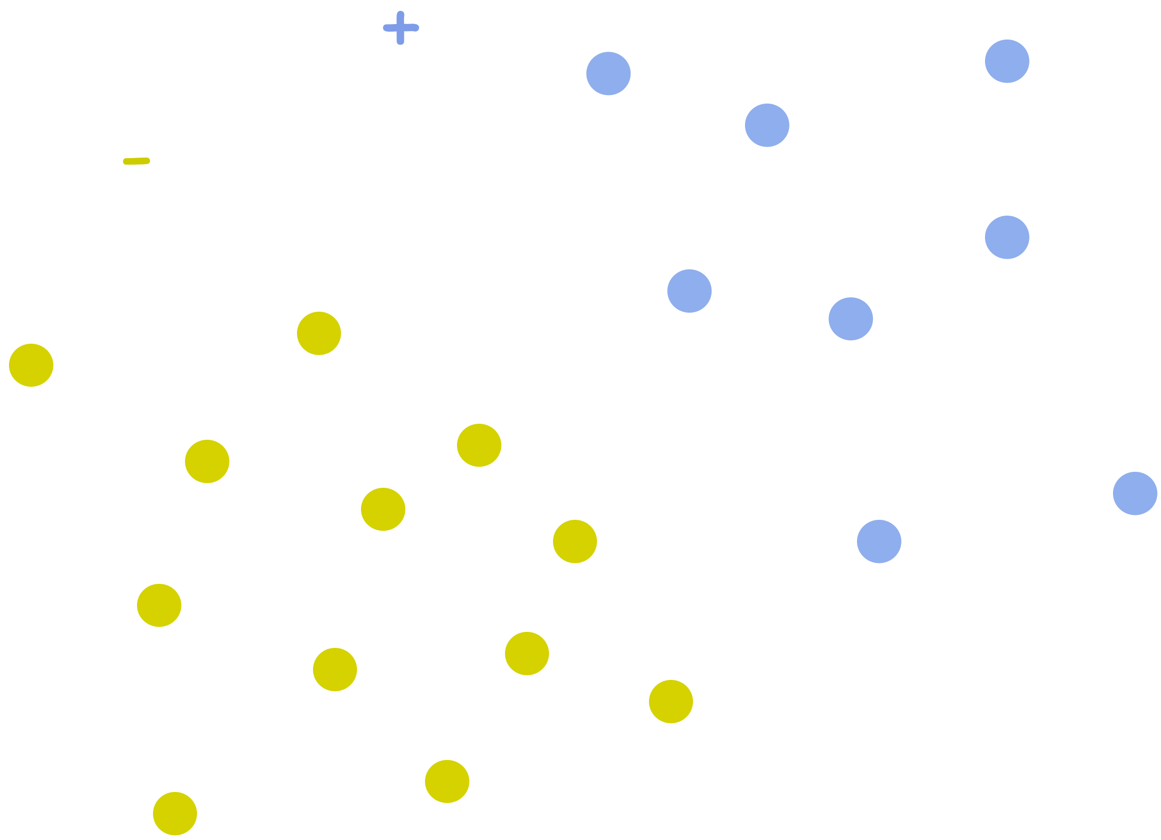
Binary classification problem

How to label
this new point?

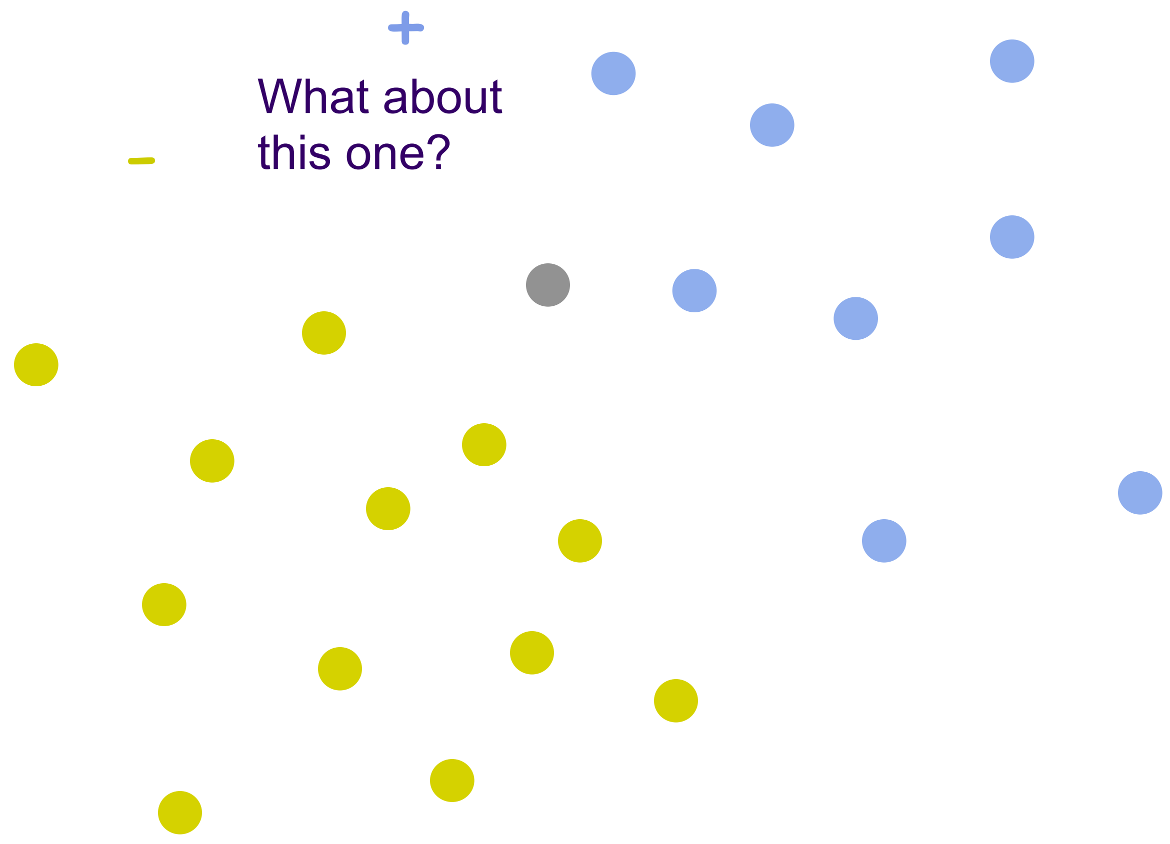


Binary classification problem

Probably green



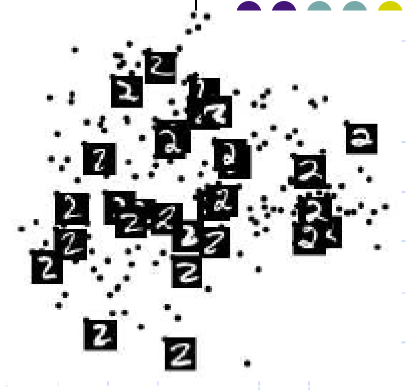
Binary classification problem



Binary classification problem

Or this one?

