ML – Supervised Learning (SVM)

Fakultas Teknologi Industri UII
14 Februari 2020

Septia Rani, S.T., M.Cs.

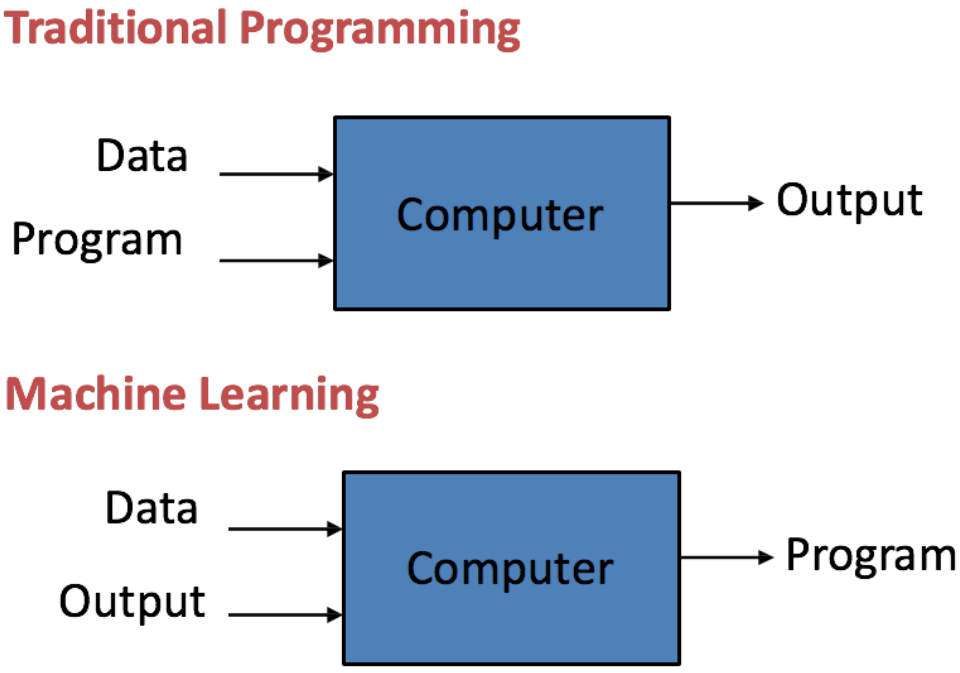
UNIVERSITAS ISLAM INDONESIA

VALUES | INNOVATION | PERFECTION

- Member of Center of Data Science, UII
- Lecturer at Department of Informatics, UII
- Education
  - Institut Teknologi Telkom (S1)
  - Universitas Gadjah Mada (S2)
  - Politecnico di Torino, Italy (S2 – exchange program)
- Research Interest
  - Artificial intelligence, machine learning
- septia.rani@uii.ac.id

Introduction to Machine Learning

- We would like machines to *learn* from data, instead of being explicitly programmed.



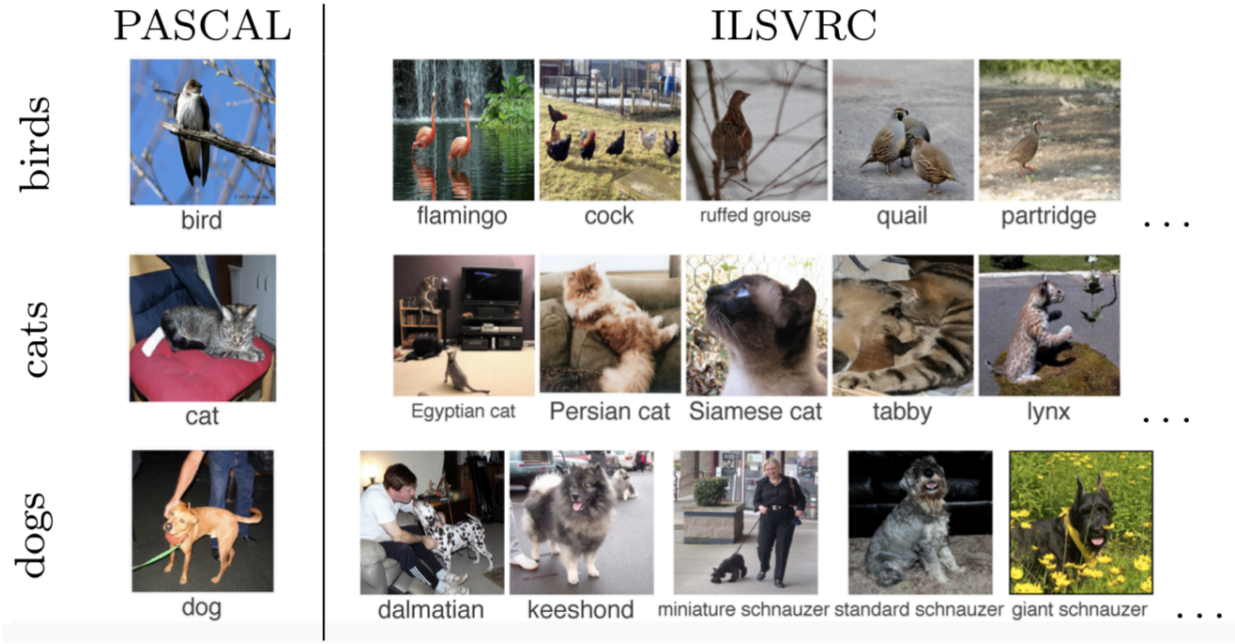
Common Learning Problems

- Common learning problems include:
  - Supervised learning
  - Unsupervised learning
  - Reinforcement learning

Supervised Learning

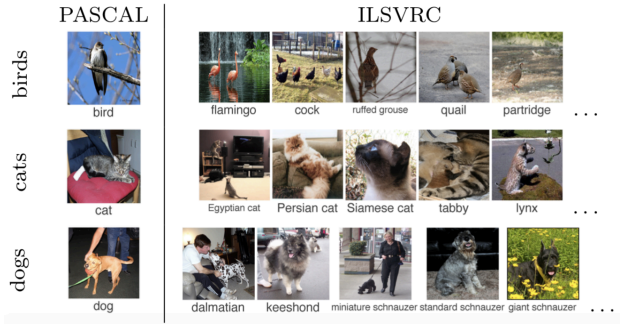
- Given a training set $S = ((x_1, y_1), \ldots, (x_m, y_m))$ drawn from $X \times Y$, the learning algorithm outputs a predictor $h : X \rightarrow Y$ that gives accurate prediction of y given x.
  - When y is categorical, we are doing **classification** and h is often called a classifier.

