An Overview of SVM

Statistics Summer School, July 30th-Aug 10th, 2018

I-Ping Tu

Institute of Statistical Science,

Academia Sinica

July 28, 2018

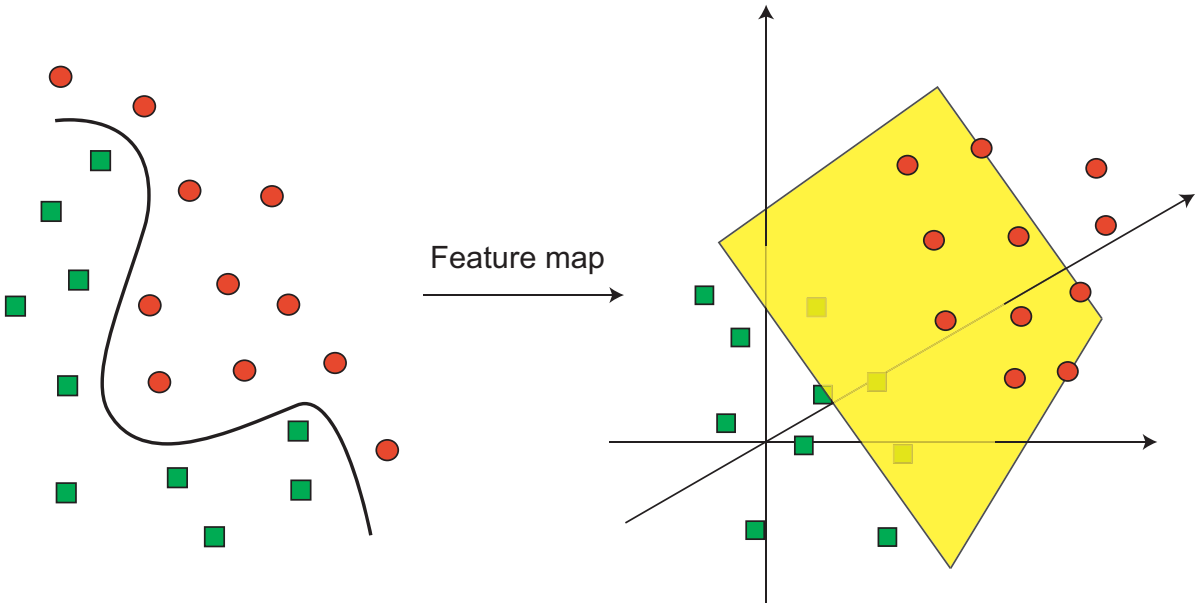
Learning Goals of this Lecture:

To answer these questions.

1. Who is SVM?
2. What is LDA? A model-base classification method, far before SVM.
3. Where did SVM come from?
4. What are the characters of SVM?
5. SVM and its sibling–Logistic Regression

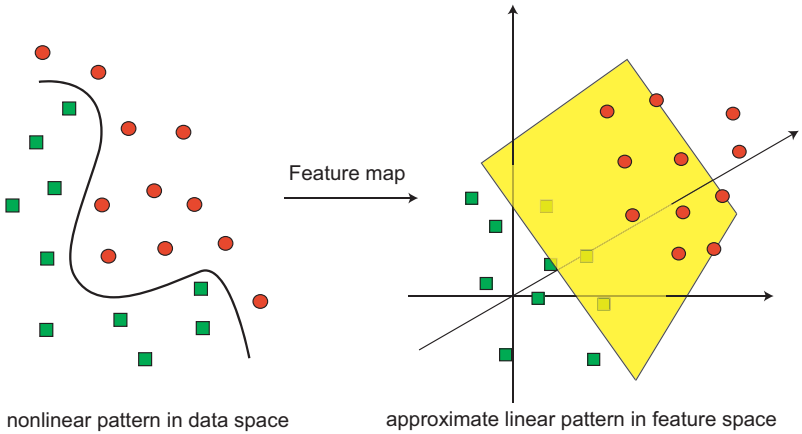
Support Vector Machine (SVM)

- A Classification Machine: to find a boundary for decision making.