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

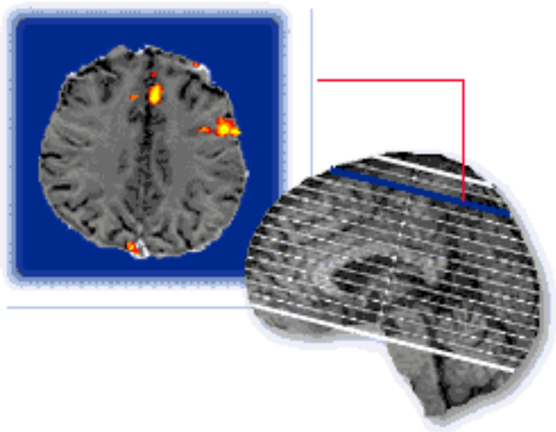
Examples from text and internet

- 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

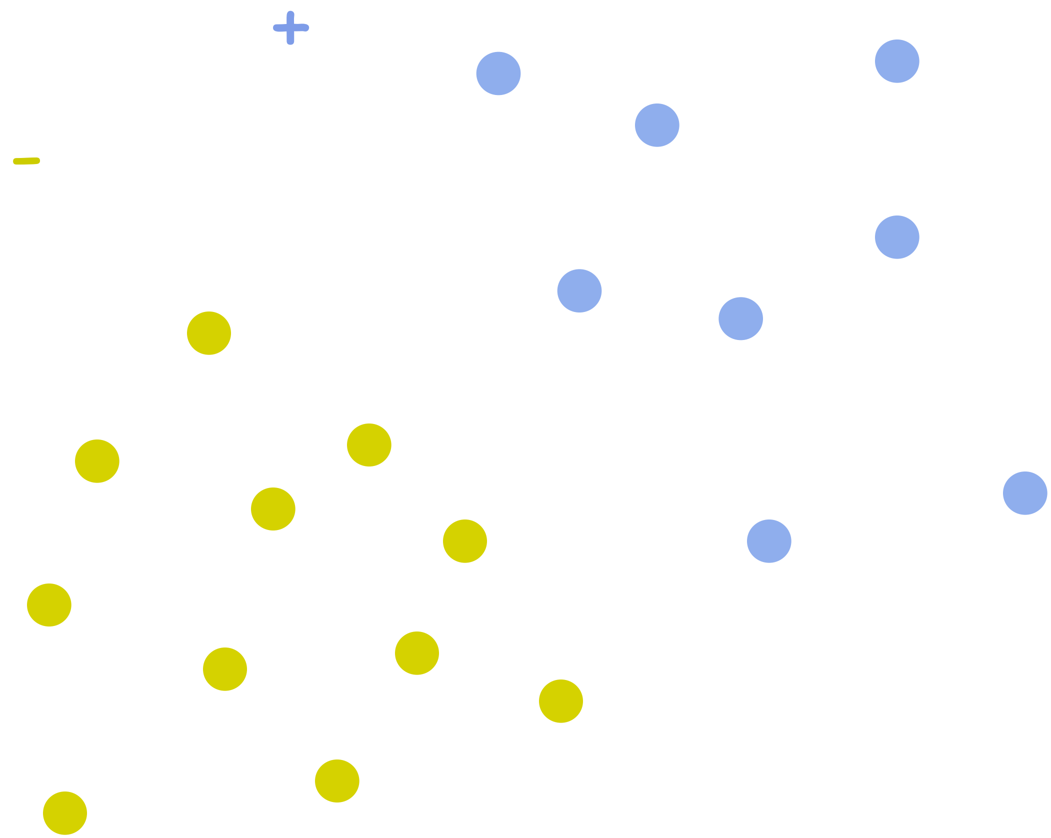
- The universe of data-label pairs (x, y),
- $y \in \{+1, -1\}$ for all $x \in \mathbf{R}^m$.
- Given a set $X \subset \mathbf{R}^m$ of n vectors.
- For each $x_i \in X$ the label y_i is known.
- Find a function $f(x) \approx y$

Example 1

SUPPORT VECTOR MACHINES

Linear classifier

Idea: separate a space into two half-spaces

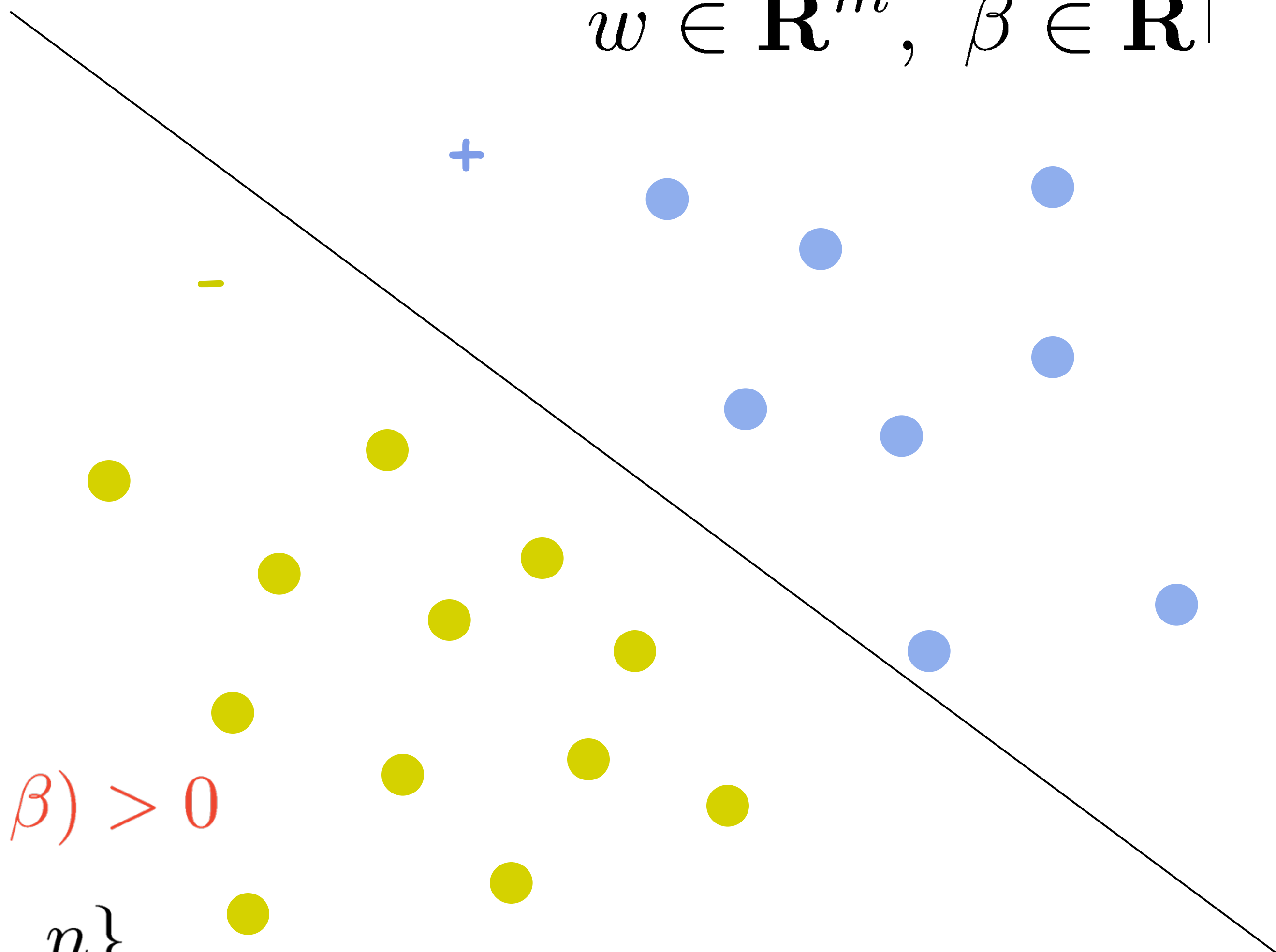


Linear classifier

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

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

Like this:



