Mollabashi house Isfahan

Introduction to Optimization for Machine Learning

Hossein Moosaei, PhD

Postdoctoral Researcher
Department of Applied Mathematics
School of Computer Science
Charles University

4 November 2019

Optimization

Optimization Problem

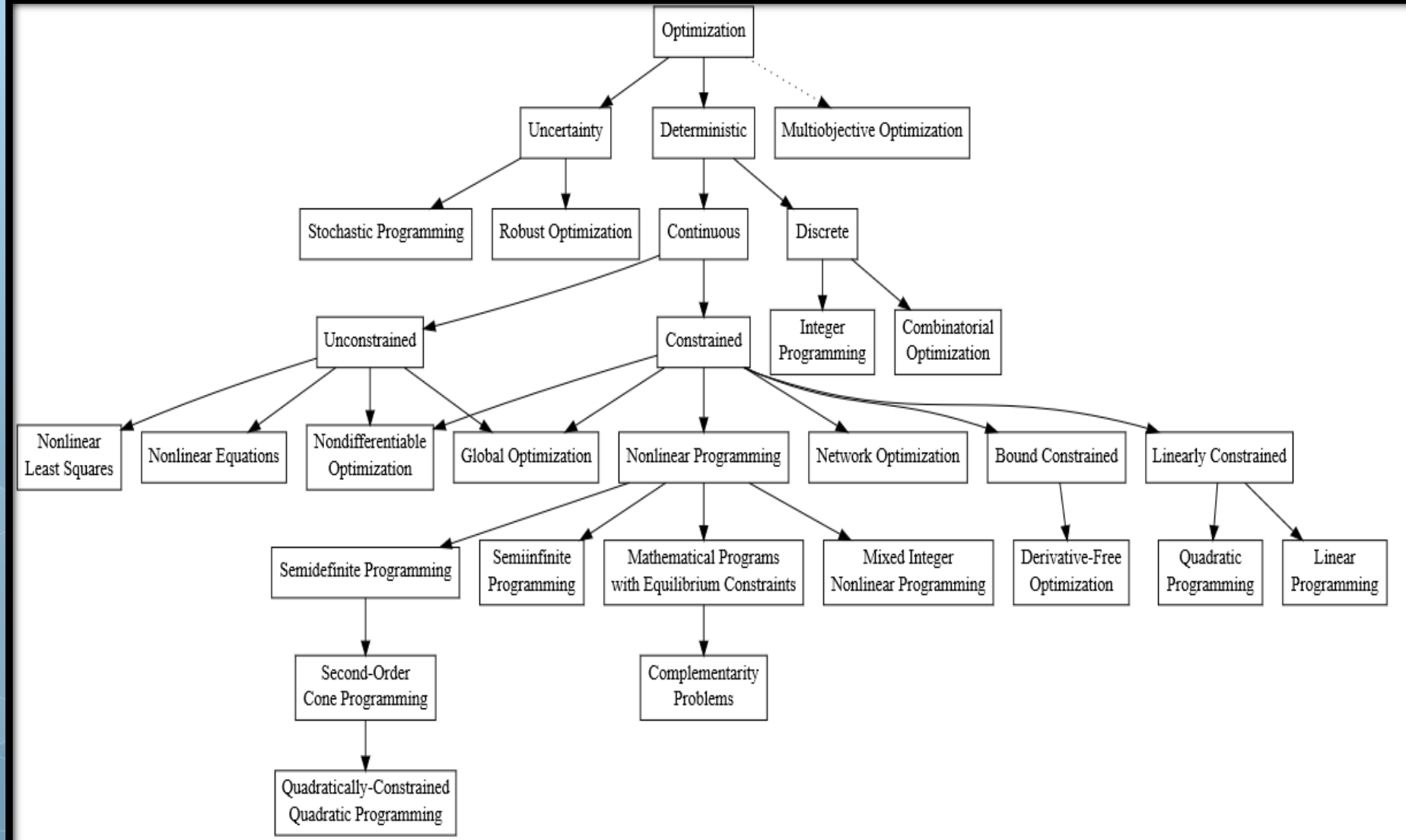
- Finding the minimizer of a function subject to constraints:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \; i = \{1, \ldots, k\} \\ & h_j(x) = 0, \; j = \{1, \ldots, l\} \end{aligned}$$

Some different types of optimization problems

?

Optimization Taxonomy



Applications of Optimization

?

Karush-Kuhn-Tucker Optimality Conditions

Optimality Criteria

- Big question: *How do we know that we have found the "optimum" for min f(x)?*

Answer: Test the solution for the "necessary and sufficient conditions"

Optimality Conditions – Unconstrained Case

- Let x* be the point that we think is the minimum for f(x)

Necessary condition (for optimality):

$$\nabla f(x^*) = 0$$

- A point that satisfies the necessary condition is a stationary point It can be a minimum, maximum, or saddle point

- *How do we know that we have a minimum?*
- Answer: Sufficiency Condition:

  The sufficient conditions for x* to be a strict local minimum are:

$$\nabla f(x^*) = 0$$

$$\nabla^2 f(x^*) \text{ is positive definite}$$

Constrained Case – KKT Conditions

- To proof a claim of optimality in constrained minimization (or maximization), we have to check the found point with respect to the (Karesh) Kuhn Tucker conditions.

- Kuhn and Tucker extended the Lagrangian theory to include the general classical single-objective nonlinear programming problem:

minimize $f(x)$

Subject to $g_j(x) \geq 0$ for $j = 1, 2, ..., J$

$h_k(\mathbf{x}) = 0$ for $k = 1, 2, ..., K$

$x = (x_1, x_2, ..., x_N)$

Necessary KKT Conditions

For the problem:

Min $f(x)$

s.t. $g(x) \leq 0$

(n variables, m constraints)

The necessary conditions are:

$\nabla f(x) + \Sigma \mu_i \nabla g_i(x) = 0$ (optimality)

$g_i(x) \leq 0$ for $i = 1, 2, ..., m$ (feasibility)

$\mu_i g_i(x) = 0$ for $i = 1, 2, ..., m$ (complementary slackness condition)

$\mu_i \geq 0$ for $i = 1, 2, ..., m$ (non-negativity)

Note that the first condition gives n equations.

Necessary KKT Conditions (General Case)

- For general case (n variables, M Inequalities, L equalities):
  Min f(x)
  s.t.
  $g_i(x) \leq 0$ for i = 1, 2, ..., M
  $h_j(x) = 0$ for J = 1, 2, ..., L

- In all this, the assumption is that $\nabla g_j(x^*)$ for j belonging to active constraints and $\nabla h_k(x^*)$ for k = 1, ...,K are linearly independent

- The necessary conditions are:
  $\nabla f(x) + \Sigma \mu_i \nabla g_i(x) + \Sigma \lambda_j \nabla h_j(x) = 0$ (optimality)
  $g_i(x) \leq 0$ for i = 1, 2, ..., M (feasibility)
  $h_j(x) = 0$ for j = 1, 2, ..., L (feasibility)
  $\mu_i g_i(x) = 0$ for i = 1, 2, ..., M (complementary slackness condition)
  $\mu_i \geq 0$ for i = 1, 2, ..., M (non-negativity)
  (Note: λ_j is unrestricted in sign)

Restating the Optimization Problem

- Kuhn Tucker Optimization Problem: Find vectors $x_{(Nx1)}$, $\mu_{(1xM)}$ and $\lambda_{(1xK)}$ that satisfy:

  $\nabla f(x) + \Sigma \mu_i \nabla g_i(x) + \Sigma \lambda_j \nabla h_j(x) = 0$ (optimality)

  $g_i(x) \leq 0$ for i = 1, 2, ..., M (feasibility)

  $h_j(x) = 0$ for j = 1, 2, ..., L (feasibility)

  $\mu_i g_i(x) = 0$ for i = 1, 2, ..., M (complementary slackness condition)

  $\mu_i \geq 0$ for i = 1, 2, ..., M (non-negativity)

- If x* is an optimal solution to NLP, then there exists a (μ^*, λ^*) such that (x*, μ^*, λ^*) solves the Kuhn–Tucker problem.

- Above equations not only give the necessary conditions for optimality, but also provide a way of finding the optimal point.

Limitations

- Necessity theorem helps identify points that are not optimal. A point is not optimal if it does not satisfy the Kuhn–Tucker conditions.
- On the other hand, not all points that satisfy the Kuhn-Tucker conditions are optimal points.
- The Kuhn–Tucker sufficiency theorem gives conditions under which a point becomes an optimal solution to a single-objective NLP.