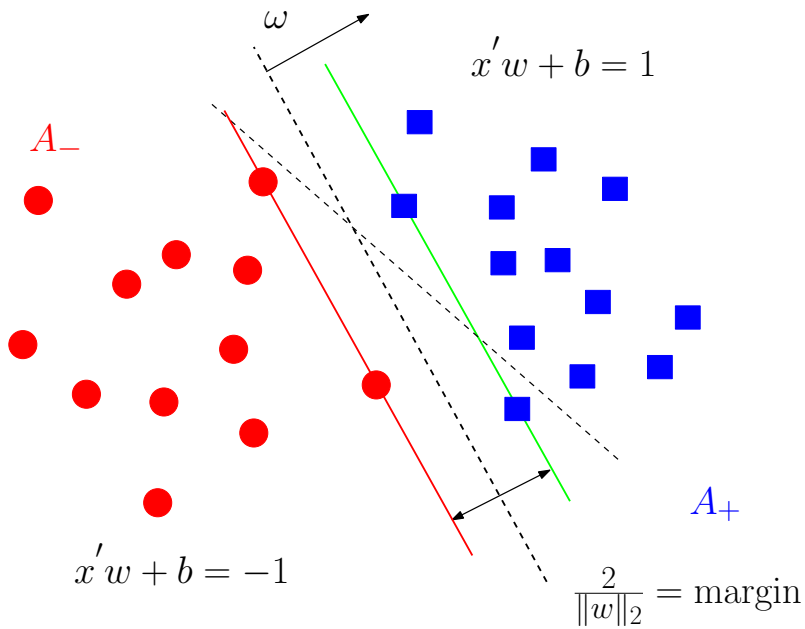
How to apply SVM

- Given a training set.
- Find the decision boundary (made by the support vectors).
- Apply on the testing data or future data points.



Why is it called SVM

- The decision function has the form $f(x) = w'x + b$.
- w is the linear combination of the "support vectors".



R. A. Fisher (1890-1962)

Linear Discriminant Analysis

- Fisher was born in London, England in 1890, with lifelong poor eyesight.
- He had strong geometric intuition.
- His works has great impacts on the fields of Statistics, Genetics, and Evolutionary biology.

Statistical Modeling $Y \sim X$:

Y : response variable,

X : predictor(s) or independent variable(s).

- Y continuous, X continuous: Regression Analysis.
- Y discrete, X continuous:
  - X is Gaussian: Linear Discriminant Analysis.
  - X is non-Gaussian: Logistic Regression.
- Y discrete, X discrete: Discriminant Correspondent Analysis.
- Y continuous, X discrete: ANOVA.

An Example of Simple Linear Regression

$$Y = \beta_0 + \beta_1 X + \epsilon,\ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