Object Recognition [2]



- In object recognition, we want a classifier that takes in an image and outputs the class of the object shown in the image.

Object Recognition [2]



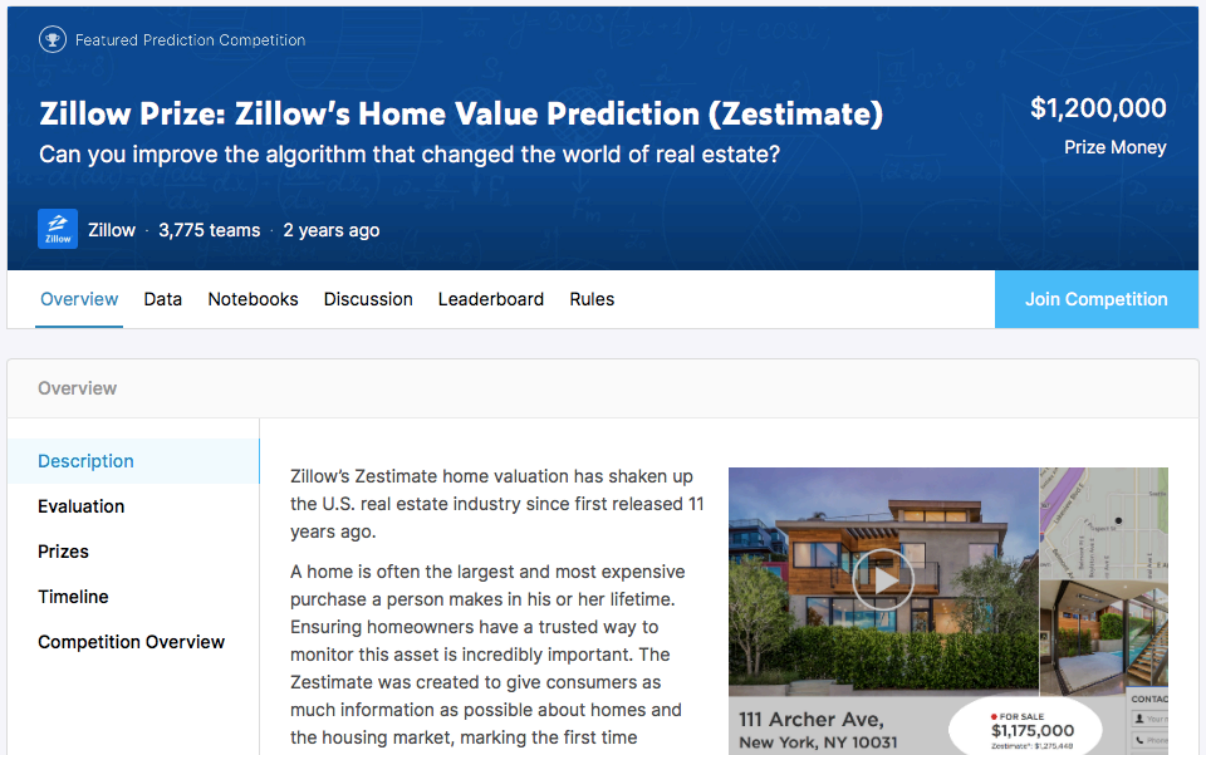
- The classifier is often learned using supervised learning.
  - Deep convolutional neural networks has been very successful.
  - ImageNet competition: 1000 classes, more than 1 million training images
  - 2010 to 2015 error rates: 28.2, 25.8, 16.4, 11.7, 6.7, 3.6
  - Around human level performance on ImageNet now.

Spam Filtering

- In spam filtering, we want a classifier that takes in an email and output whether it is spam or ham (non-spam).
  - Often created by learning.
  - Widely used on most people's email account.

  - Instead of using only spam or ham as output, we can output a real value representing the probability of spam.
  - The output can be thresholded using different thresholds to minimize false positive.
  - Problems requiring real-valued outputs are often referred to as **regression**, solved e.g. using logistic regression.

Home Price Prediction [3]

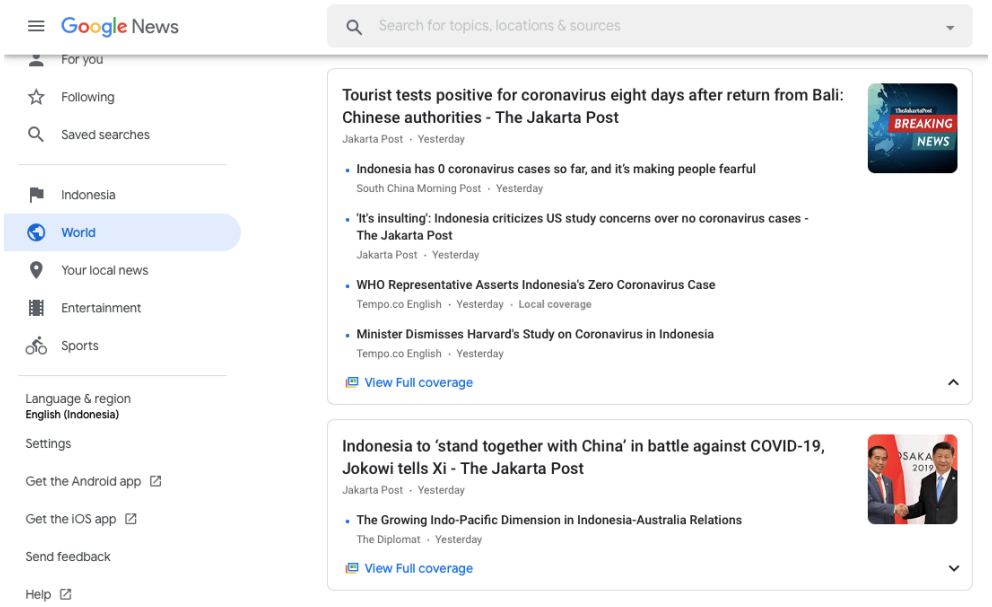


- Predict home price given location, size, number of rooms, etc.
- This is a regression problem.

Unsupervised Learning

- Given a training set $S = (x_1, \ldots, x_m)$, without a labeled output, construct a "good" model/ description of the data.
  - Look at the model/description, and find "interesting structure", e.g. clustering.
  - Can be used for dimension reduction to find the essential parts of the data and remove noise, e.g. PCA.
  - Unsupervised learning often minimizes description length of data: useful for efficient data transmission/storage.
  - Model of the data can also be used scoring how likely the data is, and for generating similar data.