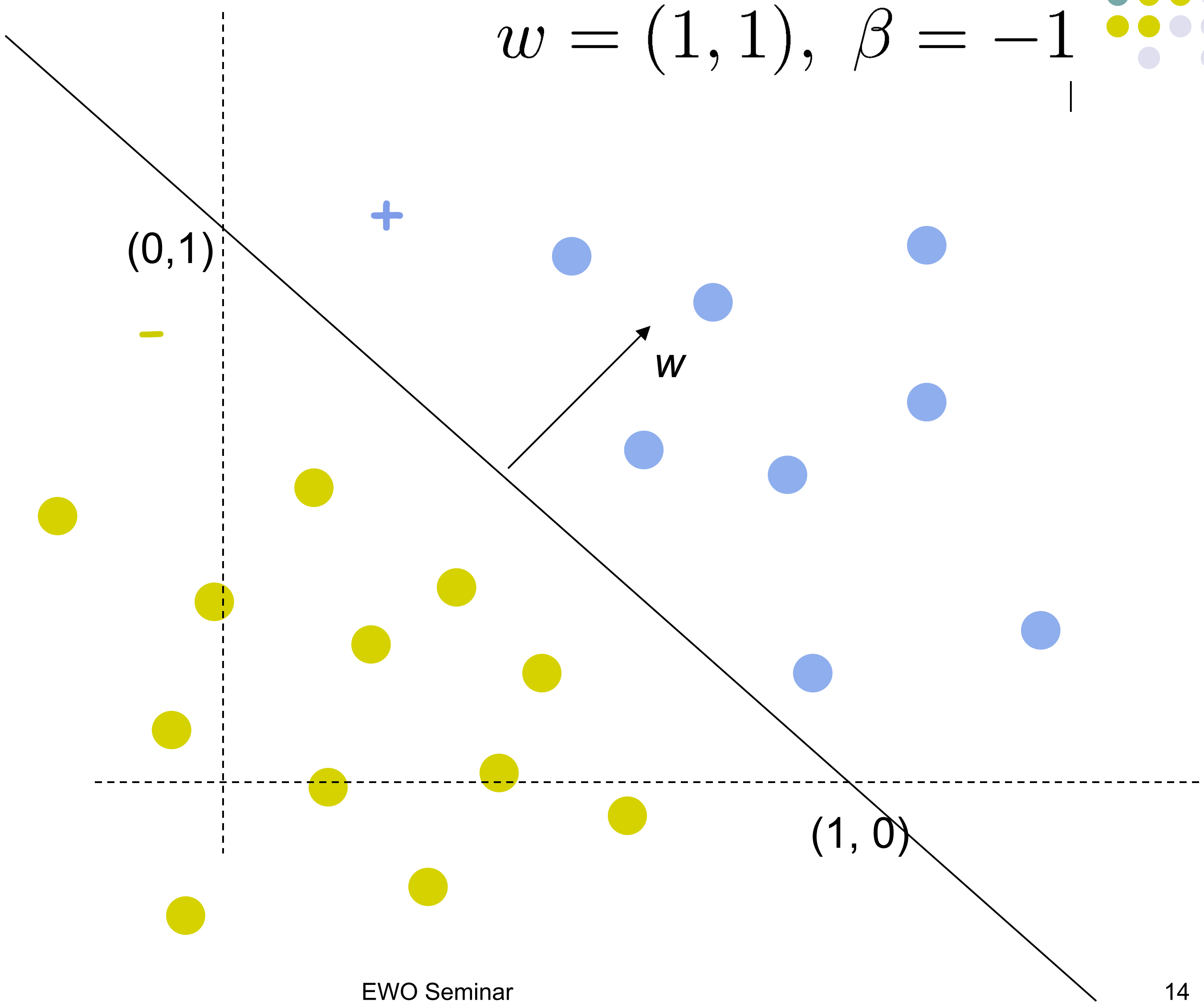
$$y_i(w^\top x_i + \beta) > 0$$

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

Linear classifier

$$x_1 + x_2 - 1 = 0$$

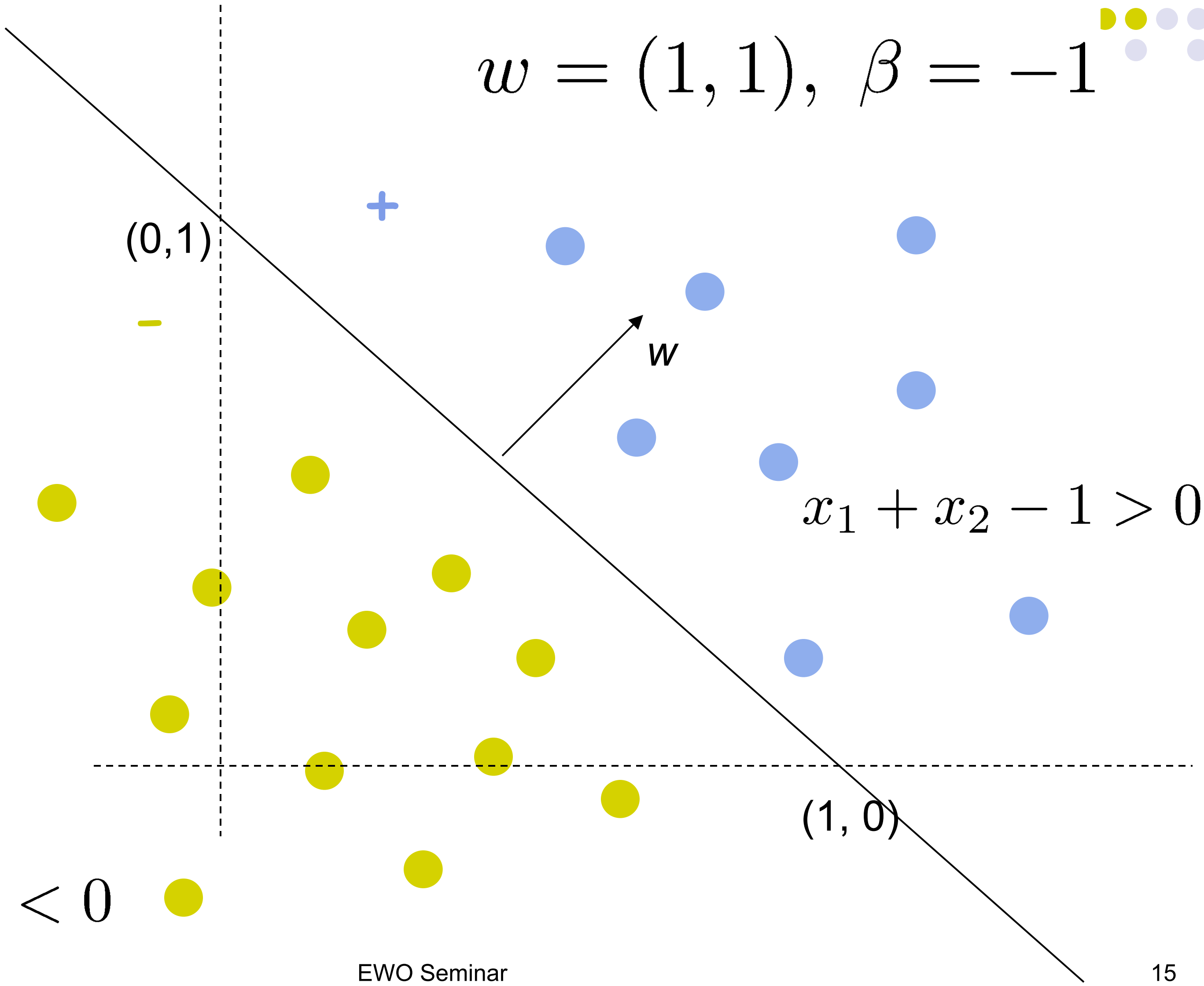
$$w = (1,1),\ \beta = -1$$



Linear classifier

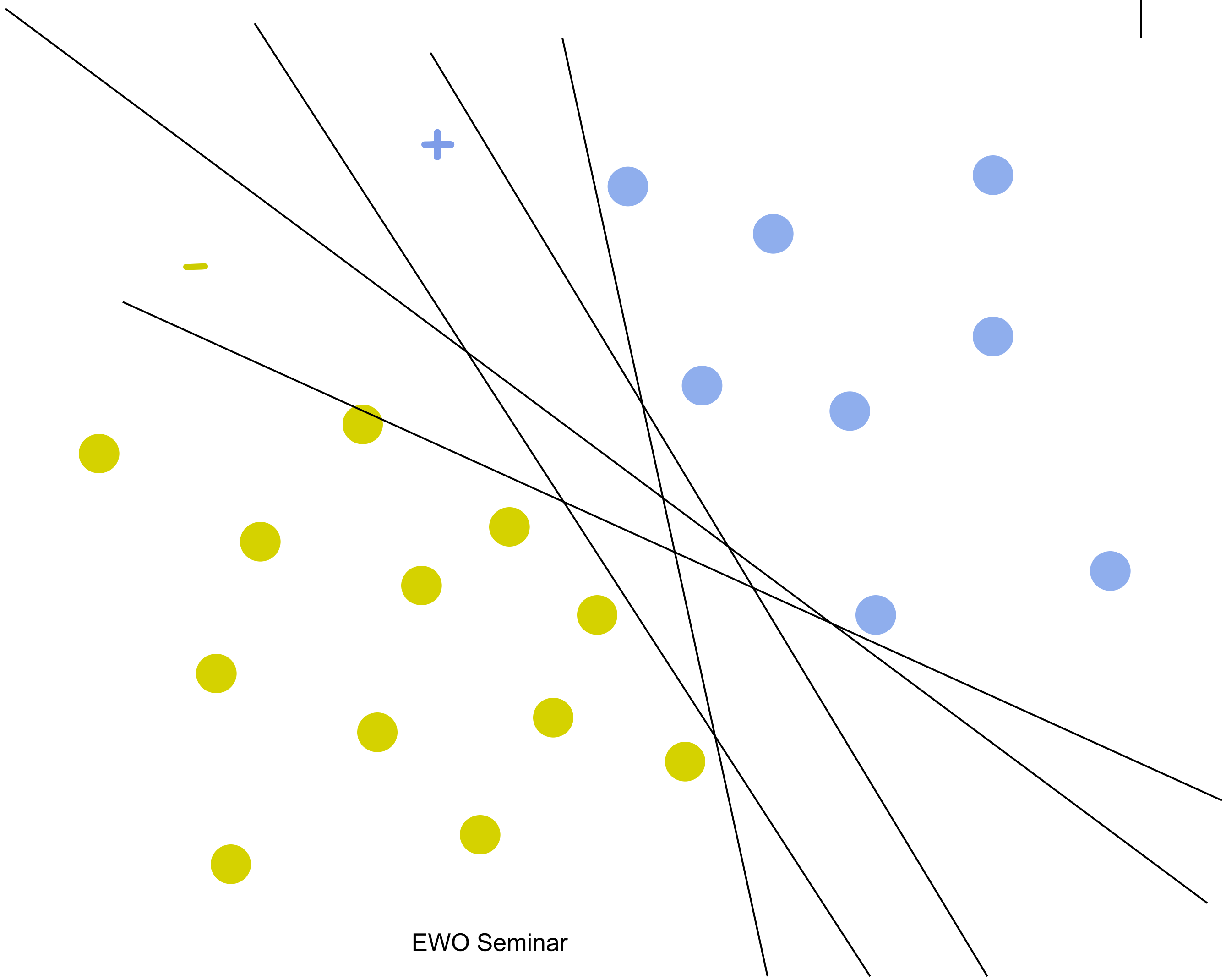
$$x_1 + x_2 - 1 = 0$$

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



$$x_1 + x_2 - 1 < 0$$

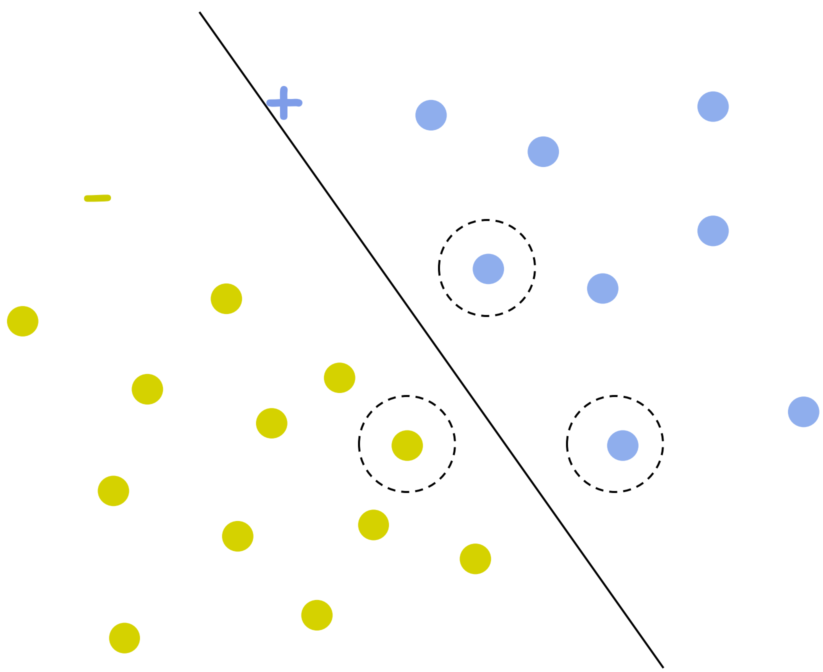
Linear classifier



Support vector machines

Assume each x_i is not known exactly,

but $z_i \in B(x_i, r)$