Sufficiency Condition

- Sufficient conditions that a point x* is a strict local minimum of the NLP problem, where f, g_j, and h_k are twice differentiable functions are that

1) The necessary KKT conditions are met.
2) The Hessian matrix $\nabla^2 L(x^*) = \nabla^2 f(x^*) + \Sigma \mu_i \nabla^2 g_i(x^*) + \Sigma \lambda_j \nabla^2 h_j(x^*)$ is positive definite on a subspace of R^n as defined by the condition:

   $y^T \nabla^2 L(x^*)\, y \geq 0$ is met for every vector $y_{(1xN)}$ satisfying:

   $\nabla g_j(\mathbf{x^*})y < 0$ for j belonging to $I_1 = \{ j \mid g_j(\mathbf{x^*}) = 0, u_j^* > 0\}$ (active constraints)

   $\nabla h_k(\mathbf{x^*})y = 0$ for k = 1, ..., K

   $y \neq 0$

KKT Sufficiency Theorem (Special Case)

- Consider the classical single objective NLP problem.

  minimize $f(x)$

  Subject to $g_j(x) \leq 0$ for $j = 1, 2, ..., J$

  $h_k(\mathbf{x}) = 0$ for $k = 1, 2, ..., K$

- Let the objective function $f(x)$ be convex, the inequality constraints $g_j(x)$ be all convex functions for $j = 1, ..., J$, and the equality constraints $h_k(x)$ for $k = 1, ..., K$ be linear.
- If this is true, then the necessary KKT conditions are also sufficient.
- Therefore, in this case, if there exists a solution x^* that satisfies the KKT necessary conditions, then x^* is an optimal solution to the NLP problem.
- In fact, it is a global optimum.

Dual Problem

Generalized Lagrangian Function

- Consider the general (primal) optimization problem

$$\begin{aligned} &minimize \quad f(w) \\ &subject\ to \quad g_i(w) \le 0, i = 1, \cdots, k \\ &\qquad\qquad h_j(w) = 0, \ j = 1, \cdots, m \end{aligned}$$

where the functions f, $g_i, i = 1, \cdots, k$, *and* $h_i, i = 1, \cdots, m$, are defined on a domain Ω. The generalized Lagrangian was defined as

$$\begin{aligned} L(w, \alpha, \beta) &= f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{j=1}^{m} \beta_j h_j(w) \\ &= f(w) + \alpha^T g(w) + \beta^T h(w) \end{aligned}$$

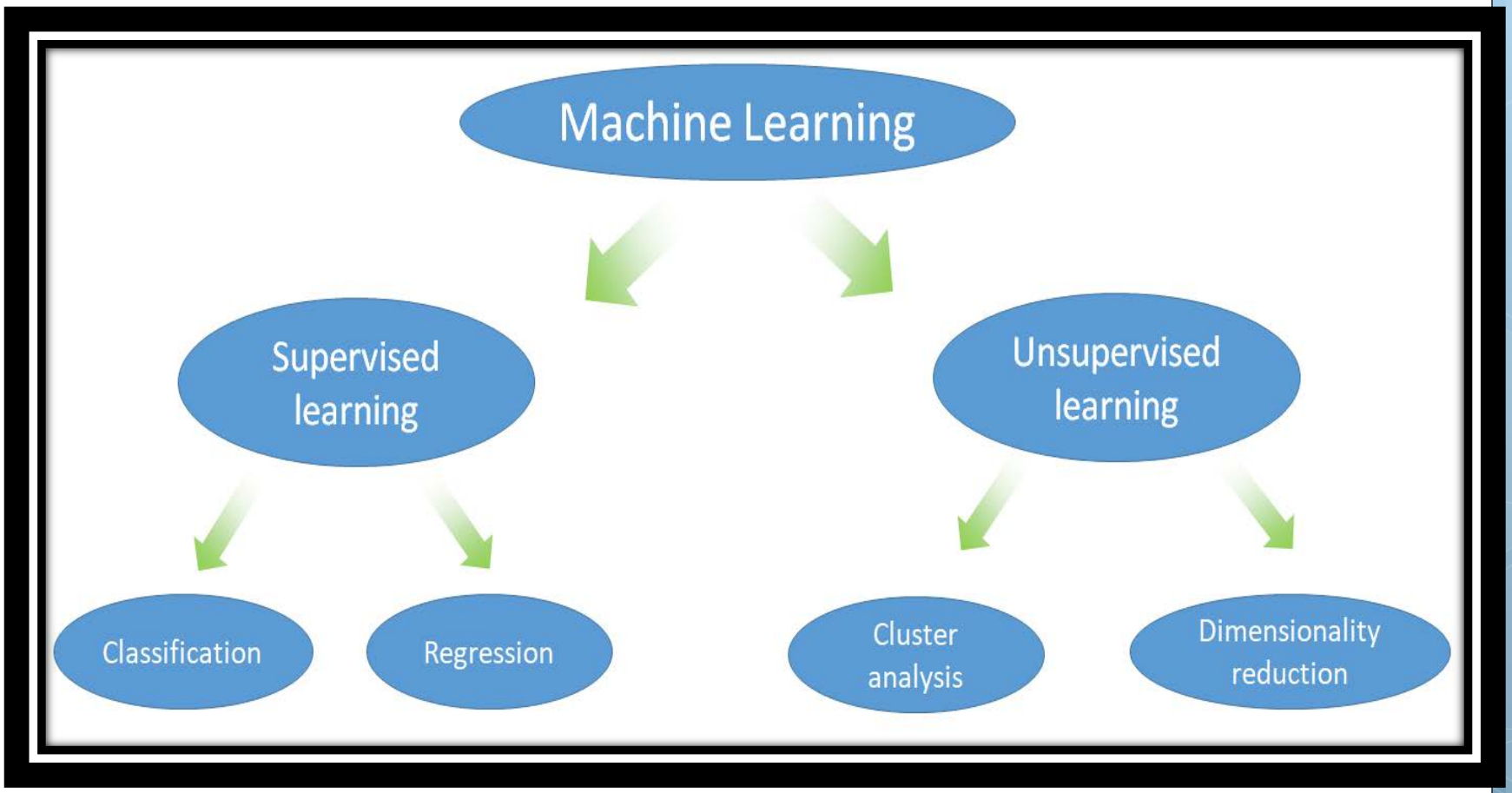
Dual Problem and Strong Duality Theorem

- Given the primal optimization problem, the dual problem of it was defined as

$$maximize \quad \theta(\alpha,\beta) = \inf_{w\in\Omega} L(w,\alpha,\beta)$$
$$subject \ \ to \qquad \alpha>0$$

- **Strong Duality Theorem**: Given the primal optimization problem, where the domain Ω is convex and the constraints g_i *and* h_i are affine functions. Then the optimum of the primal problem occurs at the same values as the optimum of the dual problem .

Machine Learning

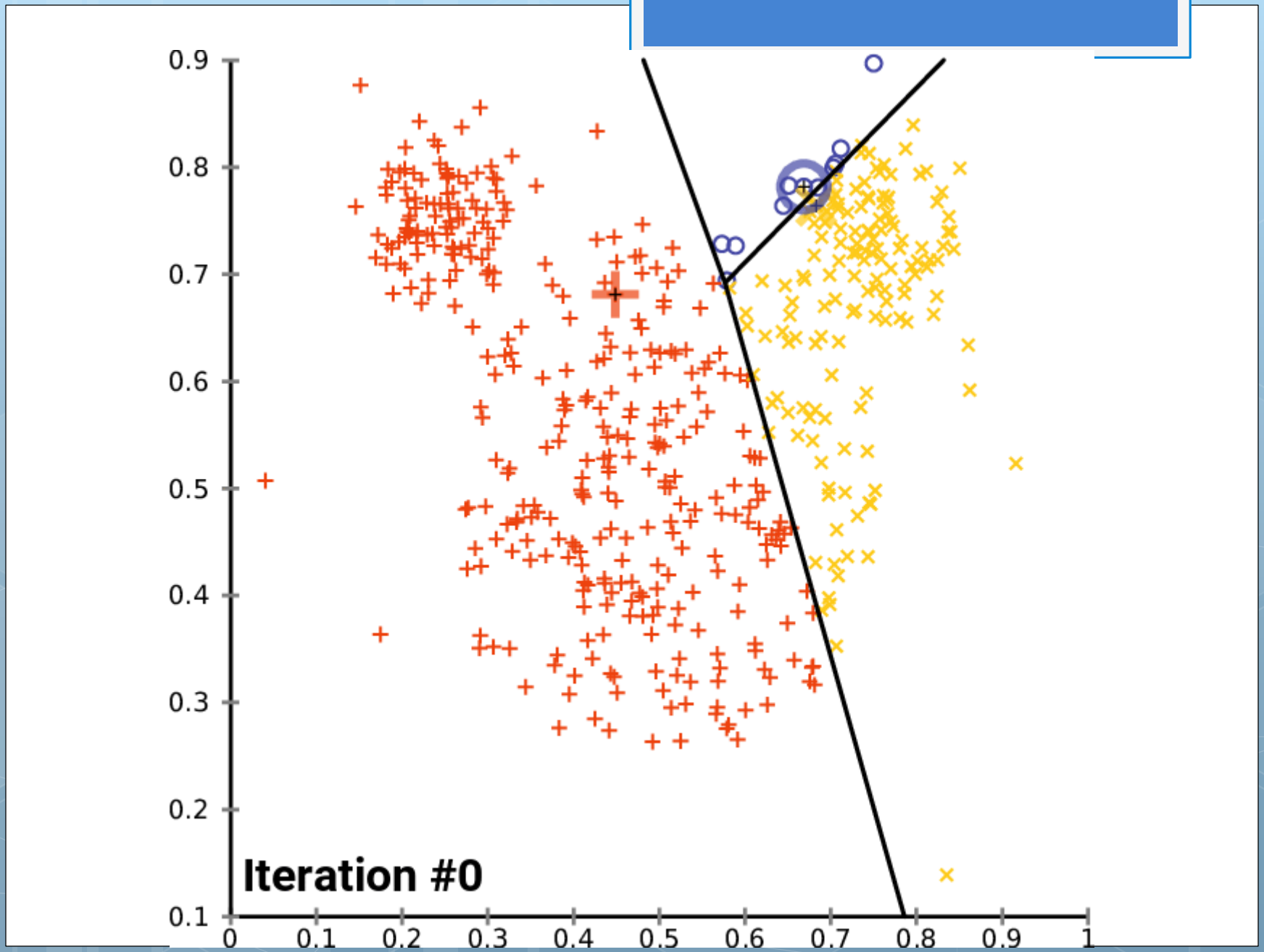

Machine Learning
Supervised learning
Unsupervised learning
Classification
Regression
Cluster analysis
Dimensionality reduction