Organizing News



- Google News groups all the articles about the same topic together into clusters to organize the articles for the readers.
  - Articles within the same cluster are similar to each other.
  - Articles in different clusters are different compared to articles from the same clusters.
- **Clustering** is often used to organize data into groups that are (hopefully) meaningful to users.

Reinforcement Learning (RL)

- RL is a very general framework for learning sequential decision making tasks.
- RL is a technique to allow an agent to take actions and interact with an environment so as to maximize the total rewards.

- RL is usually modeled as a Markov Decision Process (MDP) [4].

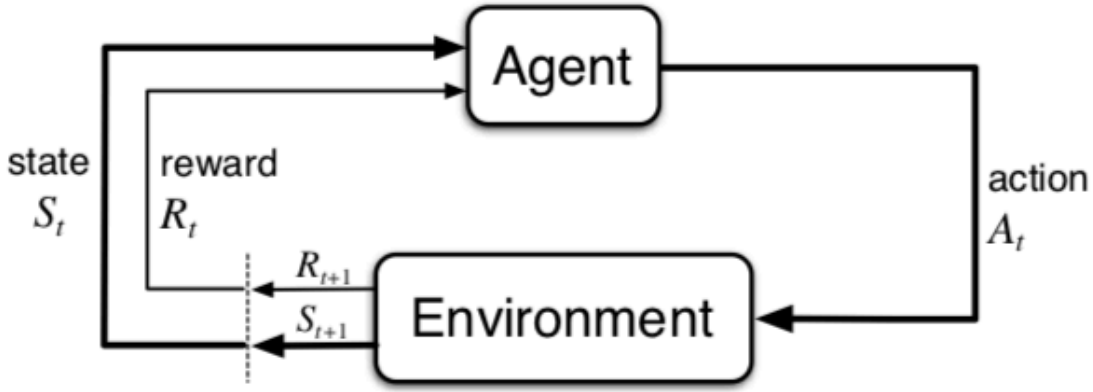


Figure 3.1: The agent–environment interaction in a Markov decision process.

RL Example [5]

- Imagine a baby is given a TV remote control at your home (environment).
  - In simple terms, the baby (agent) will first observe and construct his/her own representation of the environment (state).
  - Then the curious baby will take certain actions like hitting the remote control (action) and observe how would the TV response (next state).
  - As a non-responding TV is dull, the baby dislike it (receiving a negative reward) and will take less actions that will lead to such a result (updating the policy) and vice versa.
  - The baby will repeat the process until he/she finds a policy (what to do under different circumstances) that he/she is happy with (maximizing the total (discounted) rewards).
- How to construct a mathematical framework to solve this problem?

RL Applications

Autonomous Helicoptor Flight

AlphaGo

Semi-Supervised Learning (additional)

- In semi-supervised learning (SSL), an algorithm learns from a dataset that includes both labeled and unlabeled data, usually mostly unlabeled.
- Not enough labeled data to produce an accurate model and not have the ability or resources to get more data, then increase the size of your training data using SSL.

Detect Fraud For A Large Bank [6]

Name	Loan Amount	Loan Repaid	Fraud
Ashley	100000	1	1
Chuck	25000	0	0
Tim	4000	1	1
Mike	150000	1	1
Colin	200000000	0	
Libby	400400	1	0
Sheila	3200	1	1
Mandi	34850	1	
Gareth	6570	0	0

Name	Loan Amount	Loan Repaid	Fraud
Ashley	100000	1	1
Chuck	25000	0	0
Tim	4000	1	1
Mike	150000	1	1
Colin	200000000	0	0
Libby	400400	1	0
Sheila	3200	1	1
Mandi	34850	1	1
Gareth	6570	0	0

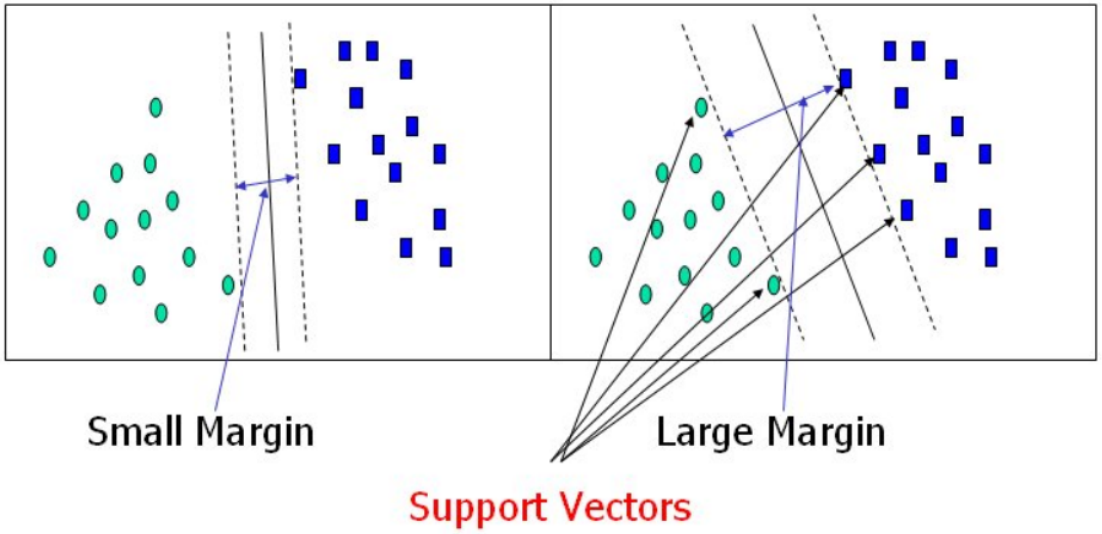
- You can use a semi-supervised learning algorithm to label the data, and retrain the model with the newly labeled dataset.
- However, there is no way to verify that the algorithm has produced labels that are 100% accurate, resulting in less trustworthy outcomes than traditional supervised techniques.

Support Vector Machine (SVM)

- SVMs can actually be used for both classification or regression problems, but most of the time you will find them used for classification.
- An SVM is an example of what's called a *constrained optimisation* problem.

Constrained Optimization Problem in SVM

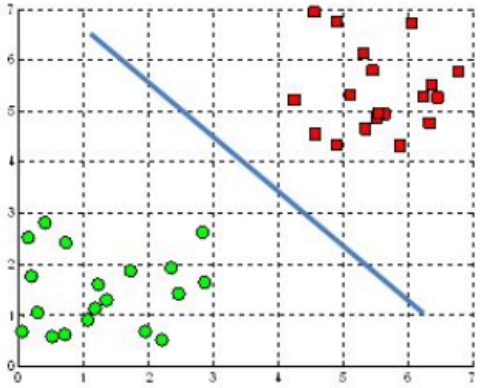
- It *optimises* the space of the margin between classes, but also seeks to *constrain* the extent of this margin in relation to data points.