ISLR Sixth Printing.pdf（頁面 76/440）

62 3. Linear Regression

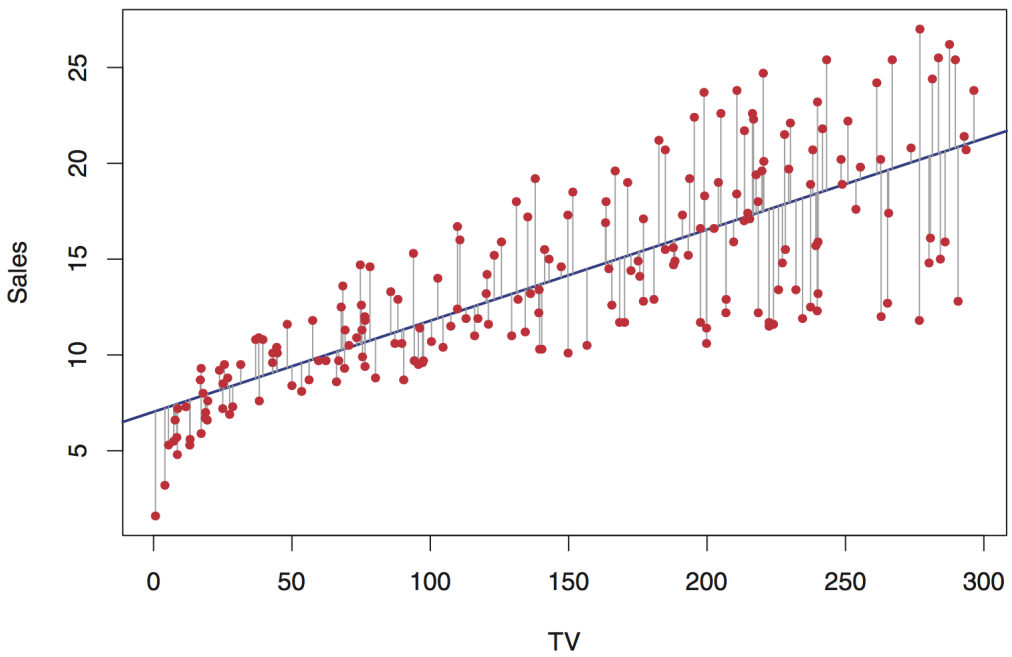
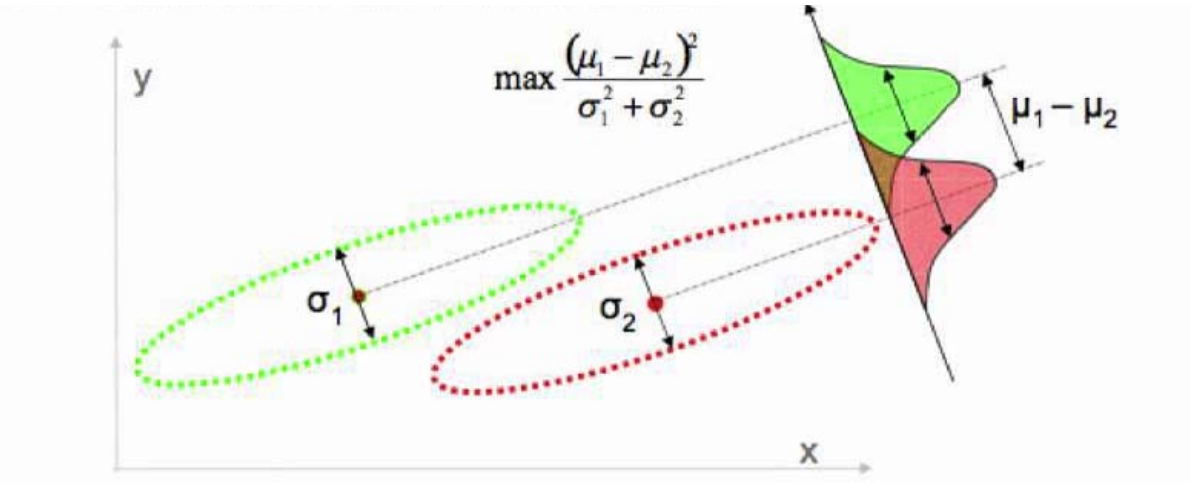


FIGURE 3.1. *For the* `Advertising` *data, the least squares fit for the regression*

Linear Discriminant Analysis

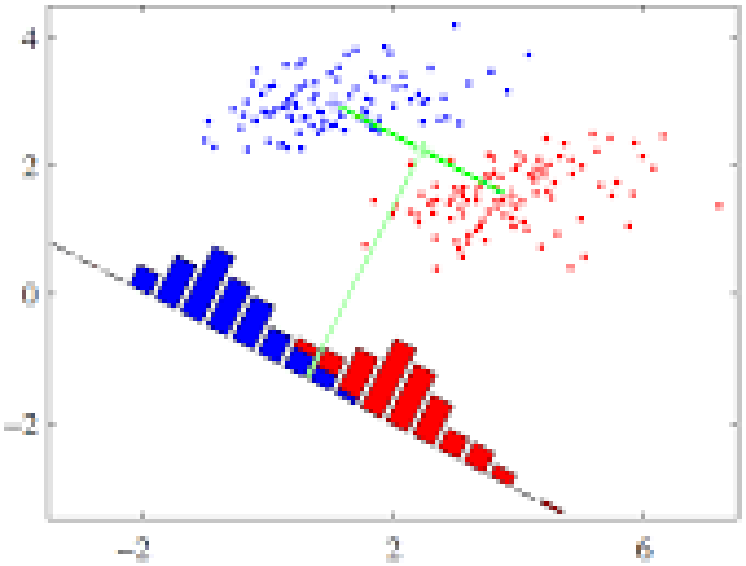
- Gaussian Mixture Model
- Concept: Use parameters (μ_1, σ_1), (μ_2, σ_2) to describe the groups.
- Its decision is based on a minimum Mahalanobis distance:

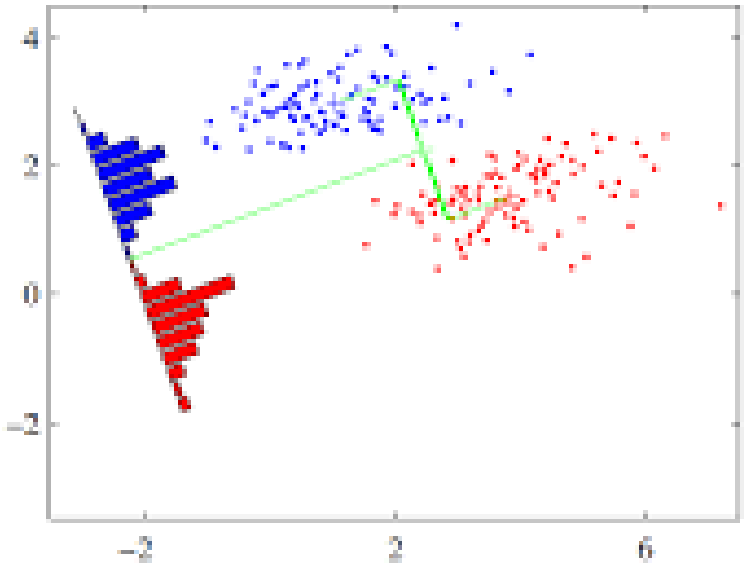
  Given a data point x, choose group 1 if $\frac{|x-\mu_1|}{\sigma_1} < \frac{|x-\mu_2|}{\sigma_2}$.



Learning from LDA

- All the points with equal distances to two center points form a hyperplane.





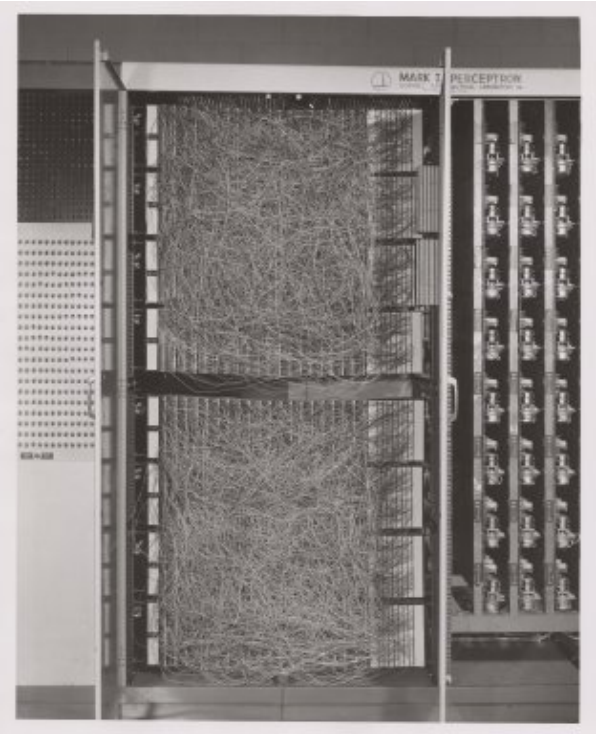
Frank Rosenblatt (1928-1971)

Perceptron (parallel to Neuron and Electron)

- Mark I Perceptron (1957-1959): A hardware construction of a visual pattern classifier.
- He built a telescope for astronomical observations and composed music.