Unsupervised Learning

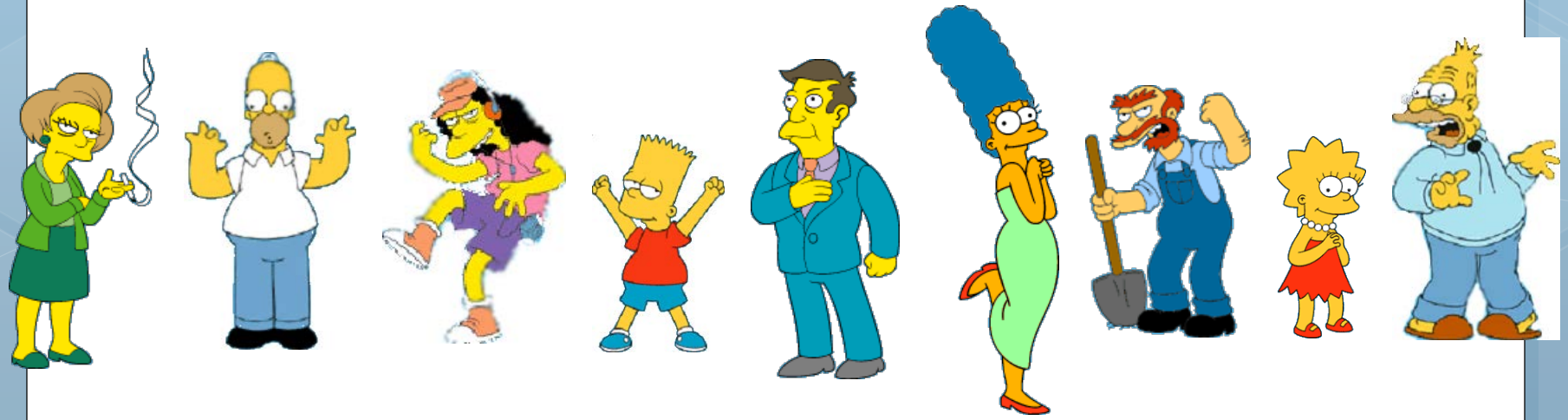
What is Clustering?

Also called *unsupervised learning*, sometimes called *classification* by statisticians and *sorting* by psychologists and *segmentation* by people in marketing

- Organizing data into classes such that there is
- high intra-class similarity
- low inter-class similarity
- Finding the class labels and the number of classes directly from the data (in contrast to classification).

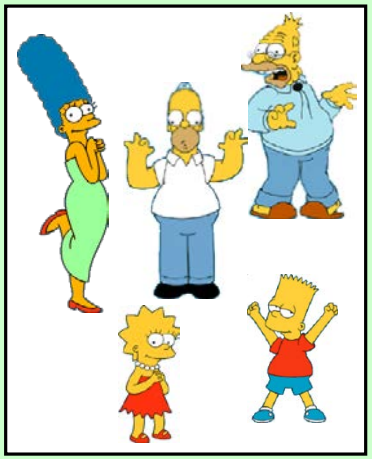

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Iteration #0

What is a natural grouping among these objects?



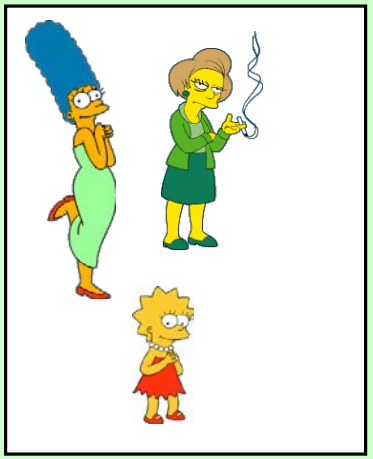
What is a natural grouping among these objects?