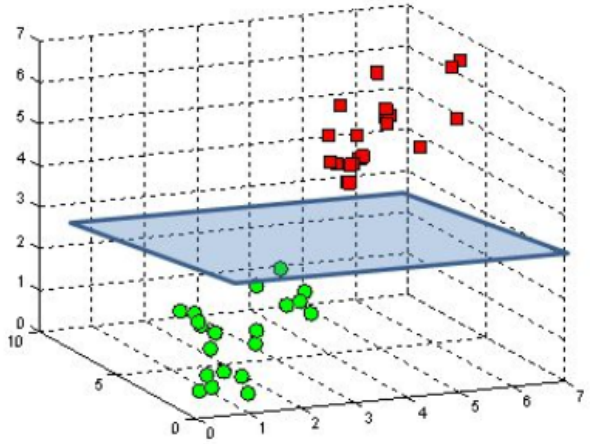
- You can think of the problem in terms of how you can construct a street between the points, that is as wide as possible.
- SVM's strategy: find the best hyperplane!

Hyperplane

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



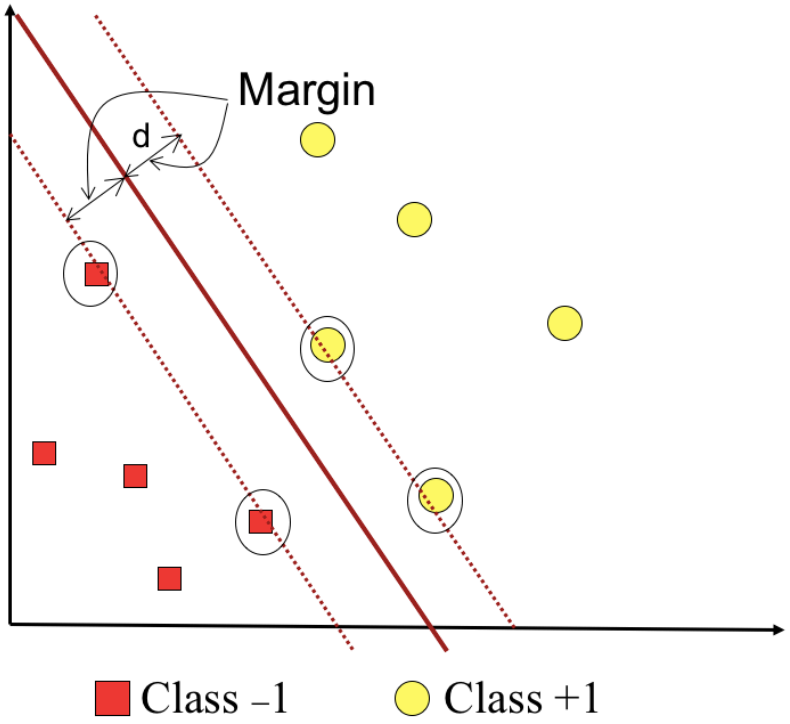
A hyperplane in $\mathbb{R}^3$ is a plane



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

- An hyperplane is a $n-1$ subspace in a n dimensional vector space.
  - For 2D spaces, the set of points in a line define an hyperplane.
  - For 3D, the points in a 2D surface (plane) define an hyperplane.

Margin



- Margin (d) = minimum distance between hyperplane and training samples

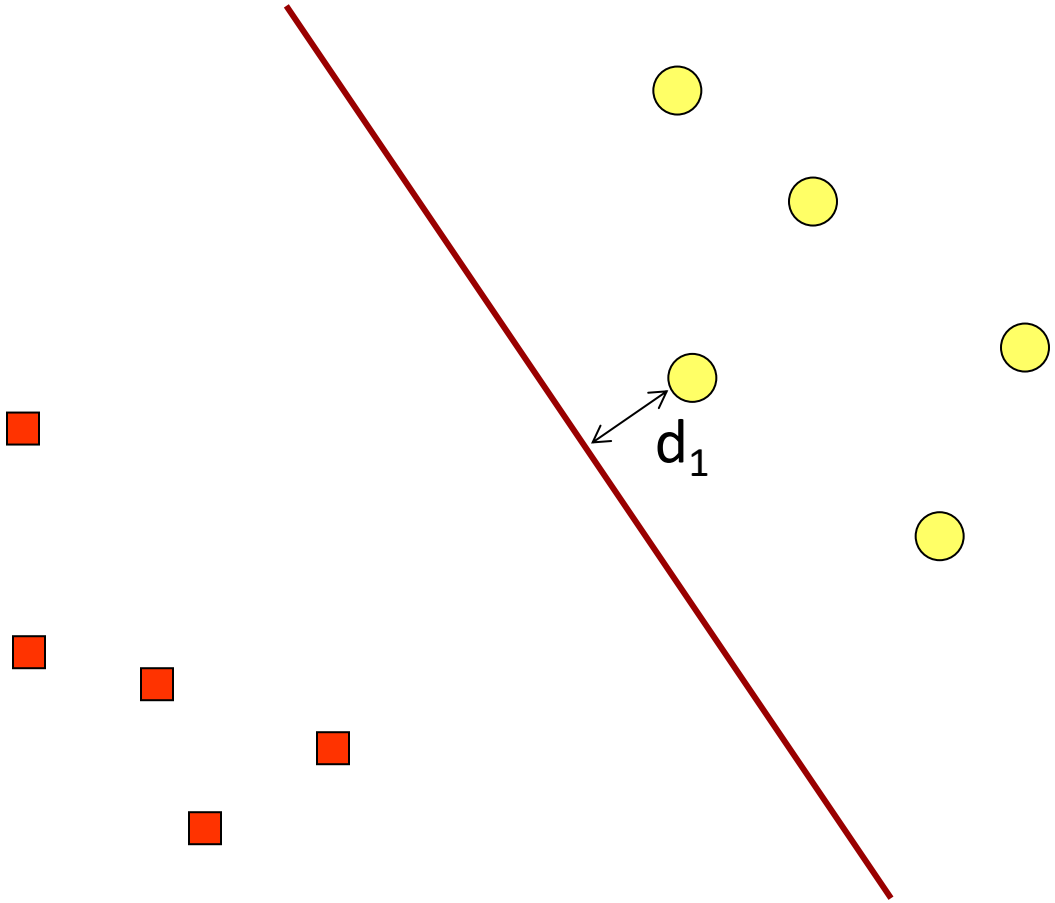
SVM's Approaches

- Hard margin
  - The two classes can be completely separated by the hyperplane.
  - In many cases this condition is not satisfied.
- Soft margin
  - Add slack variable $\xi_i \ \ (\xi_i > 0)$ that measures the distance of the point to its *marginal hyperplane.*