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

$$\Downarrow$$

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

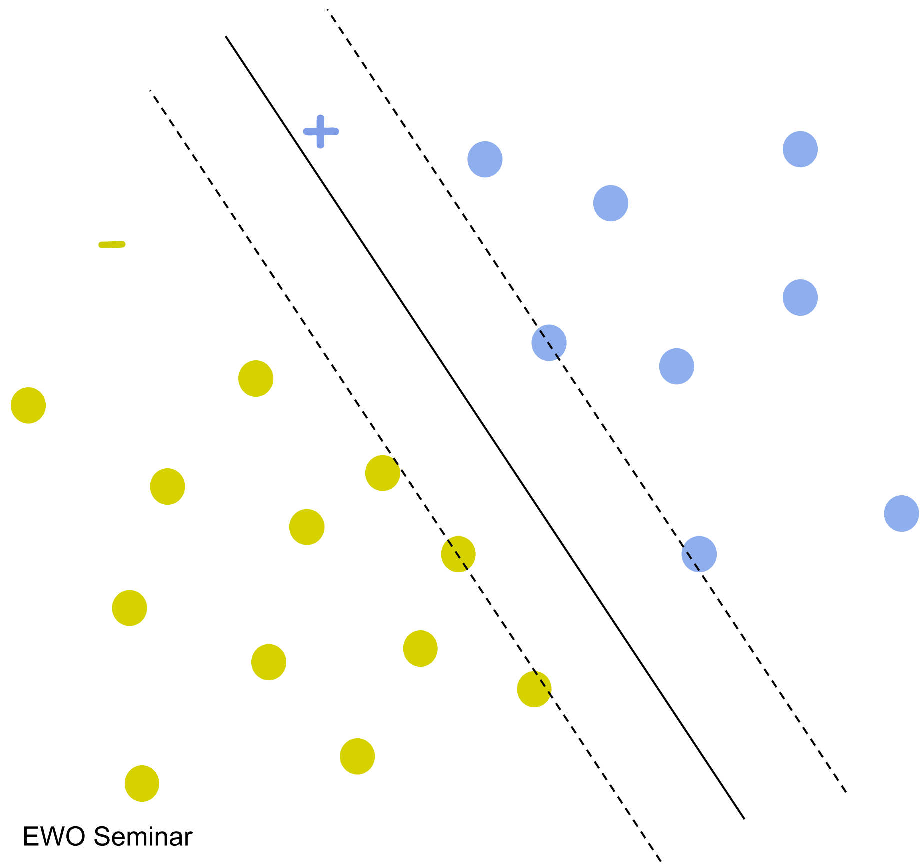
$$\Downarrow$$

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

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

Support vector machines

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



Optimization Problem

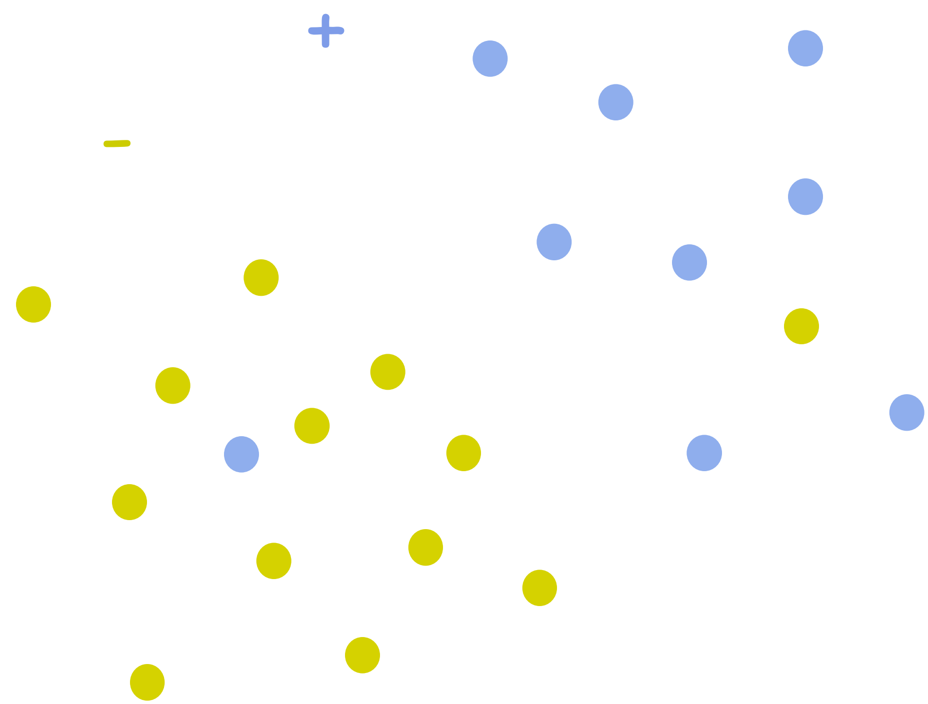
Total number of data points: n

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

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