Clustering is subjective



Simpson's Family

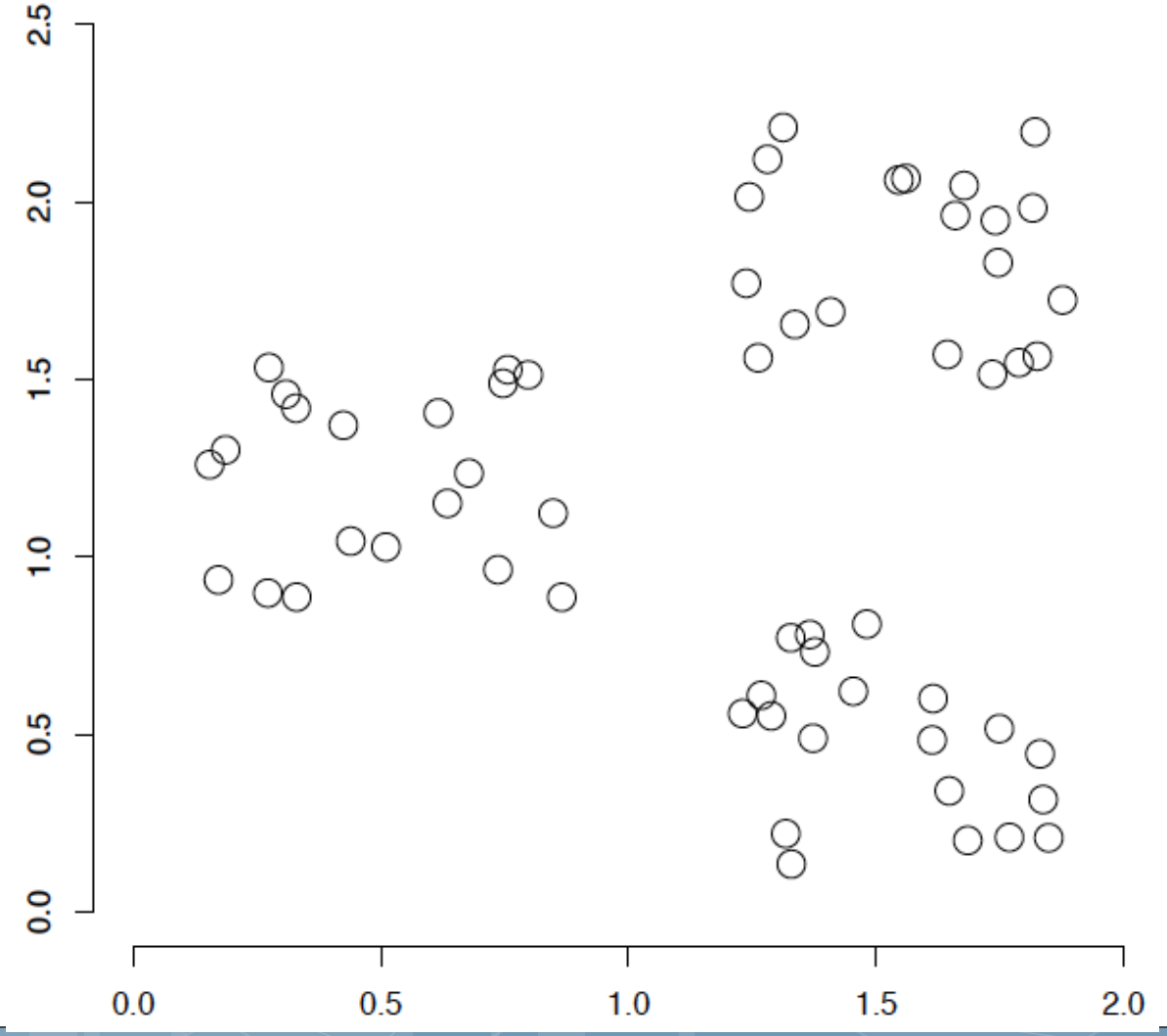
School Employees



Females

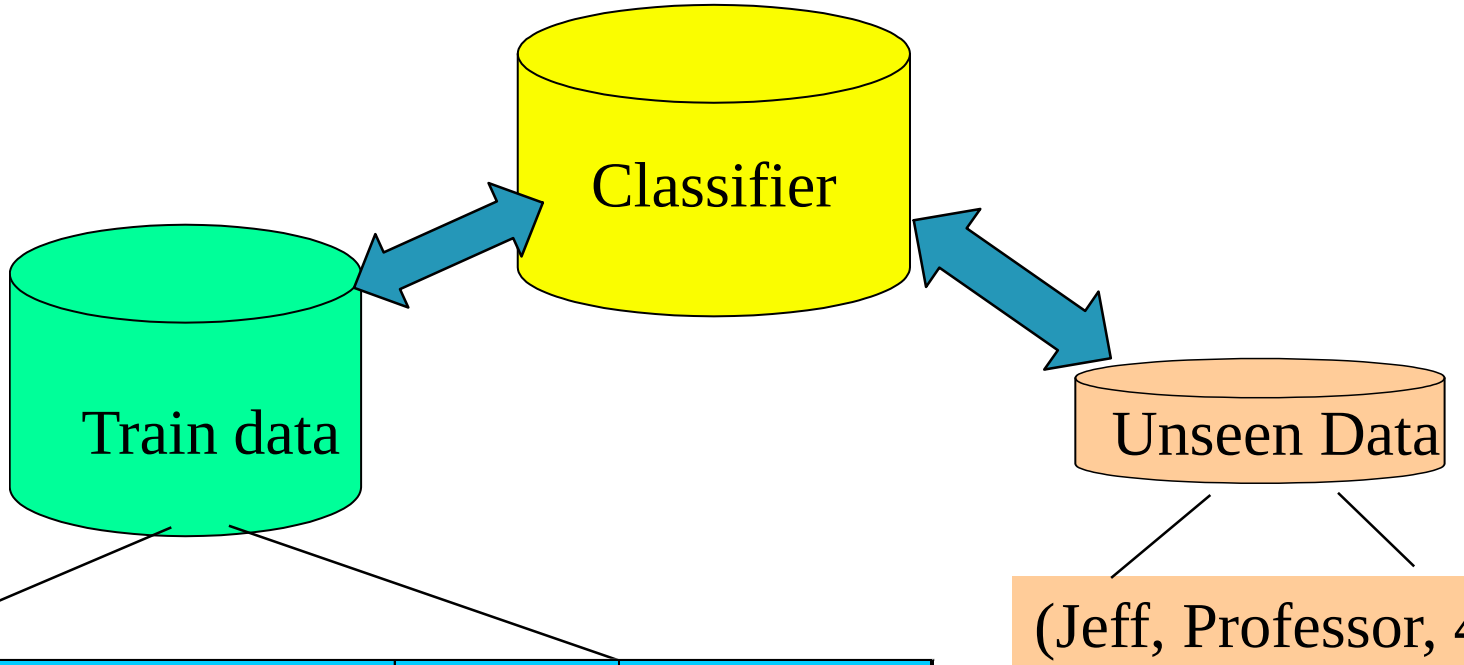
Males

A data set with clear cluster structure



- How would you design an algorithm for finding the three clusters in this case?

Supervised Learning

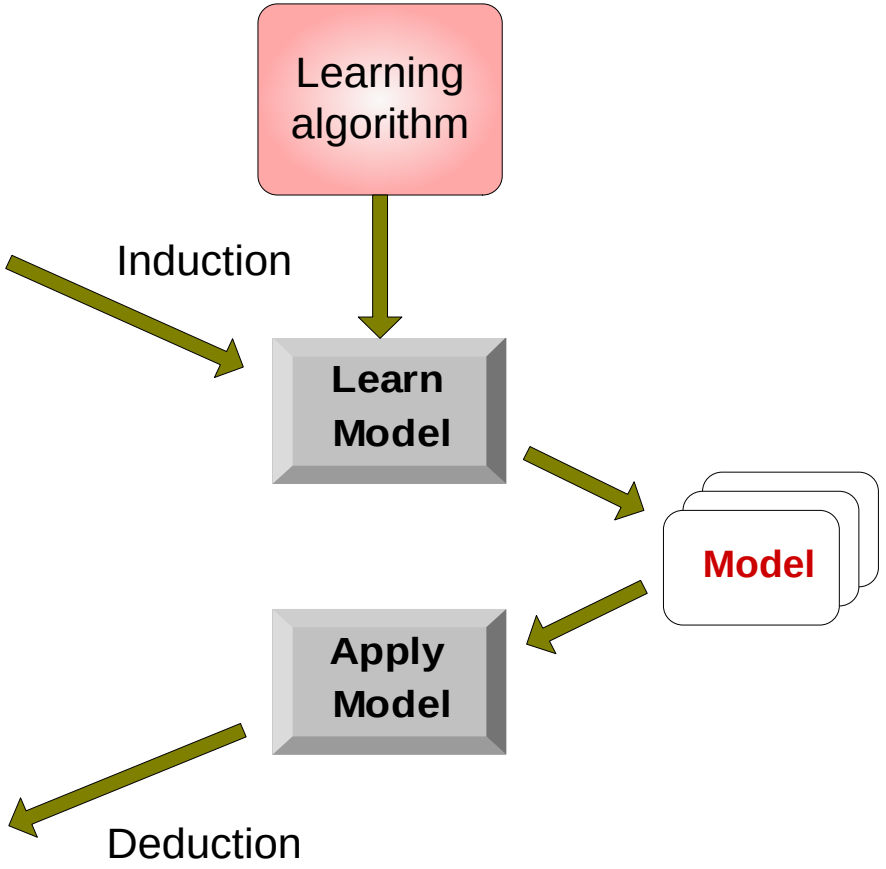


NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

General Approach for Building Classification Model

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set


EMAIL FILTER
inbox
Spam

Supervised vs. Unsupervised Learning

- Supervised learning (classification)
  - Supervision: The training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (clustering)
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Classification Techniques

- Base Classifiers
  - Decision Tree based Methods
  - Rule-based Methods
  - Nearest-neighbor
  - Neural Networks
  - Deep Learning
  - Naïve Bayes and Bayesian Belief Networks
  - Support Vector Machines
- Ensemble Classifiers
  - Boosting, Bagging, Random Forests

Support Vector Machine (SVM)

What is a good Decision Boundary?

- Consider a two-class, linearly separable classification problem. Construct the hyperplane

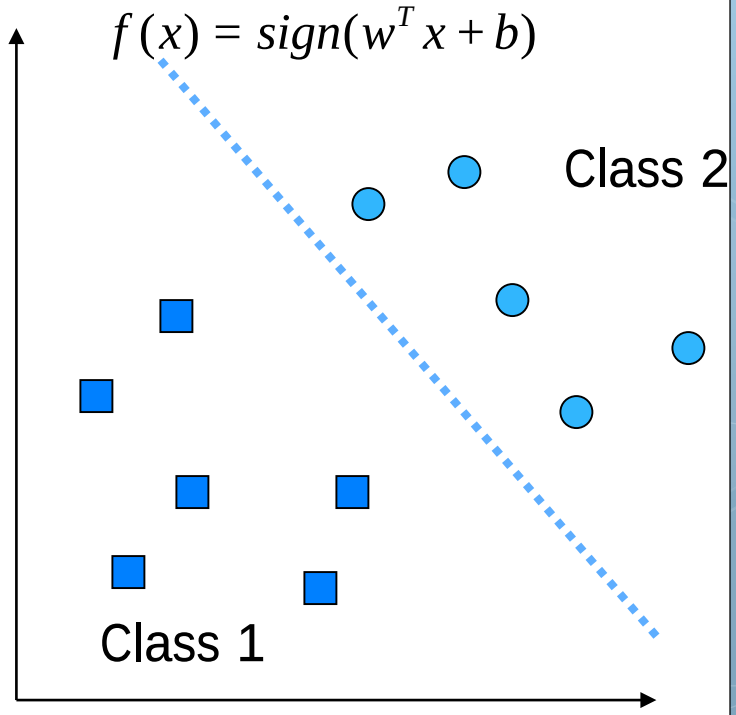
$$w^T x + b = 0, \qquad x \in R^n$$

to make

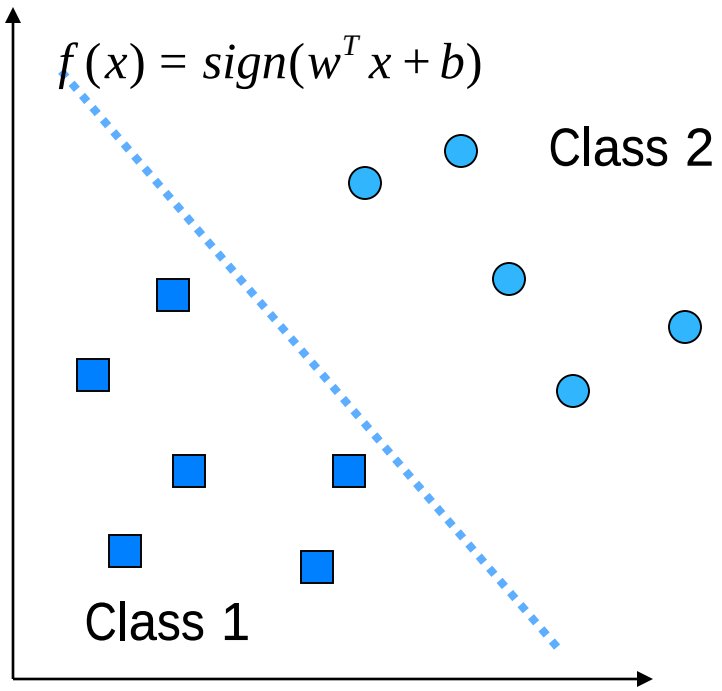
$$w^T x_i + b > 0, \qquad for \quad y_i = +1$$