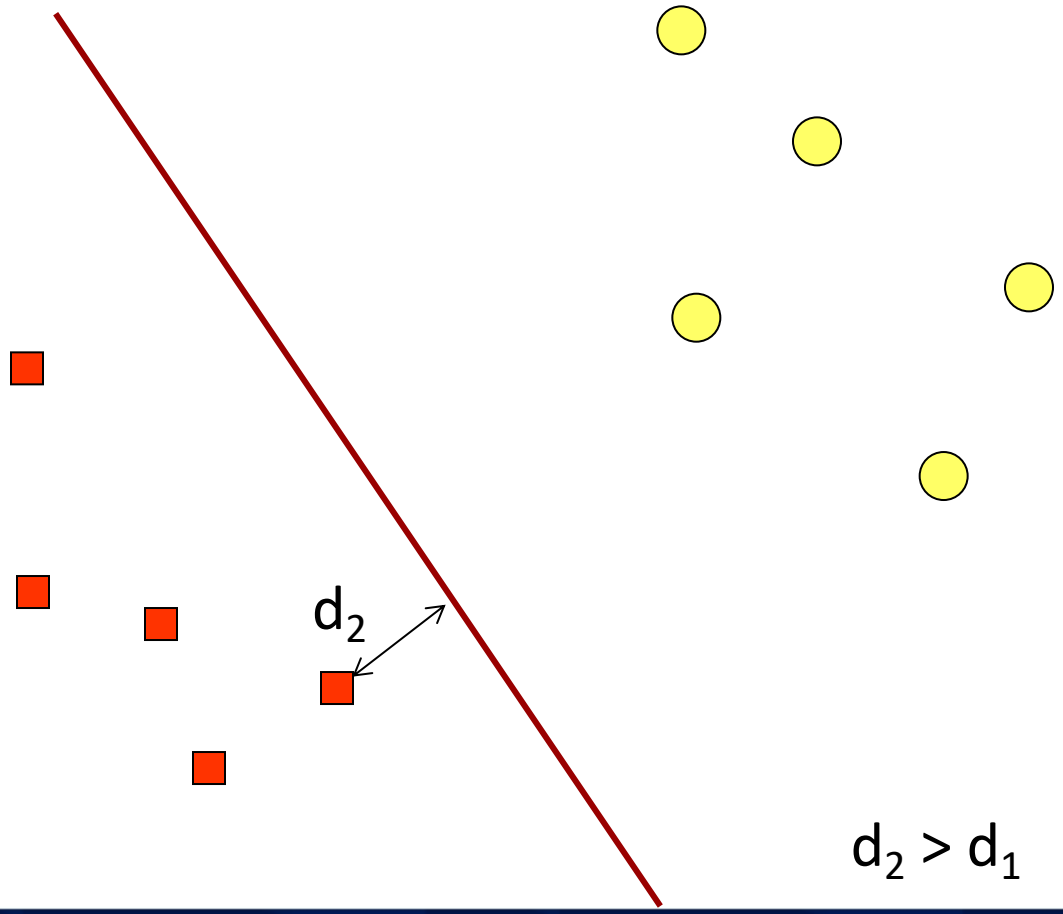
Hard Margin SVM

- Find the optimal hyperplane that maximize the margin.
- And separated the two classes perfectly.
- The effort to find the optimal hyperplane location is the *core* of the learning process in SVM.