$$\min_{\xi,w,\beta} \quad \frac{1}{2}w^\top w + c\sum_{i=1}^{n}\xi_i$$

$$\text{s.t.} \quad y_i(w^\top x_i + \beta) \geq 1 - \xi_i, \quad i = 1,\ldots,n$$

$$\xi \geq 0, \quad i = 1,\ldots,n.$$

How many variables? Constraints?

Soft margin SVM

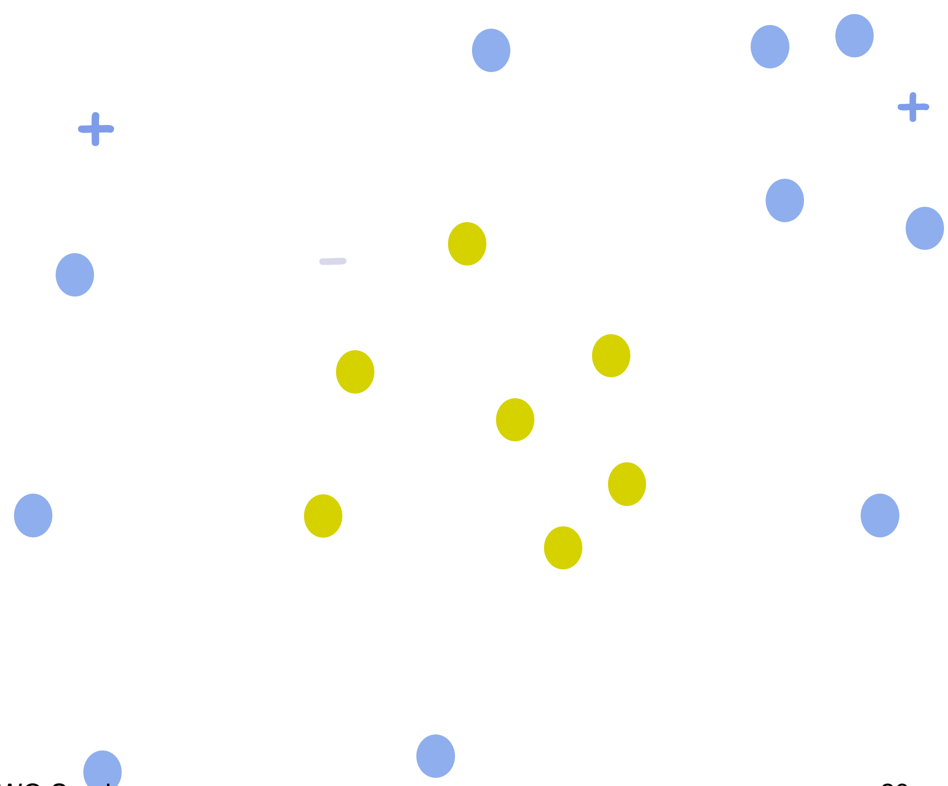
Total number of data points: n

$$\min_{w,\beta} \qquad \frac{1}{2} w^\top w + c \sum_{i=1}^{n} \max\{0, 1 - y_i(w^\top x_i + \beta)\}$$

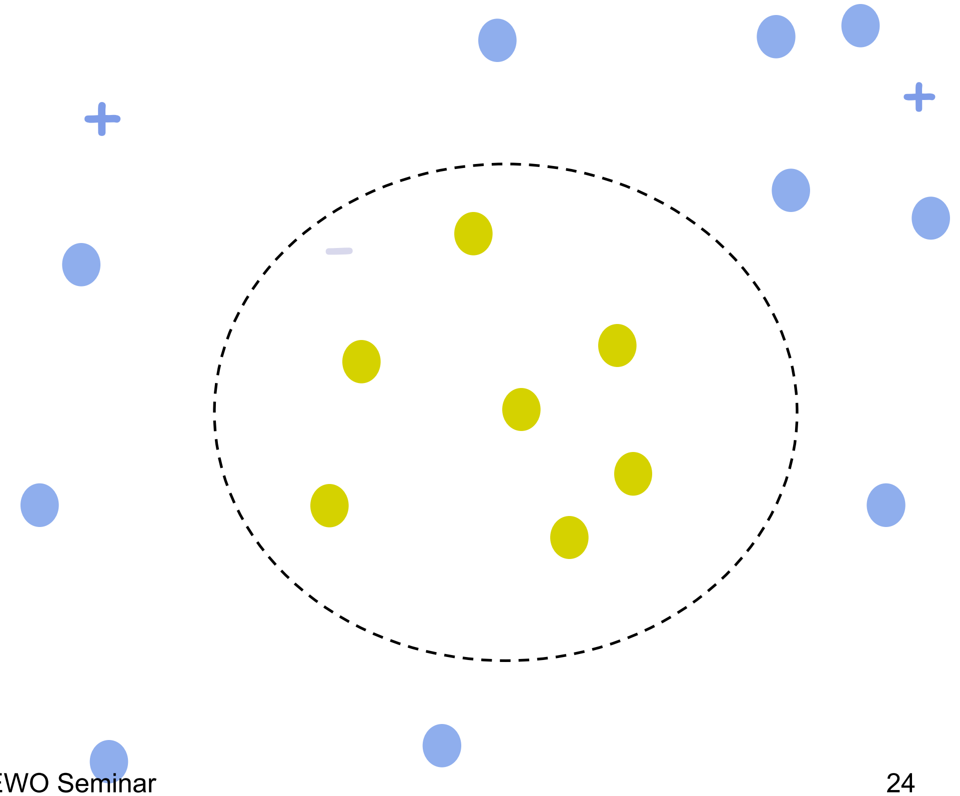
No constraints, but nonsmooth objective

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

Oh, no! What do we do now?