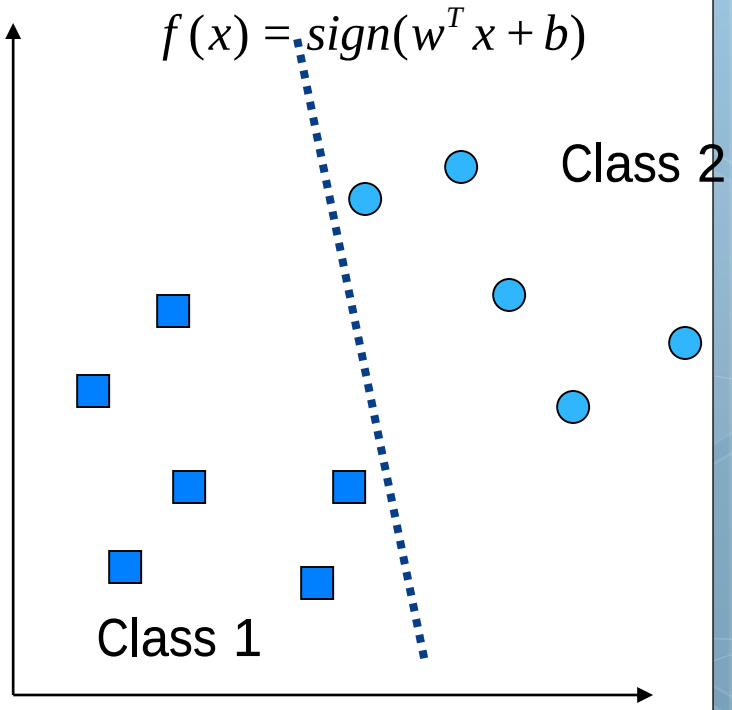
$$w^T x_i + b < 0, \qquad for \quad y_i = -1$$

- Many decision boundaries! Are all decision boundaries equally good?



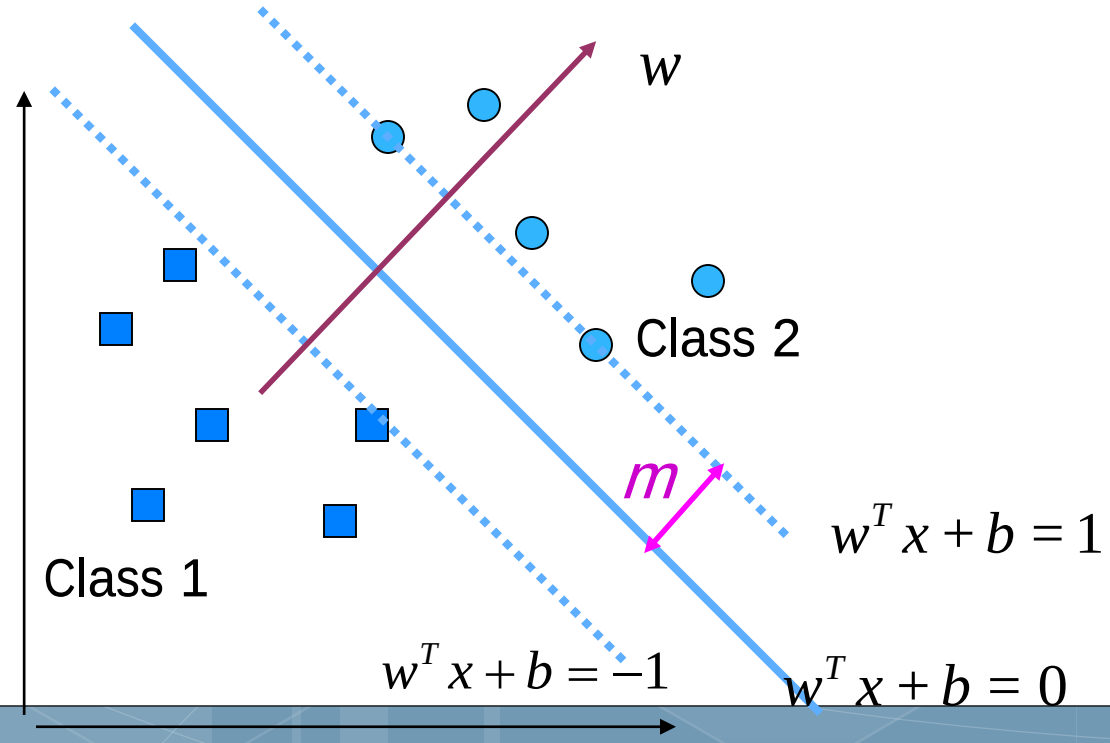
Examples of Bad Decision Boundaries





Optimal separating hyperplane

- The optimal separating hyperplane



- For the hyperplane, it can be proved that the margin *m* is

$$m = \frac{1}{\|w\|^2}$$

Hence, maximizing margin is equivalent to minimizing the square of the norm of w.

Finding the optimal decision boundary

- Let $\{x_1, ..., x_n\}$ be our data set and let $y_i \in \{1,-1\}$ be the class label of x_i
- The optimal decision boundary should classify all points correctly $\Rightarrow$ $y_i(w^T x_i + b) \geq 1, \quad \square i$
- The decision boundary can be found by solving the following constrained optimization problem

$$minimize \quad \frac{1}{2}\|w\|^2$$

$$subject\ to\ y_i(w^T x_i + b) \geq 1 \qquad \forall i$$

Lagrangian of the optimization problem

$$\text{minimize} \quad \frac{1}{2}\|w\|^2$$

$$\text{subject to } y_i(w^T x_i + b) \geq 1 \qquad \forall i$$

- The Lagrangian is

$$L = \frac{1}{2} w^T w + \sum_{i=1}^{n} \alpha_i (1 - y_i (w^T x_i + b))$$

- Setting the gradient of L w.r.t. w and be to zero, we have

$$w + \sum_{i=1}^{n} \alpha_i (-y_i) x_i = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

The Dual Problem

- If we substitute $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ into Lagrangian L, we have

$$L = \frac{1}{2}\sum_{i=1}^{n} \alpha_i y_i x_i^T \sum_{j=1}^{n} \alpha_j y_j x_j + \sum_{i=1}^{n} \alpha_i \left(1 - y_i (\sum_{j=1}^{n} \alpha_j y_j x_j^T x_i + b)\right)$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i y_i \sum_{j=1}^{n} \alpha_j y_j x_j^T x_i - b\sum_{i=1}^{n} \alpha_i y_i \quad (0)$$

$$= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{n} \alpha_i$$

- Note that $\sum_{i=1}^{n} \alpha_i y_i = 0$, and the data points appear in terms of their inner product; this is a quadratic function of α_i only.

The Dual Problem

- The dual problem is therefore:

$$\textit{maxmize} \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\textit{subject to} \quad \alpha_i \geq 0, \qquad \sum_{i=1}^{n} \alpha_i y_i = 0$$

The Dual Problem

$$\text{minimize} \quad W(\alpha) = \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{n} \alpha_i$$

$$\text{subject to} \quad \alpha_i \geq 0, \qquad \sum_{i=1}^{n} \alpha_i y_i = 0$$

- This is a quadratic programming (QP) problem, and therefore a global minimum of α_i can always be found

- w can be recovered by $w = \sum_{i=1}^{n} \alpha_i y_i x_i$, and

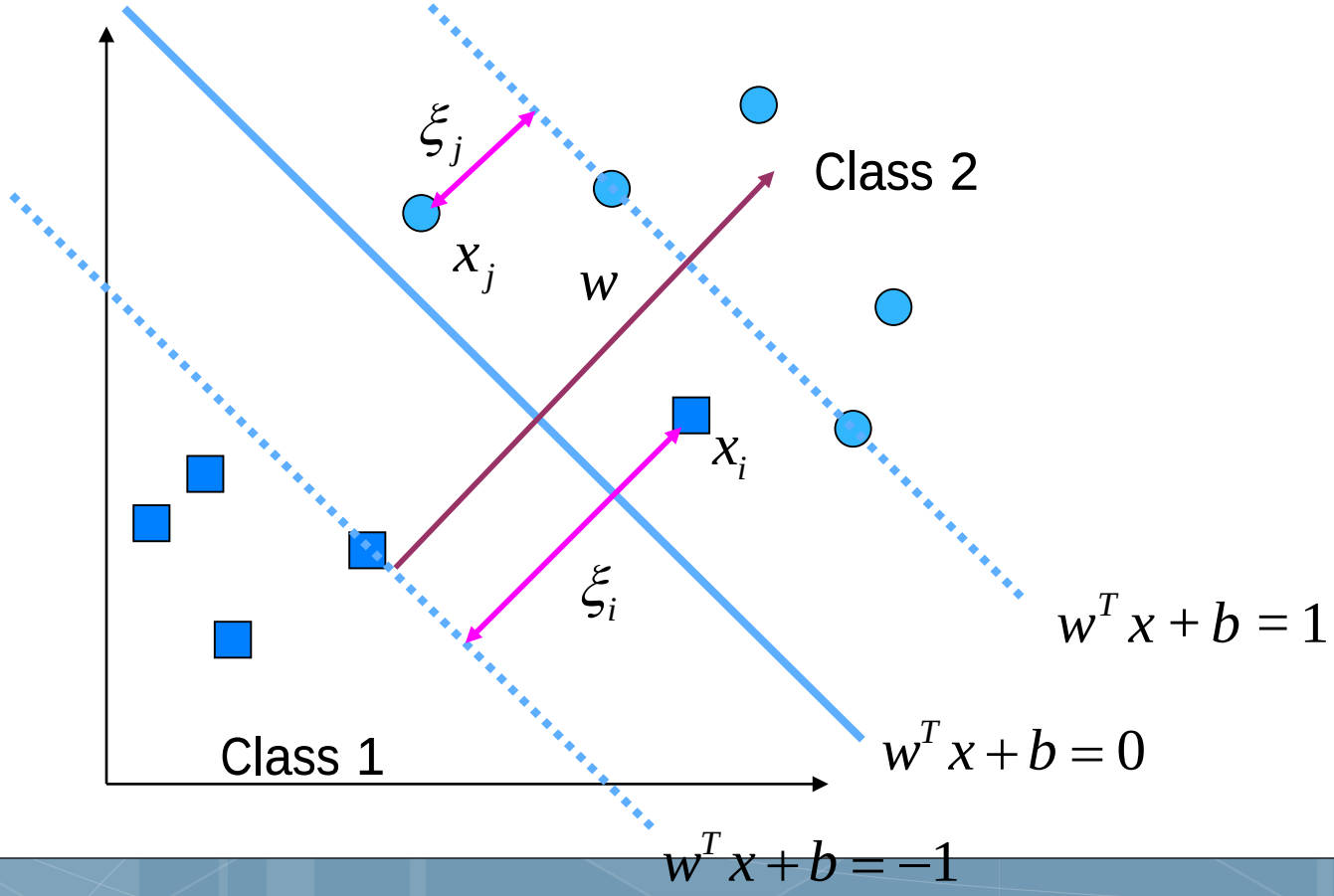
$$b = y_k - \sum_{i=1}^{n} \alpha_i y_i x_i^T x_k \qquad \text{for any} \quad \alpha_k > 0$$