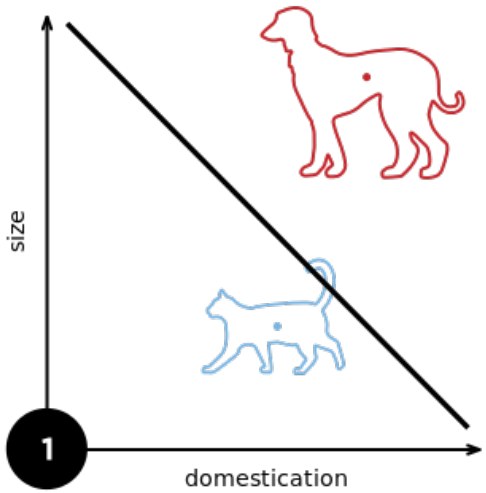
Mark I Perceptron

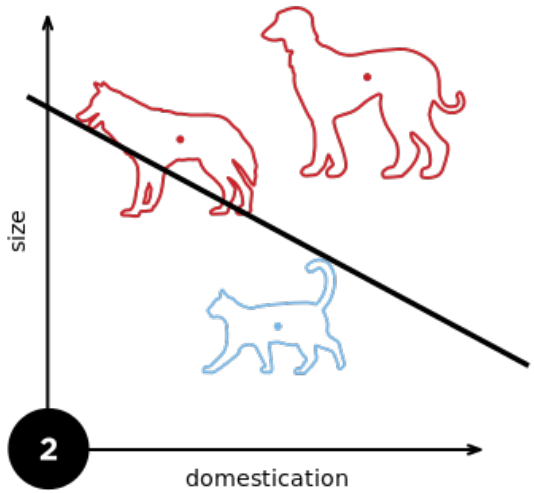
- Mark I Perceptron (1957-1959): A hardware construction of a visual pattern classifier.

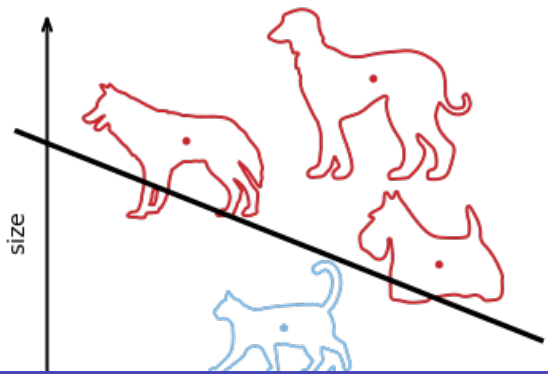


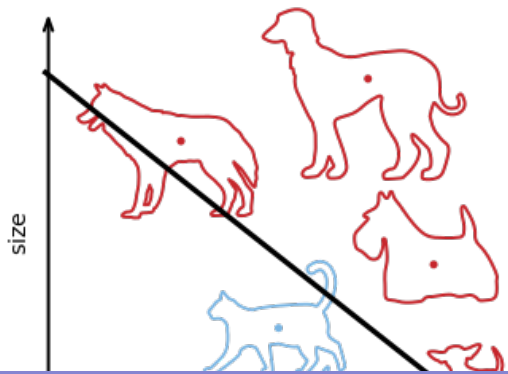
Perceptron: I only need to know the hyperplane!

Why bother a model?









Perceptron: On-Line learning Classification

- The hyperplane is sequentially updated by the training set. Calculate the actual output and then update the weight whose increment is proportional to the difference of the actual output and the desired output.
- Given an initial $\mathbf{w}$, for each training pair $(\mathbf{x}_i, y_i)$ where $1 \leq i \leq n$, do the two steps:
  - $o_i = f(\mathbf{w}^T(t)\mathbf{x}_i)$ (the decision rule)
  - $\mathbf{w}(t+1) = \mathbf{w}(t) + \alpha(o_i - y_i)\mathbf{x}_i$

From Perceptron to SVM

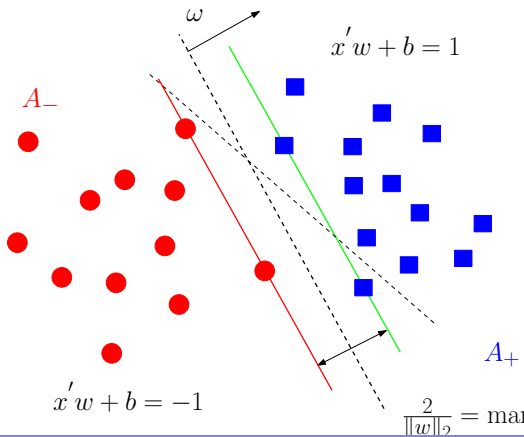
- Introduce the concept of Margin $\rightarrow$ a unique solution.
- Introduce the slack variables $\rightarrow$ soft margin.
- Introduce Kernel method $\rightarrow$ non-linear feature mapping.