Kernel SVM

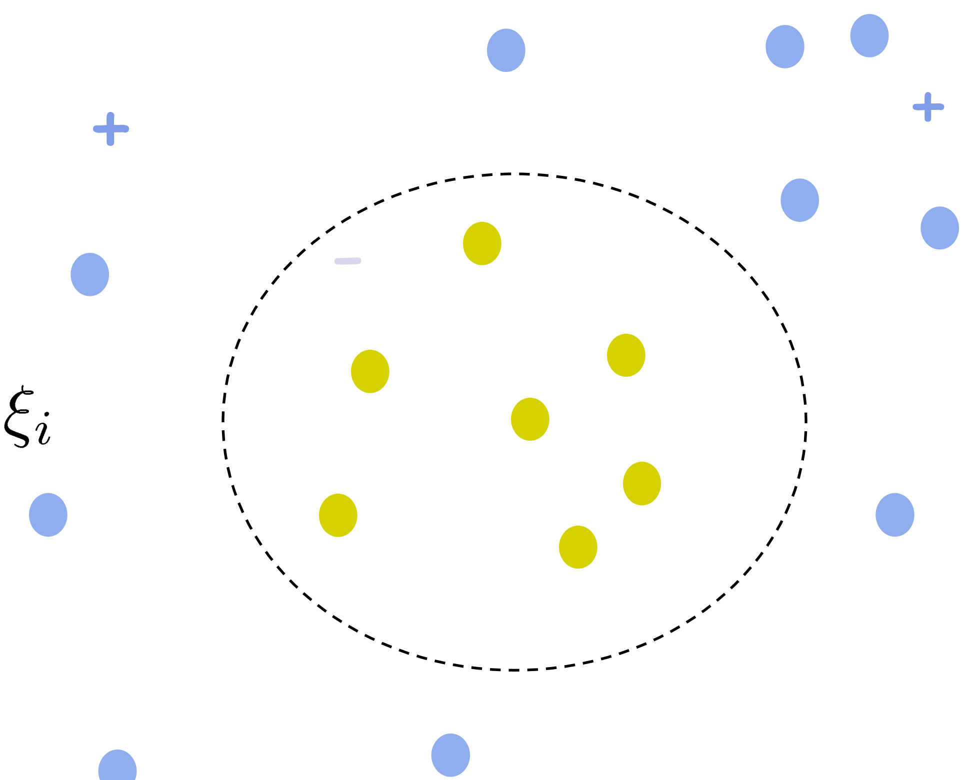


Kernel SVM

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

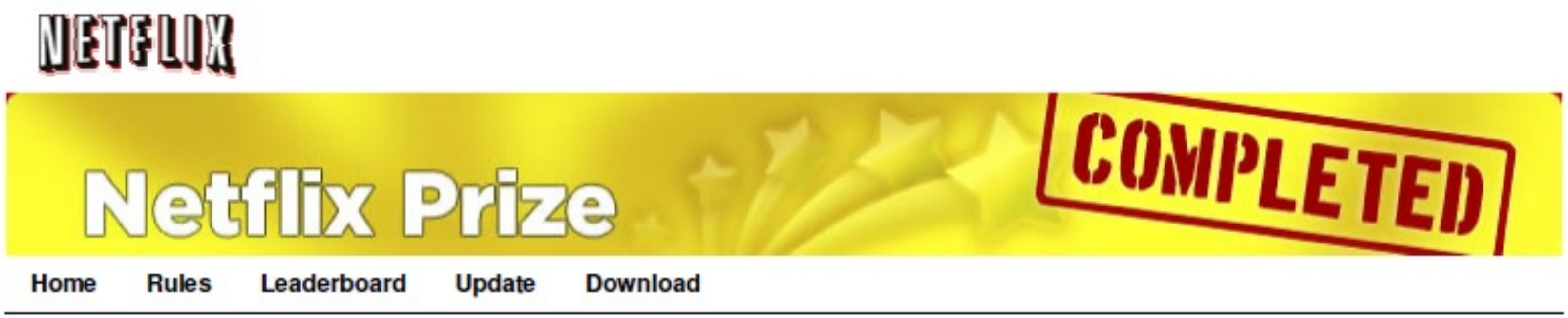
$$w^\top \phi(x) + \beta, \;\; \phi(x) = (x_1, x_2, x_1^2, x_1 x_2, x_2^2) \in \mathbf{R}^5$$

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



Example 2

COLLABORATIVE FILTERING, NETFLIX CHALLENGE



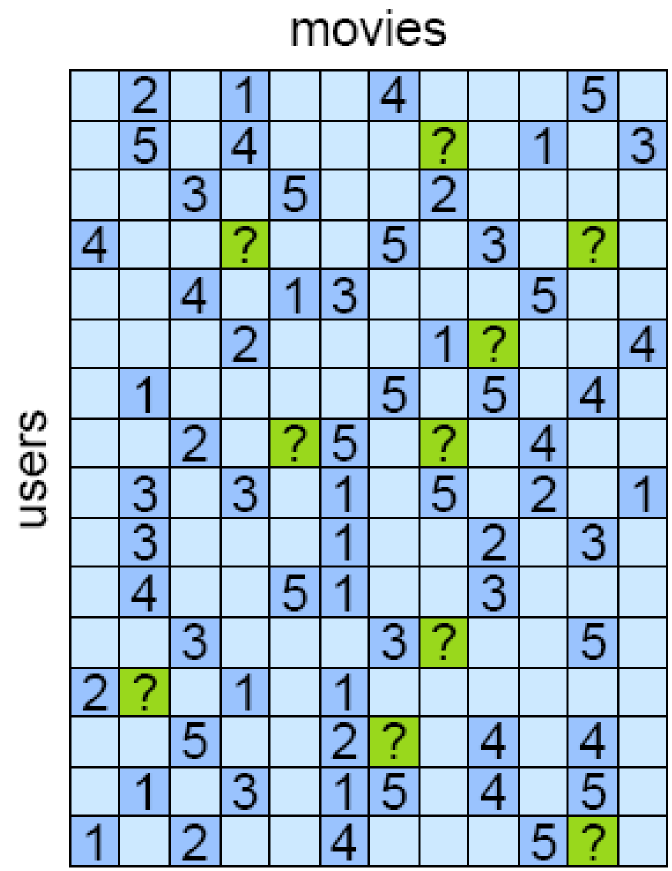
- Some users rate some movies they watched (or didn't!)
- Predict the rating (1..5) for each user/ movie pair.
- Use this prediction to recommend users the movies that they would like

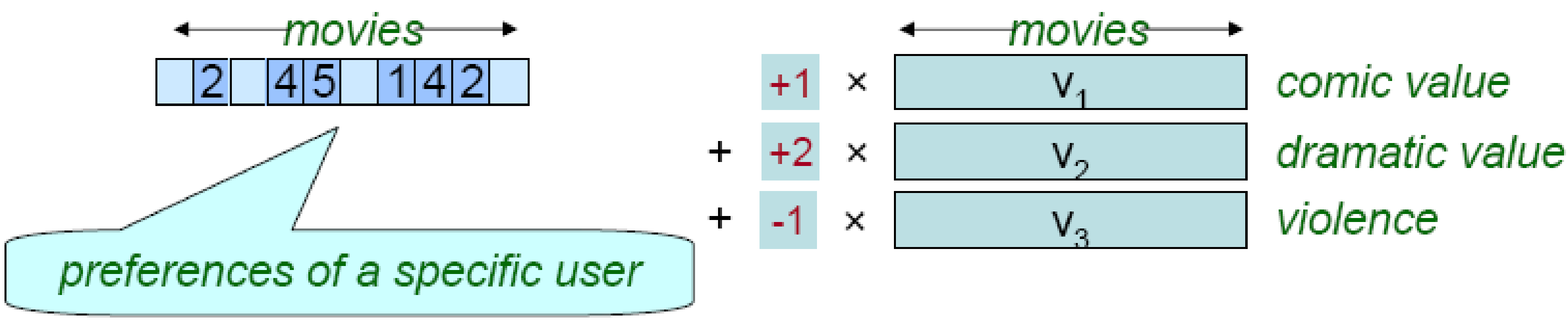
Matrix completion problem, collaborative filtering