- so the decision function can be written

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i x_i^T x + b\right)$$

The use of slack variables

- We allow "errors" ξ_i in classification for noisy data



Soft Margin Hyperplane

- The use of slack variables ξ_i enable the soft margin classifier

$$\begin{cases} w^T x_i + b \geq 1 - \xi_i & y_i = 1 \\ w^T x_i + b \leq -1 + \xi_i & y_i = -1 \\ \xi_i \geq 0 & \forall i \end{cases}$$

  - ξ_i are "slack variables" in optimization
  - Note that ξ_i=0 if there is no error for x_i

- The objective function $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$

  C : tradeoff parameter between error and margin

- The primal optimization problem becomes

$$\begin{aligned} &\textit{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\textit{subject to} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{aligned}$$

Dual Soft-Margin Optimization Problem

- The dual of this new constrained optimization problem is

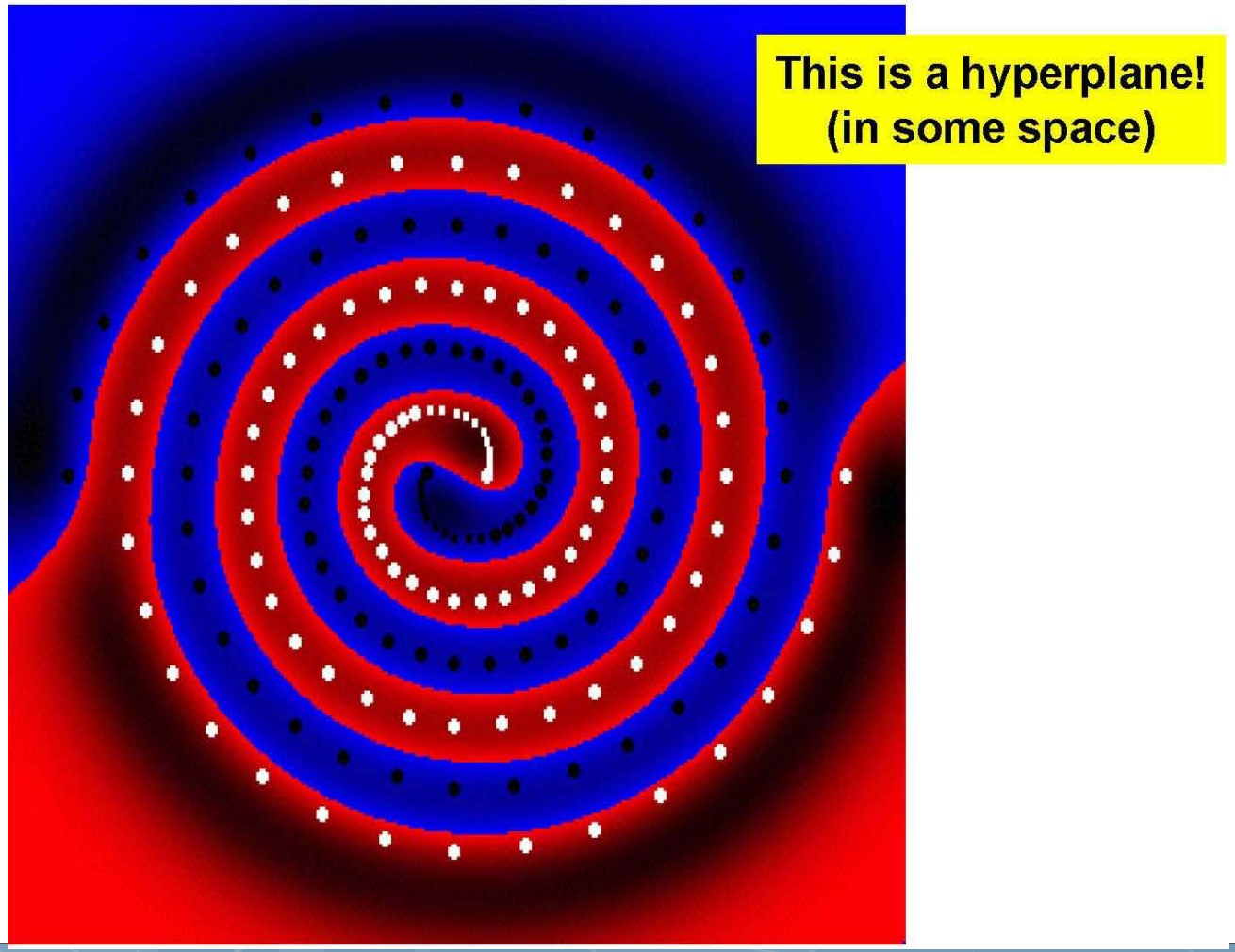
$$maxmize \;\; W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$subject \;\; to \;\; C \geq \alpha_i \geq 0, \;\; \sum_{i=1}^{n} \alpha_i y_i = 0$$

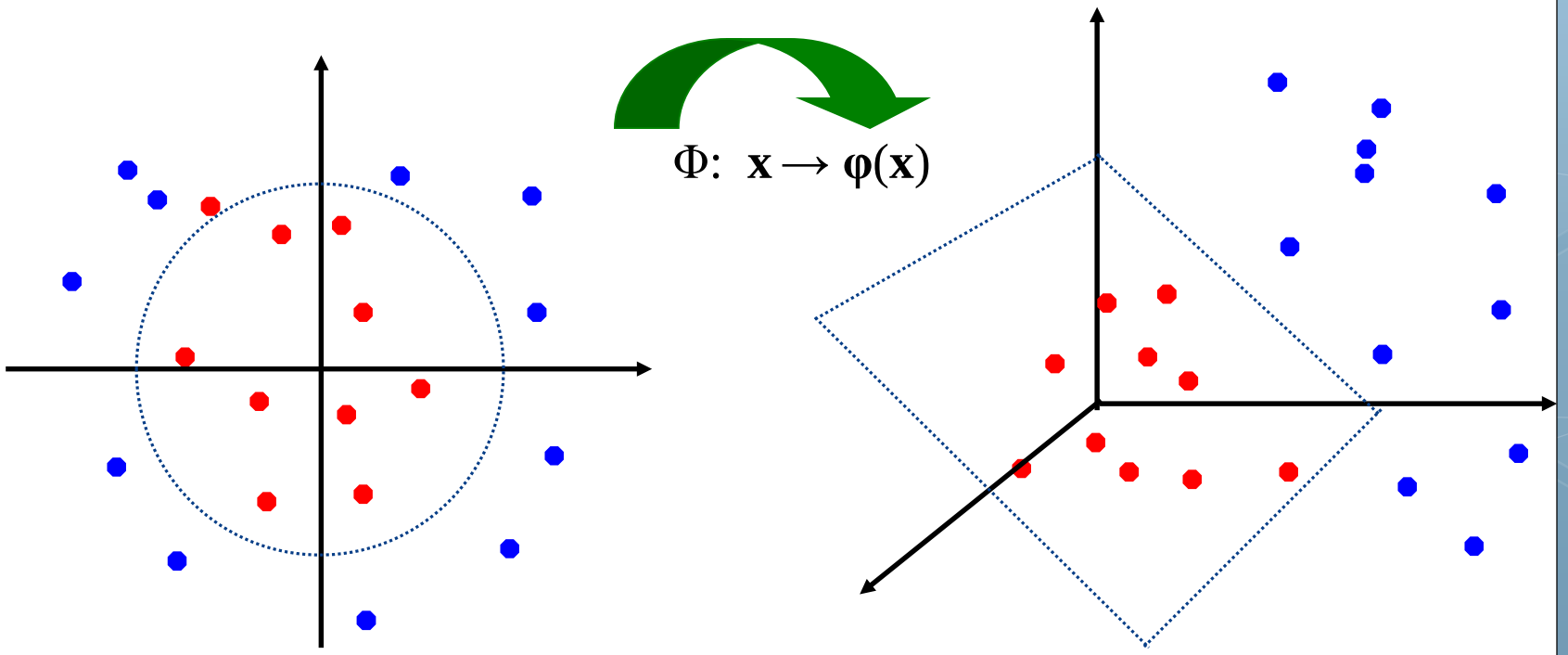
- w can be recovered as $w = \sum_{i=1}^{n} \alpha_i y_i x_i$