Biography of Vladimir N. Vapnik

- Vladimir Vapnik was born in the Soviet Union.
- He received his master's degree in mathematics at the Uzbek State University, Samarkand, Uzbek SSR in 1958.
- He got his Ph.D in statistics at the Institute of Control Sciences, Moscow in 1964.
- He worked at this institute from 1961 to 1990 and became Head of the Computer Science Research Department.
- At the end of 1990, Vladimir Vapnik moved to the USA and joined Bell Labs, where Vapnik and his colleagues developed the theory of the support vector machine.

Support Vector Machine

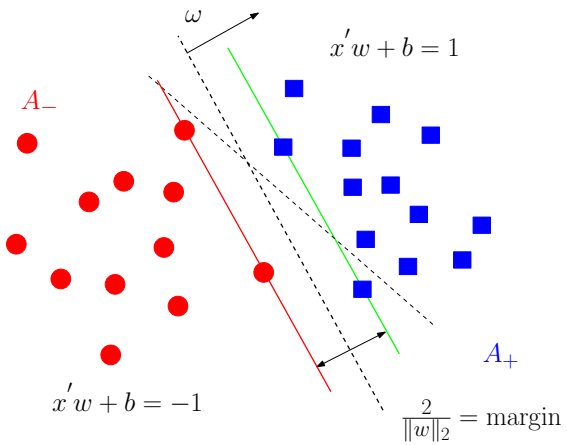
- Simple Definition: Support Vector Machine is to find a hyperplane to separate the features of the data objects for classifications.
- Advantages: *The key features of SVMs are the use of kernels, the absence of local minima, the sparseness of the solution and the capacity control obtained by optimizing the margin.* by Shawe-Taylor and Cristianini (2004)



Maximize the Margin: Elegence of SVM

$$w'x_i + b \geq +1 \quad \text{for all} \quad y_i = +1,$$

$$w'x_i + b \leq -1 \quad \text{for all} \quad y_i = -1.$$

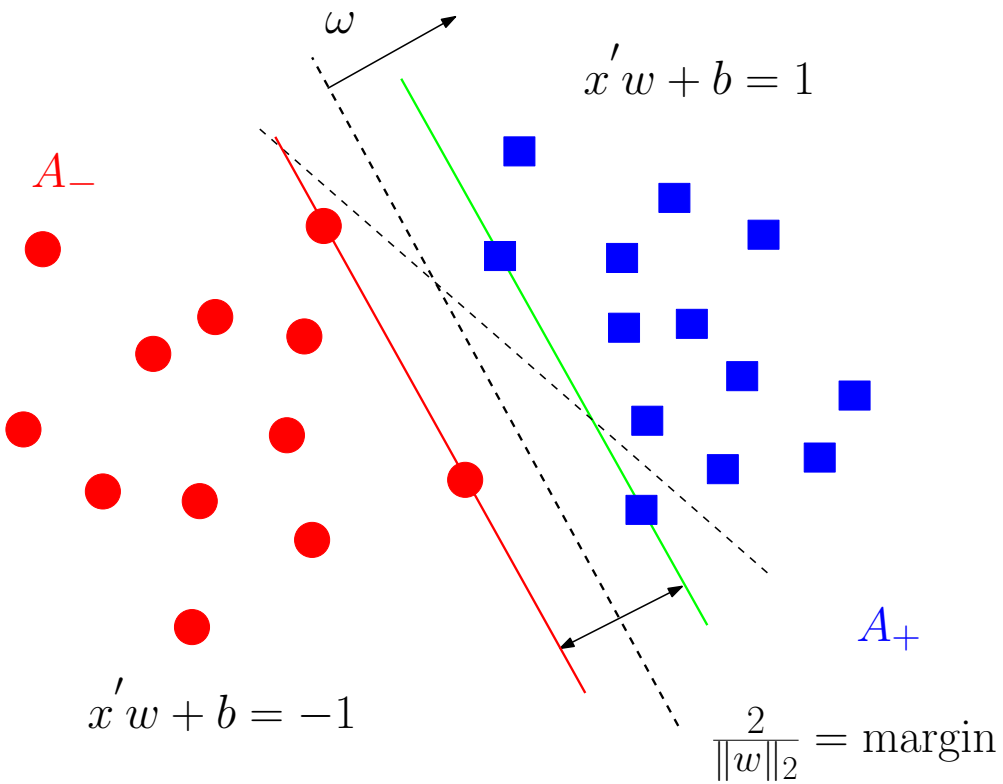


Elegence: Two become One!