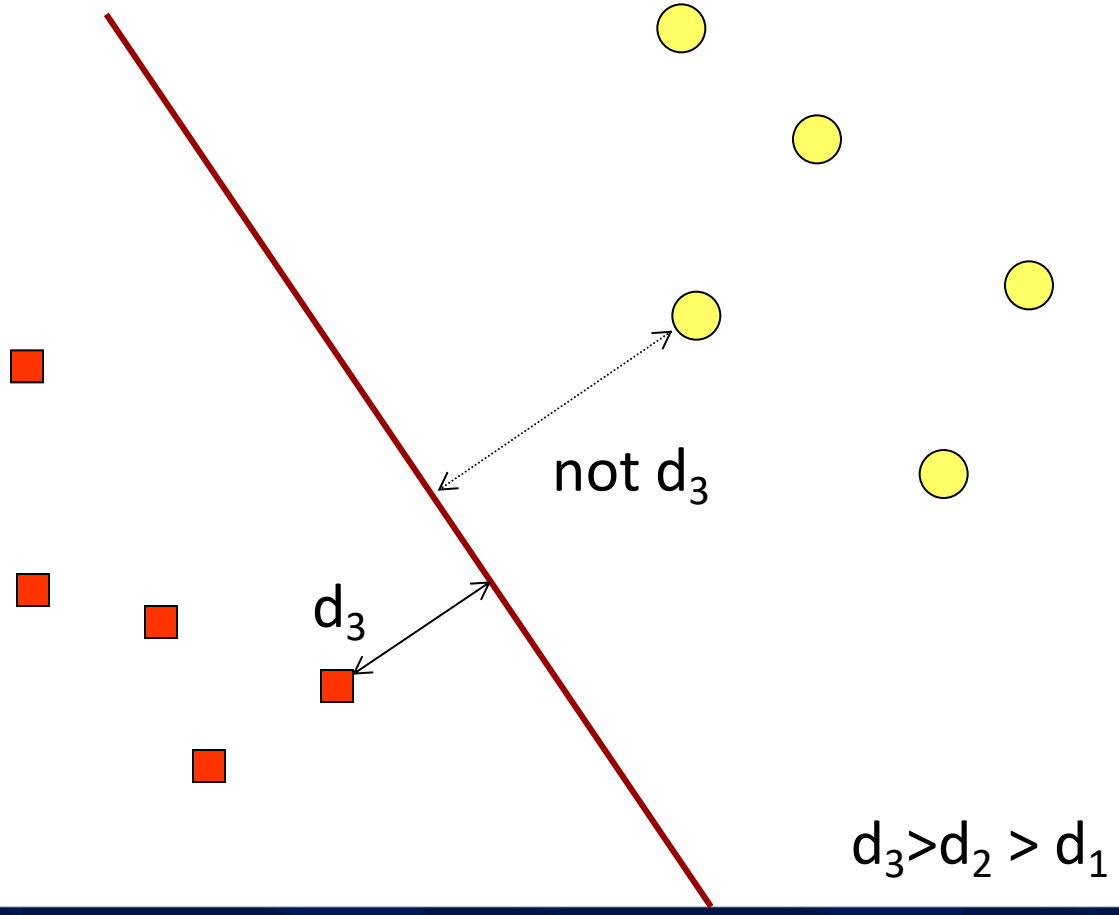
Optimal Hyperplane by SVM



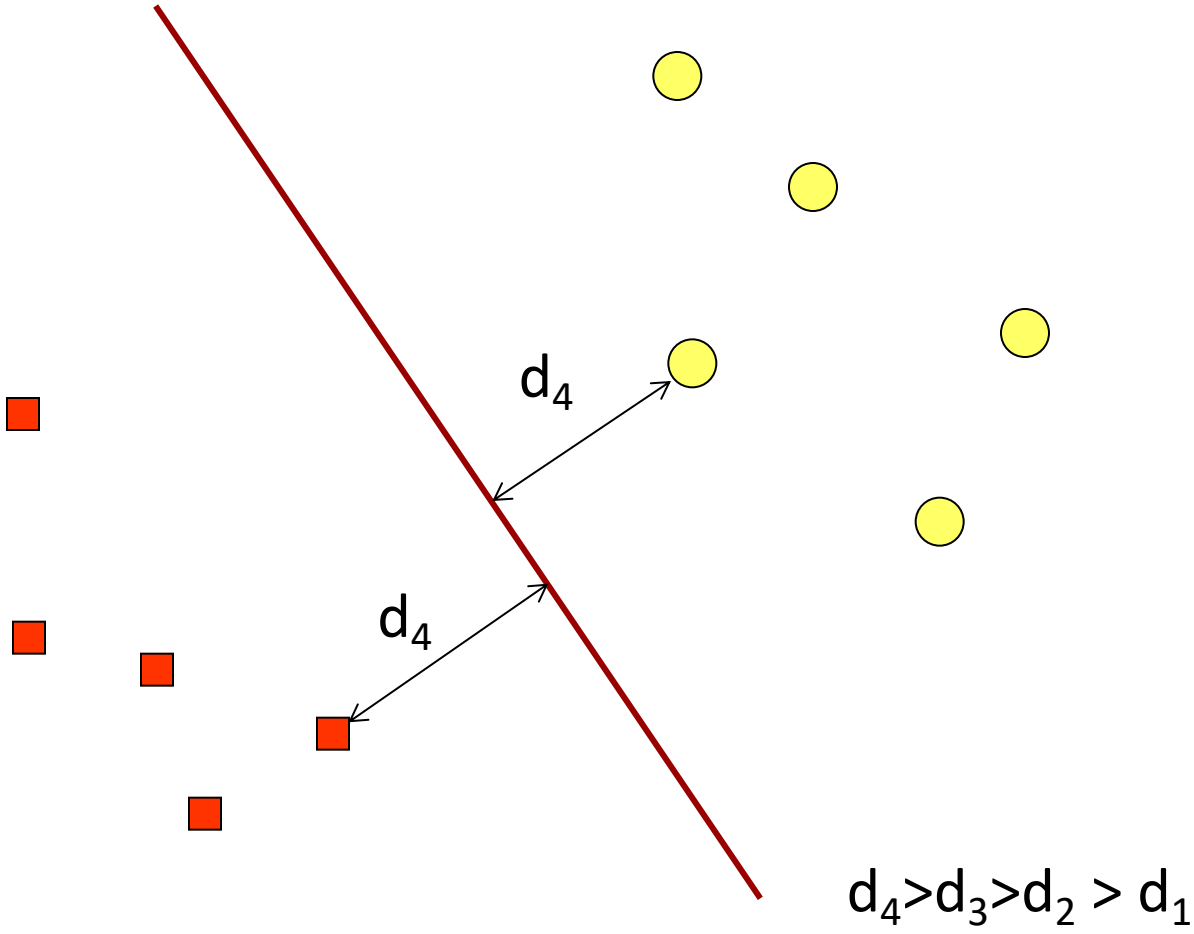
Optimal Hyperplane by SVM



Optimal Hyperplane by SVM



Optimal Hyperplane by SVM

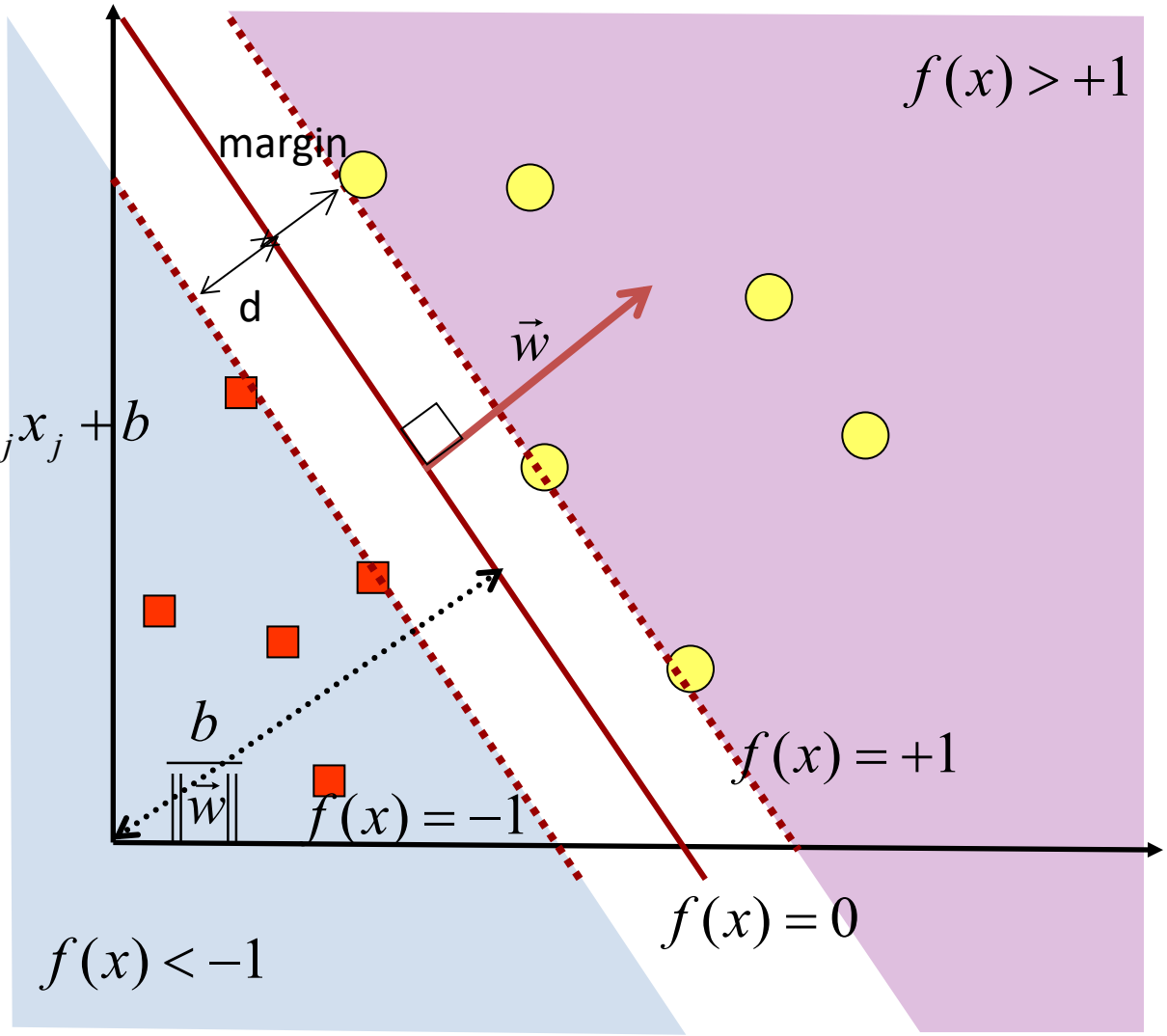


Separating Hyperplane for 2D

$$\vec{x} = (x_1, x_2, \cdots, x_d)^T$$

$$f(x) = <\vec{w} \cdot \vec{x}> + b = \sum_{j=1}^{\dim} w_j x_j + b \quad (1)$$

bias

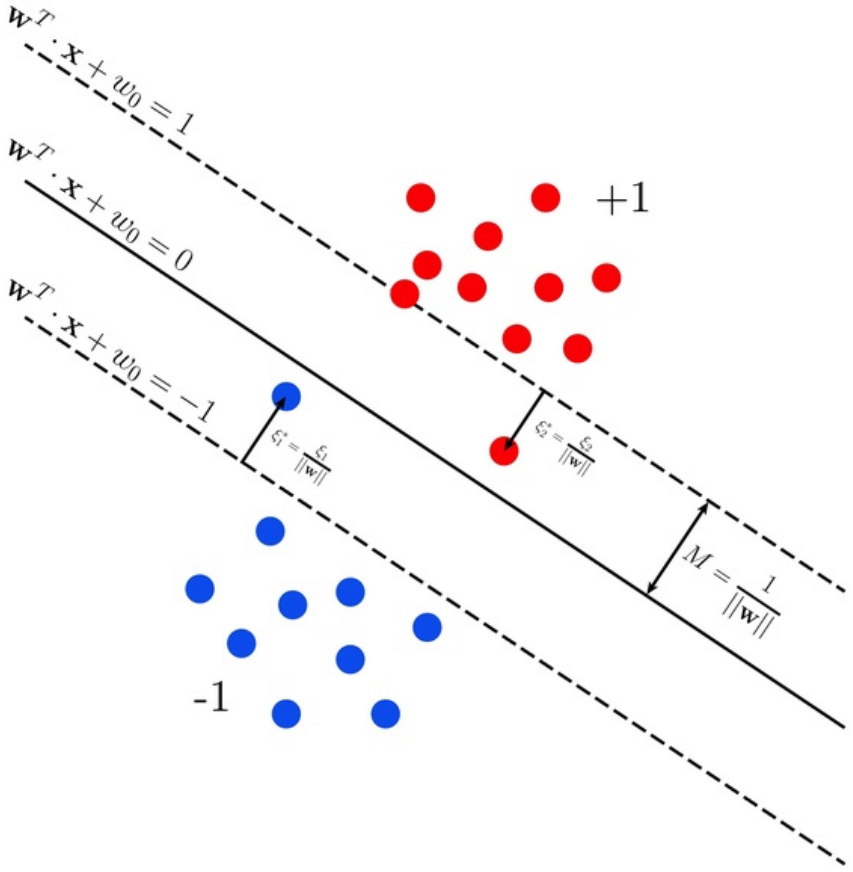


Mathematics Model

- Hyperplane: $\vec{w}.\vec{x} + b = 0$
  - $\vec{w}$: weight vector
  - b: bias
- A data $\vec{x}_i$ that belongs to class -1 (negative sample) can be formulated as data that meets the inequality: $\vec{w}.\vec{x}_i + b \leq -1$
- While data $\vec{x}_i$ which belongs to class +1 (positive sample): $\vec{w}.\vec{x}_i + b \geq +1$

Mathematics Model

- How to maximized?



Mathematics Model

- The largest margin can be found by maximizing the distance between the hyperplane and its closest point: $\frac{1}{\|\vec{w}\|}$
- Form:

PRIMAL FORM

Minimize $\|\vec{w}\|^2$ (1)

Subject to $y_i(<\vec{w}\cdot\vec{x}_i>+b)\geq 1 \quad (i=1,2,...,l)$ (2)

Mathematics Model

- This problem can be solved by various computational techniques, including **Lagrange Multiplier**:

$$L(\vec{w}, b, \alpha) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i \left((\vec{x}_i . \vec{w} + b) - 1 \right) \right) \quad (i = 1,2, \dots, n)$$

  - α_i is Lagrange multipliers, $\alpha_i \geq 0$

- Modified form:

DUAL FORM

Maximize α

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j < \vec{x}_i \cdot \vec{x}_j > \quad (3)$$

Subject to

$$\alpha_i \geq 0 \quad (i = 1,2,\dots,l) \qquad \sum_{i=1}^{l} \alpha_i y_i = 0 \quad (4)$$

Mathematics Model

DUAL FORM

Maximize
α

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j < \vec{x}_i \cdot \vec{x}_j > \quad (3)$$