- This is very similar to the optimization problem in the hard-margin case, except that there is an upper bound C on α_i now.
- Once again, a QP solver can be used to find α_i

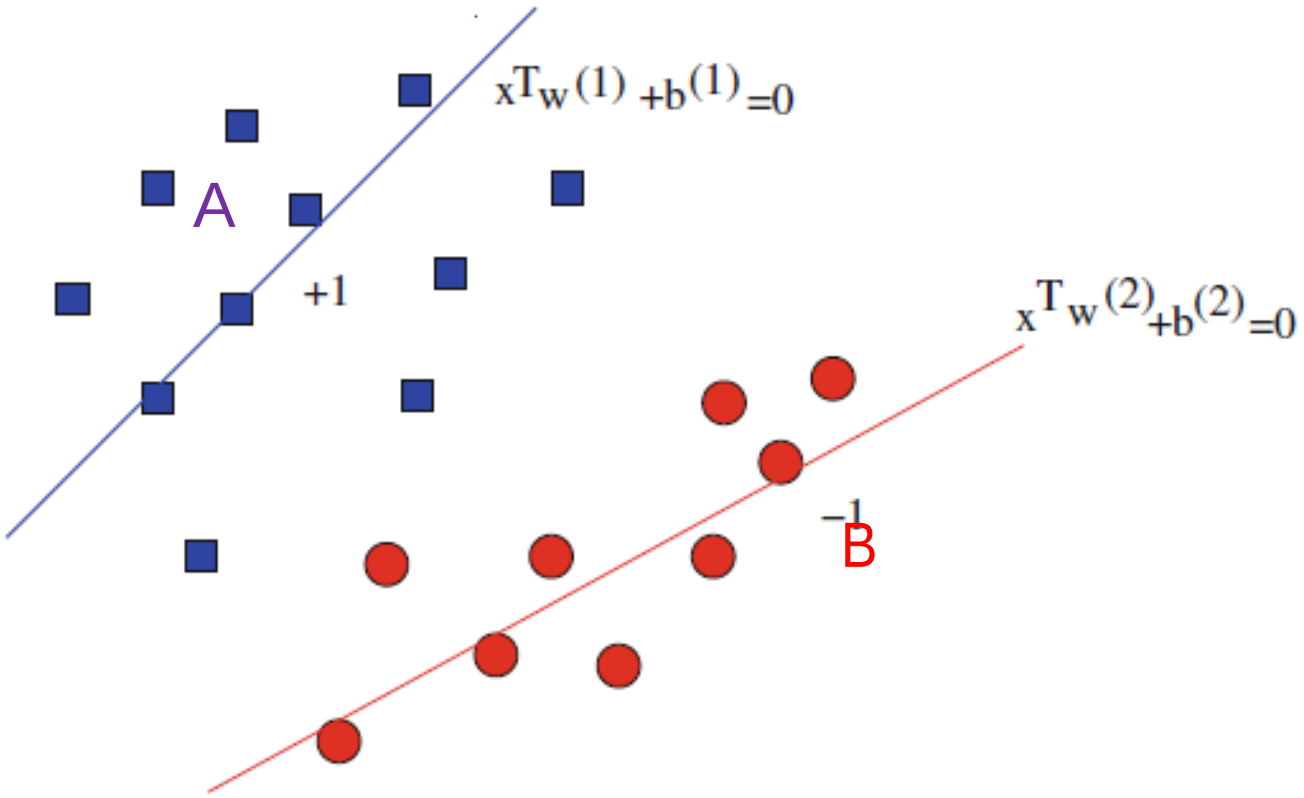
Nonlinear separable problems



Non-linear SVMs: Feature spaces



Proximal Support Vector Machine



The algorithm finds two non-parallel hyperplanes one for each class, each hyperplane should be as close as possible to one class and as far as possible from the other class.

$$\min \frac{\|AW^1 + b^1\|}{\|BW^1 + b^1\|}$$

$$\min \frac{\|BW^2 + b^2\|}{\|AW^2 + b^2\|}$$

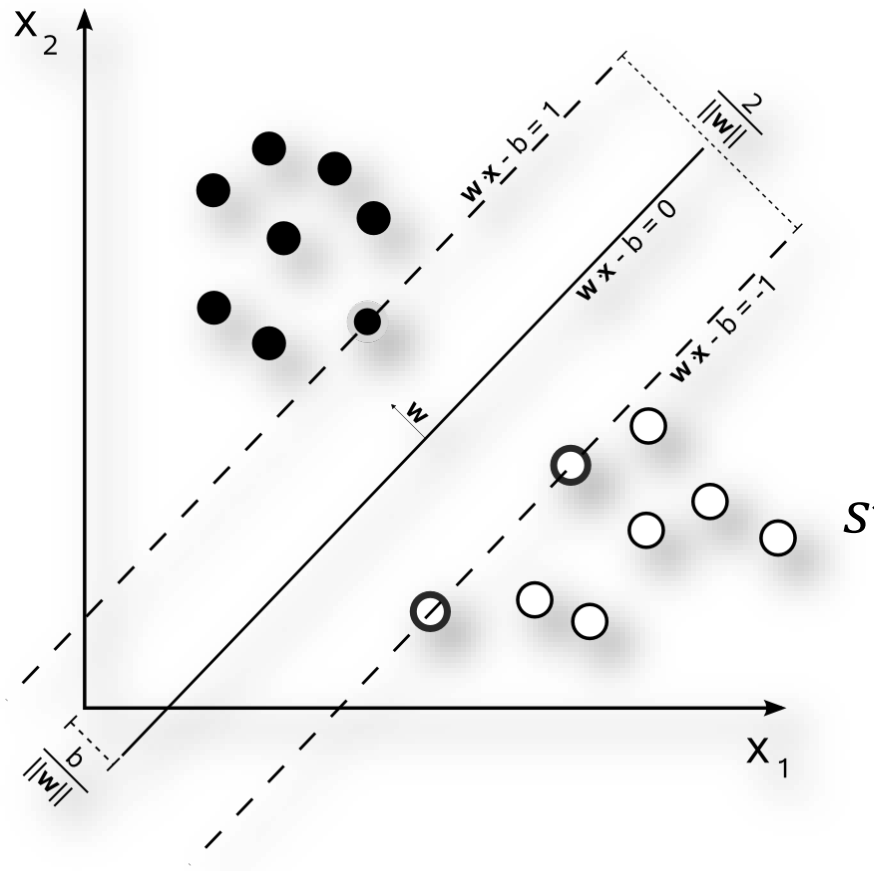
IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE,

Twin Support Vector Machines for Pattern Classification

Jayadeva, *Senior Member*, *IEEE*,
R. Khemchandani, *Student Member*, *IEEE*,
and
Suresh Chandra

Abstract—We propose Twin SVM, a binary SVM classifier that determines two nonparallel planes by solving two related SVM-type problems, each of which is smaller than in a conventional SVM. The Twin SVM formulation is in the spirit of proximal SVMs via generalized eigenvalues. On several benchmark data sets, Twin SVM is not only fast, but shows good generalization. Twin SVM is also useful for automatically discovering two-dimensional projections of the data.

Standard SVM :