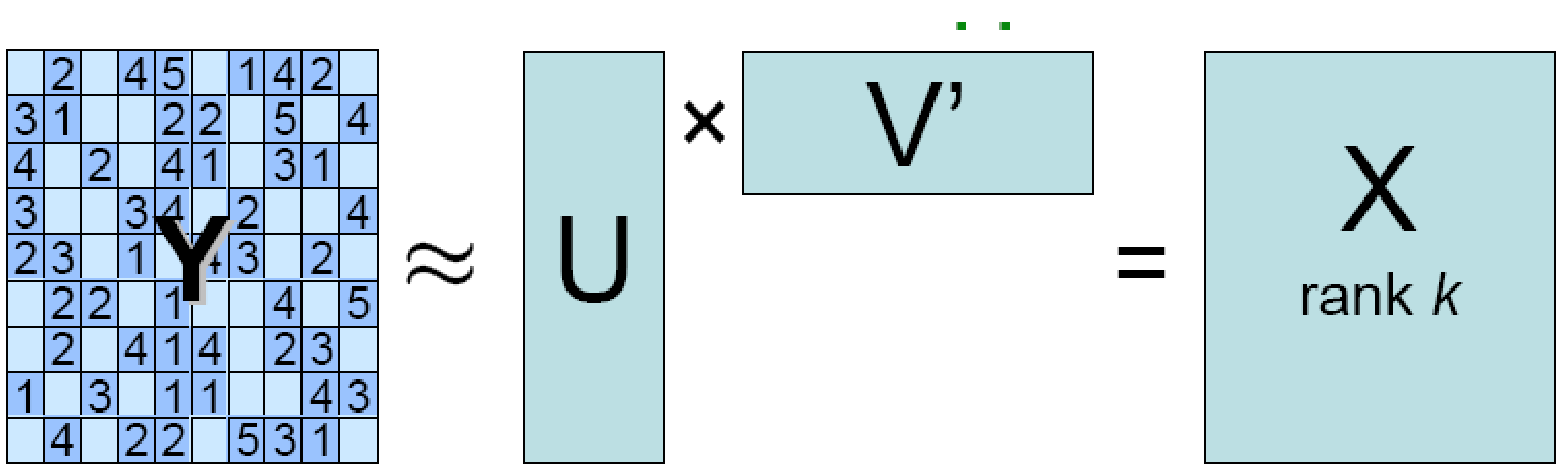
Collaborative filtering: famous Netflix challenge

Will user i like movie j?

Complete the matrix based on partially filled information.




movies
2 4 5 1 4 2
preferences of a specific user
movies
+1 × v1
+ +2 × v2
+ -1 × v3
comic value
dramatic value
violence


Y
≈
U
×
V'
=
X
rank k

Convex relaxation via nuclear norm

- Given the values for a subset of entries, find the matrix with these entries and the smallest (or given) rank.

$$\begin{aligned} \min_{X \in \mathbf{R}^{m \times n}} \quad & \mathbf{rank}(X) \\ \text{s.t.} \quad & X_{ij} = M_{ij}, \ (i,j) \in I \end{aligned}$$

- NP-hard problem.

$\mathbf{rank}(X) = \|\sigma(X)\|_0,$
where $\sigma(X)$ is the vector of the singular values.

$\|\cdot\|_0, \Rightarrow \|\cdot\|_1$ - the tightest convex relaxation.

$$\text{Nuclear norm: } \|X\|_* = \sum_{i=1}^{n} \sigma_i(X)$$

Convex relaxation via nuclear norm

- Given the values for a subset of entries, find the matrix with these entries and the smallest "nuclear norm".

$$\begin{aligned} \min_{X \in \mathbf{R}^{m \times n}} \quad & \|X\|_* \\ \text{s.t.} \quad & X_{ij} = M_{ij}, \ (i,j) \in I \end{aligned}$$

- Convex problem

Convex relaxation via nuclear norm

- Given the values for a subset of entries, find the matrix with similar entries and the smallest "nuclear norm".

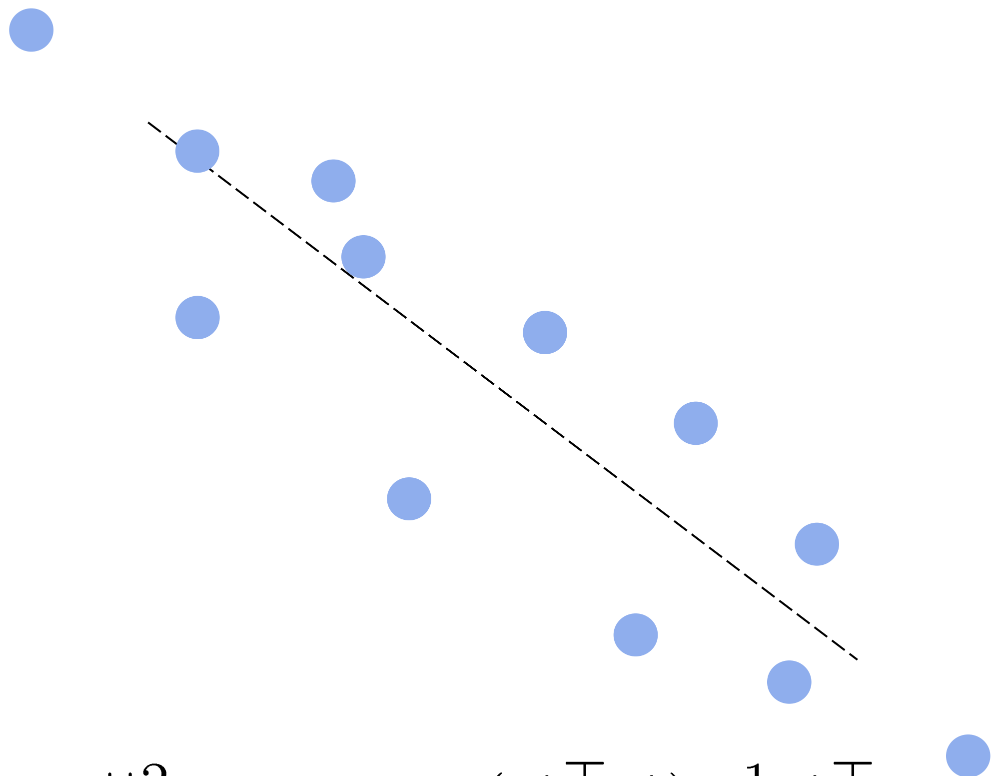
$$\begin{aligned} \min_{X \in \mathbf{R}^{m \times n}} \quad & \|X\|_* \\ \text{s.t.} \quad & |X_{ij} - M_{ij}| < \epsilon_{ij}, \ (i,j) \in I \end{aligned}$$

- Or

$$\min_{X \in \mathbf{R}^{m \times n}} \quad \|X\|_* + \rho \sum_{(i,j) \in I} (X_{ij} - M_{ij})^2$$