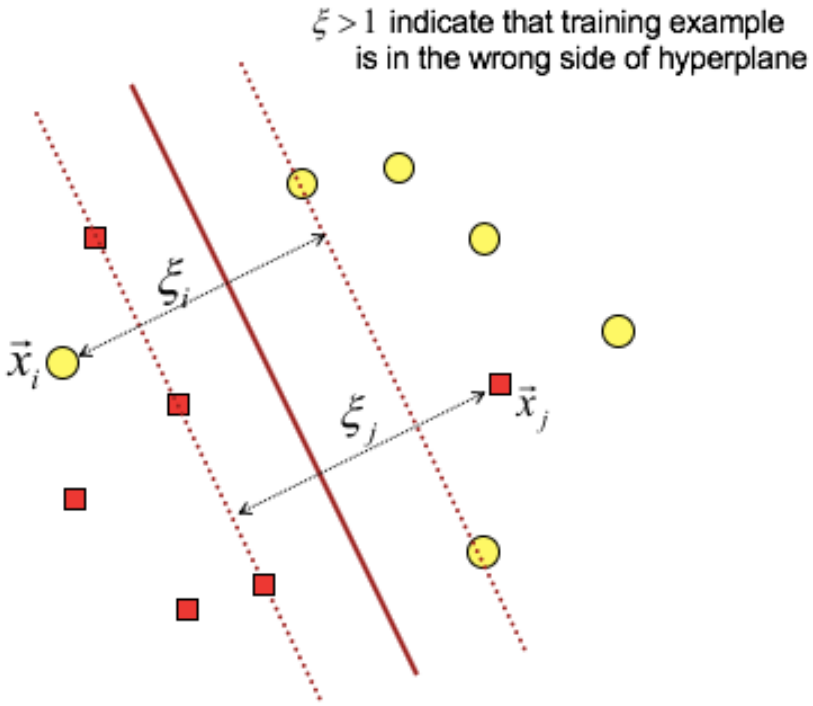
Subject to

$$\alpha_i \geq 0 \quad (i = 1,2,...,l) \qquad \sum_{i=1}^{l} \alpha_i y_i = 0 \quad (4)$$

- Quadratic Programming problem, which solution α_i is mostly 0.
- Data x_i from the training set which α_i is not 0 is called **Support Vector** (the most informative part of training set)

Soft Margin SVM

- Add slack variable ξ_i $(\xi_i > 0)$ that measures the distance of the point to its *marginal hyperplane.*



Soft Margin SVM

- Minimize: $\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{l}\xi_i$

  – Subject to: $y_i(<\vec{w}\cdot\vec{x}_i> + b) \geq 1 - \xi_i \quad (i = 1,2,...,l)$

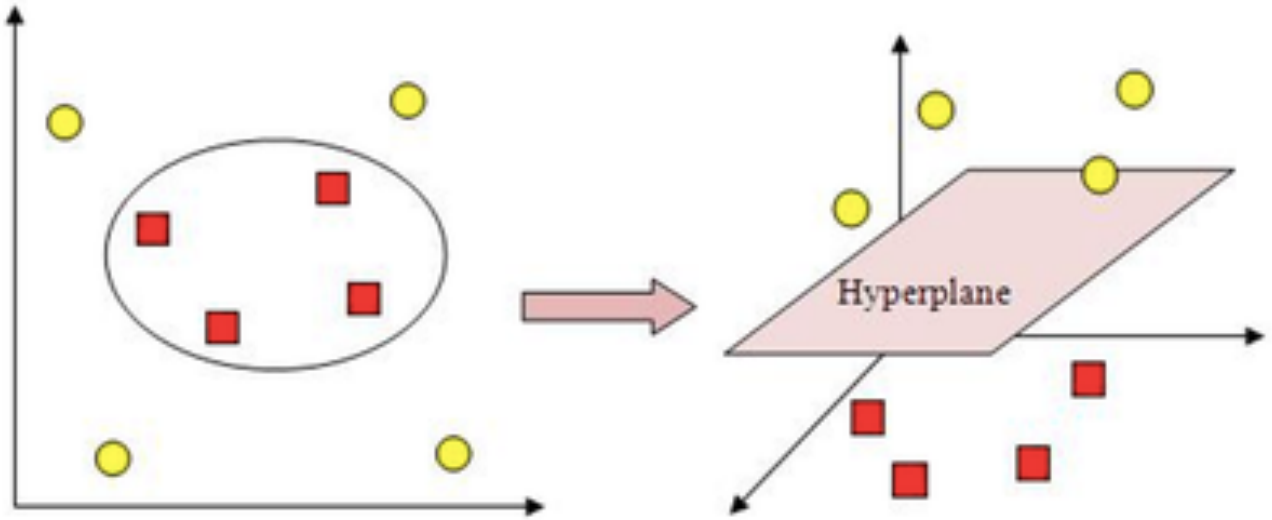
- C is a parameter that controls the trade off between margin and classification error ξ. The greater the value of C, the greater the penalty for errors, so the training process becomes tighter.

Kernel Trick

- In general, problems in the real world domain are rarely linearly separable.
- To solve the non-linear case, the SVM calculation is modified into two stages:
  - First of all the data $\vec{x}$ is mapped by the function $\Phi(\vec{x})$ to the vector space of a higher dimension.
  - In this new vector space, the hyperplane that separates the two classes linearly can be constructed.

Kernel Trick - illustration

$$\Phi : \Re^d \rightarrow \Re^q \; d < q$$



Input Space X
(a)

High-dimensional Feature Space $\Phi(X)$
(b)

Kernel Function

Kernel	Definition
Linear	$K(x_i.x_j) = (x_i.x_j)$
Polynomial	$K(x_i.x_j) = (x_i.x_j + 1)^p$
Gaussian RBF	$K(x_i.x_j) = exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$

More to Explore: Multiclass SVM

- SVM originally works for binary classification (two classes).
- Can be extended to solve multiclass problem.
- Methods:
  - One-Against-All
  - One-Against-One
  - Decision Directed Acyclic Graph (DDAG)
  - Adaptive Directed Acyclic Graph (ADAG)

Hands On [10]

- Please download the code from this link:

https://tinyurl.com/SVM-hands-on

References

[1] Wee Sun Lee. “Introduction to Machine Learning”. In Southeast Asia Machine Learning School (2019).

[2] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." *International journal of computer vision*115.3 (2015): 211-252.

[3] Zillow Prize: Zillow’s Home Value Prediction (Zestimate) [Online: https://www.kaggle.com/c/ zillow-prize-1, accessed February 2020].

[4] Sutton, Richard S., and Andrew G. Barto. "Reinforcement learning: An introduction." (2011).

[5] Applications of Reinforcement Learning in Real World [Online: https://towardsdatascience.com/applications-of-reinforcement-learning-in-real-world-1a94955bcd12, accessed February 2020].

[6] Semi-Supervised Machine Learning [Online: https://www.datarobot.com/wiki/semi-supervised-machine-learning/, accessed February 2020].

References

[7] Support Vector Machines [Online: https://rpubs.com/jgab3103/460138, accessed February 2020].

[8] Anto Satriyo Nugroho. ”Pengantar Support Vector Machine." (2008).

[9] What is the purpose for using slack variable in SVM? [Online: https://www.quora.com/What-is-the-purpose-for-using-slack-variable-in-SVM, accessed February 2020].

[10] Beyeler, Michael. *Machine Learning for OpenCV*. Packt Publishing Ltd, 2017.