$$\min_{w,r} \frac{1}{2} w^T w + \nu e^T r,$$

$$subject\ to\ \ (Aw - e\gamma) + r \geq e,$$

$$(Bw - e\gamma) - r \leq -e,$$

$$r \geq 0.$$

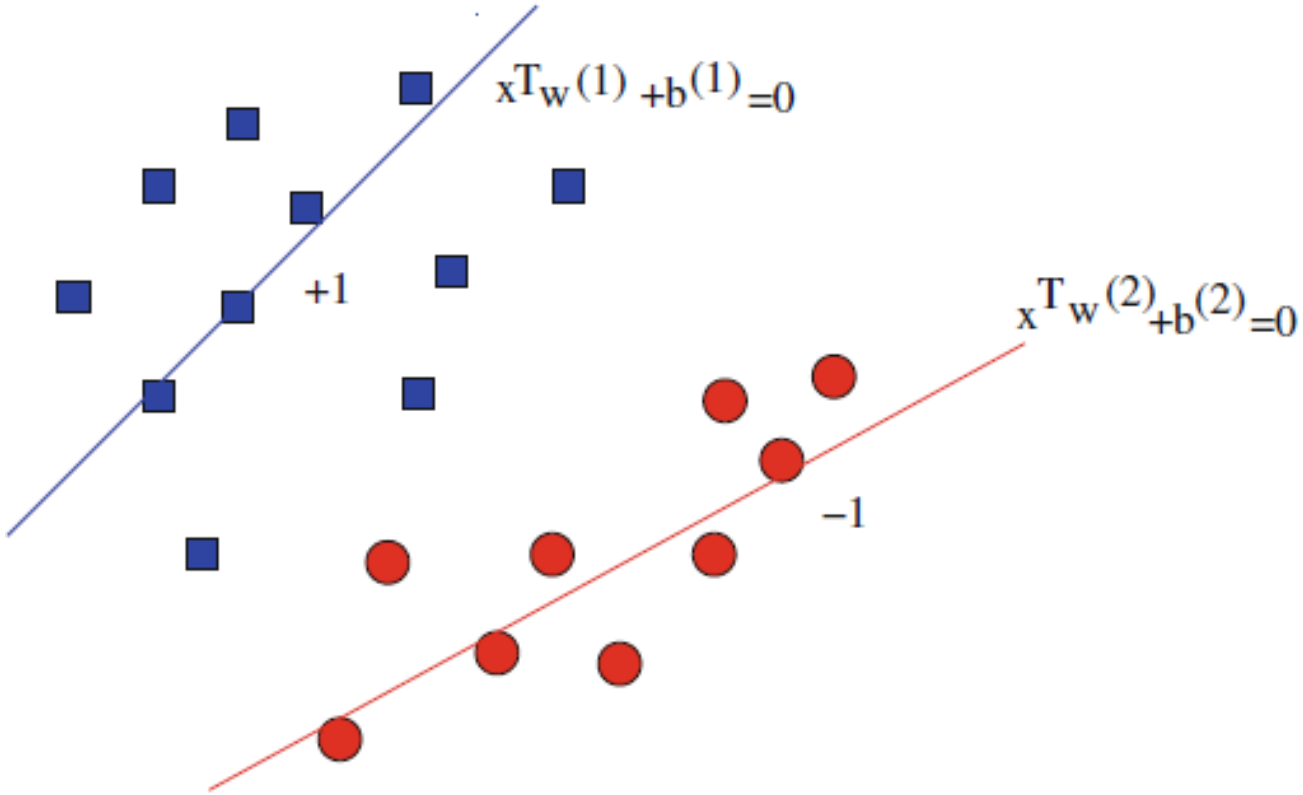
Why TWSVM?

This quadratic programming problem (QPP) is expensive to solve for large dimensions because all data points appear in the constraints.

How does it works ?

Instead of solving one large QPP, TWSVM solve two smaller QPP each of them has the formulation of standard SVM except that not all data patterns appear in the constraint at the same time.

The algorithm finds two non-parallel hyperplanes one for each class, each hyperplane should be as close as possible to one class and as far as possible from the other class.


$x^T w^{(1)} + b^{(1)} = 0$
+1
$x^T w^{(2)} + b^{(2)} = 0$
−1

Linear Classifier

TWSVM is obtained by solving the following pair of QPPs:

$$(TWSVM1) \quad \underset{w^{(1)},\, b^{(1)},\, q}{Min} \quad \frac{1}{2}(Aw^{(1)} + e_1 b^{(1)})^T (Aw^{(1)} + e_1 b^{(1)}) + c_1 e_2^T q$$

$$subject\ to \quad -(Bw^{(1)} + e_2 b^{(1)}) + q \geq e_2, \quad q \geq 0,$$

$$(TWSVM2) \quad \underset{w^{(2)},\, b^{(2)},\, q}{Min} \quad \frac{1}{2}(Bw^{(2)} + e_2 b^{(2)})^T (Bw^{(2)} + e_2 b^{(2)}) + c_2 e_1^T q$$

$$subject\ to \quad (Aw^{(2)} + e_1 b^{(2)}) + q \geq e_1, \quad q \geq 0,$$

The first term of the objective function represents the sum of square distance from the hyperplane to each pattern of one class, therefore minimizing it keeps the hyperplane close to the patterns of one class.

The constraints require the hyper plane to be far from the other class patterns at least with distance 1.

The second term of the objective function minimize the sum of error variables to minimize miss classification of patterns belongs to other class.

The Wolfe dual can be obtain as follows

$$\max_{\alpha} e_2^T \alpha - \frac{1}{2}\alpha^T G(^{H^T H)-1} G^T \alpha, \qquad G = [B \quad e_2] \quad and \quad H = [A \quad e_1]$$
$$subject\ to\ \ 0 \le \alpha \le c_1$$

$$u = -(H^T H)^{-1} G^T \alpha \quad where \quad u = [w_1^T, \qquad b_1]^T.$$

$$\max_{\alpha} e_1^T \gamma - \frac{1}{2}\gamma^T P(^{Q^T Q)-1} P^T \gamma, \qquad \text{P=[A} \quad e\_1] \quad and \quad Q=[B \quad e\_2]$$
$$subject\ to\ \ 0 \le \gamma \le c_2$$

$$v = (Q^T Q)^{-1} P^T \gamma \quad where \quad v = [w_2^T, \qquad b_2]^T$$

The first QPP TWSVM can be modified as follow:

$$\min_{w_1,b_1,q_1} \|Aw_1 + e_1 b_1\|^2 + c_1 q^T q,$$
$$subject\ to \quad -(Bw_1 + e_2 b_1) + q \geq e_2,$$
$$q \geq 0.$$

We combine constraint together, then we have

$$q = (e_2 - Bw_1 - e_2 b_1)_+$$

Then the above problem change to following unconstrained problem:

$$\min_{w_1,b_1,q_1} \|Aw_1 + e_1 b_1\|^2 + c_1 \|(e_2 - Bw_1 - e_2 b_1)_+\|^2.$$

Similarly, the second QPP TWSVM can be modified as follow:

$$\min_{w_2,b_2,q_2} \|Bw_2 + e_2 b_2\|^2 + c_2 \|(e_1 - Aw_2 - e_1 b_2)_+\|^2.$$

The above problems are piecewise, quadratic, convex, and once differentiable. The generalized Newton method can be used for solving them.

Algorithm : Generalized Newton Method with the Stepsize Armijo Rule

Choose any vector p_0 and $\epsilon > 0$, $i = 0$;

while $\|\nabla g(p_i)\|_\infty \geq \epsilon$

Choose $\alpha_i = max\{s, s\delta, s\delta^2, ...\}$ *such that*

$$g(p_i) - g(p_i + \alpha_i d_i) \geq -\delta\mu\nabla g(p_i)^T d_i,$$

where $d_i = -\partial^2 g(p_i)^{-1}\nabla g(p_i)$, $s > 0$ is a constant, $\delta \in (0, 1)$ and $\mu \in (0, 1)$.

$$p_{i+1} = p_i + \alpha_i d_i$$

$$i = i + 1;$$

Numerical Experiments

Data set	Twin SVM	New Method	
ionosphere	0.8346+-0.0617	**0.92024e-001+-3.9924e-002**	
WPBC	0.6511+-0.2512	**0.8792+-0.0757**	
WDBC	0.5778+-0.1128	**0.9526+-0.0468**	
Pima Indians	0.36309+-4.3776e-002	**0.69672+- 7.4829e-002**	
Soanr	0.61524+-7.1800e-002	**0.85024+-6.0008e-002**	
Heart-statlog	.57407+-8.4186e-002	**0.67778+-6.0607e-002**	

We can extend this method to Nonlinear Classifier

One of the hardest parts of writing a research paper can be just finding a good topic to write about.

Some ideas:

1. Finding a new method to separate data sets
2. New efficient optimization model for the previous ideas.
3. Solving the existence ideas with a new method.
4. Extending the currents methods for binary classification to

Multi-class classification

Resources: Datasets

- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- UCI KDD Archive: http://kdd.ics.uci.edu/summary.data.application.html
- Statlib: http://lib.stat.cmu.edu/
- Delve: http://www.cs.utoronto.ca/~delve/

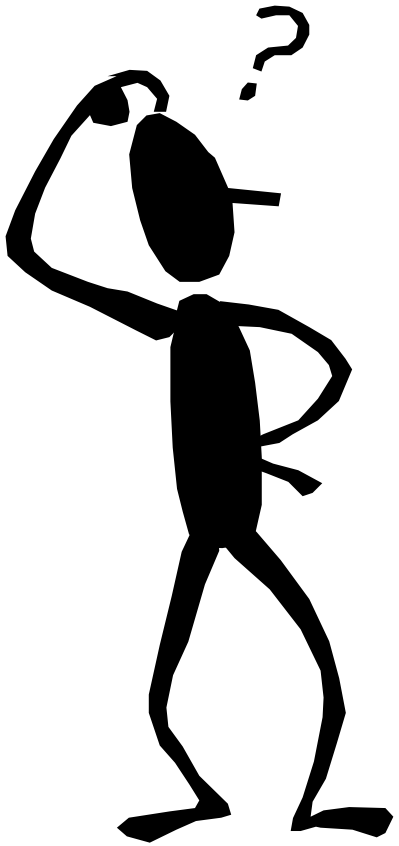
Resources: Journals

- Journal of Machine Learning Research Machine Learning
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Annals of Statistics
- Journal of the American Statistical Association
- ...

Resources: Conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NIPS)
- Computational Learning
- International Joint Conference on Artificial Intelligence (IJCAI)
- ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)
- IEEE Int. Conf. on Data Mining (ICDM)

Questions?



The End ...

Thanks for your Attention

Charles Bridge Prague