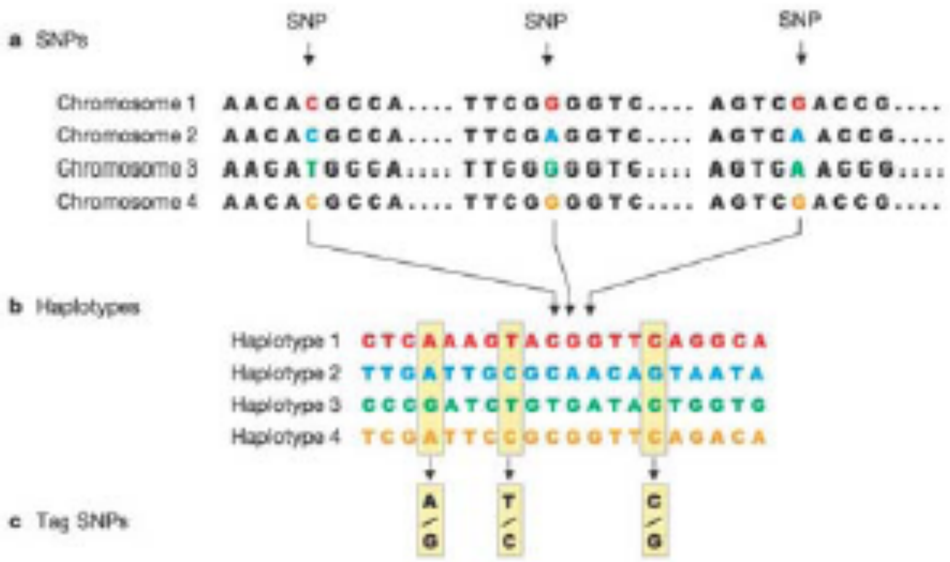
SPARSE REGRESSION, LASSO

Least Squares Linear Regression



$$\min_{x \in \mathbf{R}^n} ||Ax - b||_2^2 \;\Rightarrow\; x = (A^\top A)^{-1} A^\top b$$

Disease state prediction



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

- Data: Normalized hybridization intensities for each allele of a SNP
- Label: Disease state
- Problem size: Approx. 600,000 SNPs and 5,000 individuals[7]

Least squares problem

Standard form of LS problem

$$\min_{x \in \mathbf{R}^n} ||Ax - b||_2^2 \;\Rightarrow\; x = (A^\top A)^{-1} A^\top b$$

A has 500000 columns and 5000 rows – underdetermined.

Regularized regression can be used

$$\min_{x \in \mathbf{R}^n} ||Ax - b||_2^2 + \lambda ||x||_2^2 \;\Rightarrow\; x = (A^\top A + I)^{-1} A^\top b$$

x is going to be dense – hence linear combination of all factors (genes)

We would prefer to find a linear combinations of as few genes as possible

$$\min_{x \in \mathbf{R}^n} ||Ax - b||_2^2 + \lambda ||x||_0 \;\Rightarrow\; \mathrm{NP-hard\ problem}$$

Lasso and other formulations to recover structure

Sparse regularized regression or Lasso:

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

Sparse regressor selection

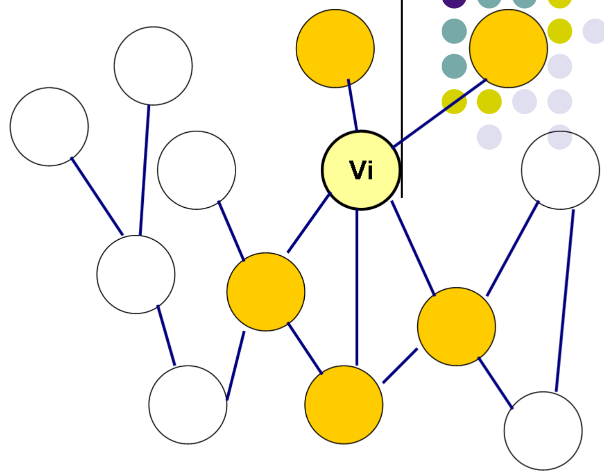
$$\begin{aligned} \min \quad & ||Ax - b|| \\ s.t. \quad & ||x||_1 \leq t. \end{aligned}$$

Noisy signal recovery

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

SPARSE INVERSE COVARIANCE SELECTION

Sparse inverse covariance selection



p random varibles

$$\boldsymbol{x} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$$

Multivariate Gaussian probability density function:

$$P(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- $\Sigma \in R^{n \times n}$ - covariance matrix
- Zeros in Σ^{-1} : conditional independence
- Sparsity of Σ^{-1} : better interpretability

Optimizing log likelihood

- $\max_\Sigma \log(P(X)) = \max_\Sigma \frac{m}{2}\log(\det(\Sigma^{-1})) - \frac{1}{2}Tr((XX^\top)\Sigma^{-1})$

- Let $A = \frac{1}{m}XX^\top$

- $\Sigma^{-1} = \arg\max_C \frac{m}{2}(\log\det C - Tr(AC))$

- Solution $\Sigma^{-1} = A^{-1}$ - typically not sparse.

- Need to enforce sparsity of Σ^{-1}: Penalize for nonzeros

Enforcing sparsity

- Convex relaxation

$$\Sigma^{-1} = \arg\max_C \frac{m}{2}(\log\det C - Tr(AC)) - \rho||C||_1$$

$$(\|C\|_1 = \sum_{ij} |C_{ij}|)$$

- Convex optimization problem with unique solution for each ρ

SOLUTION APPROACHES

Examples

- Lasso $\min_x \quad \frac{1}{2}||Ax - b||^2 + \rho||x||_1$
- SVM $\min_{w,\beta} \quad \frac{1}{2}w^\top w + \rho \sum_{i=1}^{n} \max\{0, 1 - y_i(w^\top x_i + \beta)\}$
- Collaborative filtering $\min_{X \in \mathrm{R}^{n \times m}} \rho \sum_{(i,j) \in I} (X_{ij} - M_{ij})^2 + ||X||_*$
- Robust PCA $\min_{X \in \mathrm{R}^{n \times m}} \rho \|X_{ij} - M_{ij}\|_1 + ||X||_*$
- SICS $\max_X \frac{m}{2}(\log \det X - Tr(AX)) - \rho||X||_1$

Alternating directions (splitting) method

- Consider: $\min_x F(x) = f(x) + g(x)$

$\Updownarrow$

$$\min_{x,y} \quad f(x) + g(y)$$
$$\text{s.t.} \quad y = x$$

- Relax constraints via Augmented Lagrangian technique

$$\min_{x,y} f(x) + g(y) + \lambda^\top (y - x) + \frac{1}{2\mu}||y - x||^2 = Q_\lambda(x, y)$$

In our examples *f(x)* and *g(y)* are both such that the above functions are easy to optimize in x or y

A variant of alternating directions method

- $x^{k+1} = \min_x Q_\lambda(x, y^k)$
- $\lambda^{k+\frac{1}{2}} = \lambda^k + \frac{1}{\mu}(y^k - x^{k+1})$
- $y^{k+1} = \min_y Q_\lambda(x^{k+1}, y)$
- $\lambda^{k+1} = \lambda^{k+\frac{1}{2}} + \frac{1}{\mu}(y^{k+1} - x^{k+1})$

This turns out to be equivalent to……

Goldfarb, Ma and S, '10

Alternating linearization method (ALM)

- $x^{k+1} = \min_x Q_g(x, y^k)$
- $y^{k+1} = \min_y Q_f(x^{k+1}, y)$

$$Q_g(x, y) = f(x) + \nabla g(y)^\top (x - y) + \frac{1}{2\mu}||y - x||^2 + g(y)$$

$$Q_f(x, y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2\mu}||y - x||^2 + g(y)$$

Goldfarb, Ma, S, '10

What is involved?

- Theoretical convergence guarantees and convergence rates have been developed
- The real complexity depends on the choice of μ
- Various strategies for parameter selection affect performance and have extra costs.
- Depending on application minimization and gradient computations can be expensive.
- Inexact computations may be utilized but may lead to worse convergence properties.
- Parallelization? Stochastic sampling?

THANK YOU!