$$1 - y_i(w'x_i + b) \leq 0, \quad i = 1, \ldots, n.$$

A Simple Optimization Criterion is formed for SVM

$$\min_{w\in\Re^p, b\in\Re} \frac{\|w\|_2^2}{2} \text{ subject to } 1 - y_i(w'x_i + b) \leq 0, \quad \forall i.$$



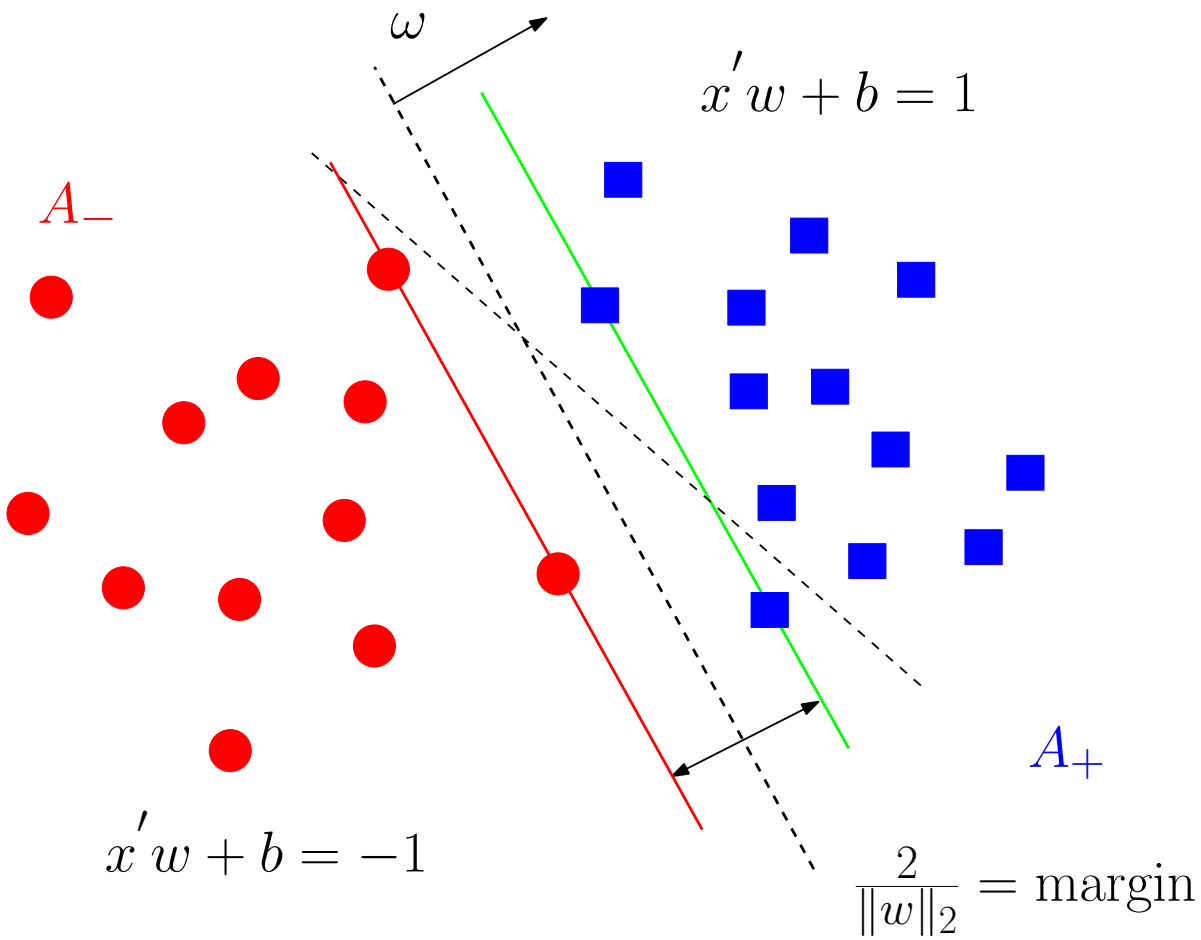
History of Support Vector Machine

Margin Concept

- Cover (1965) discussed large margin hyperplanes in the input space and sparseness.
- Duda and Hart (1973) discussed large margin hyperplanes in the input space.

Support Vector Machine

hyperplane; maximize the margin.



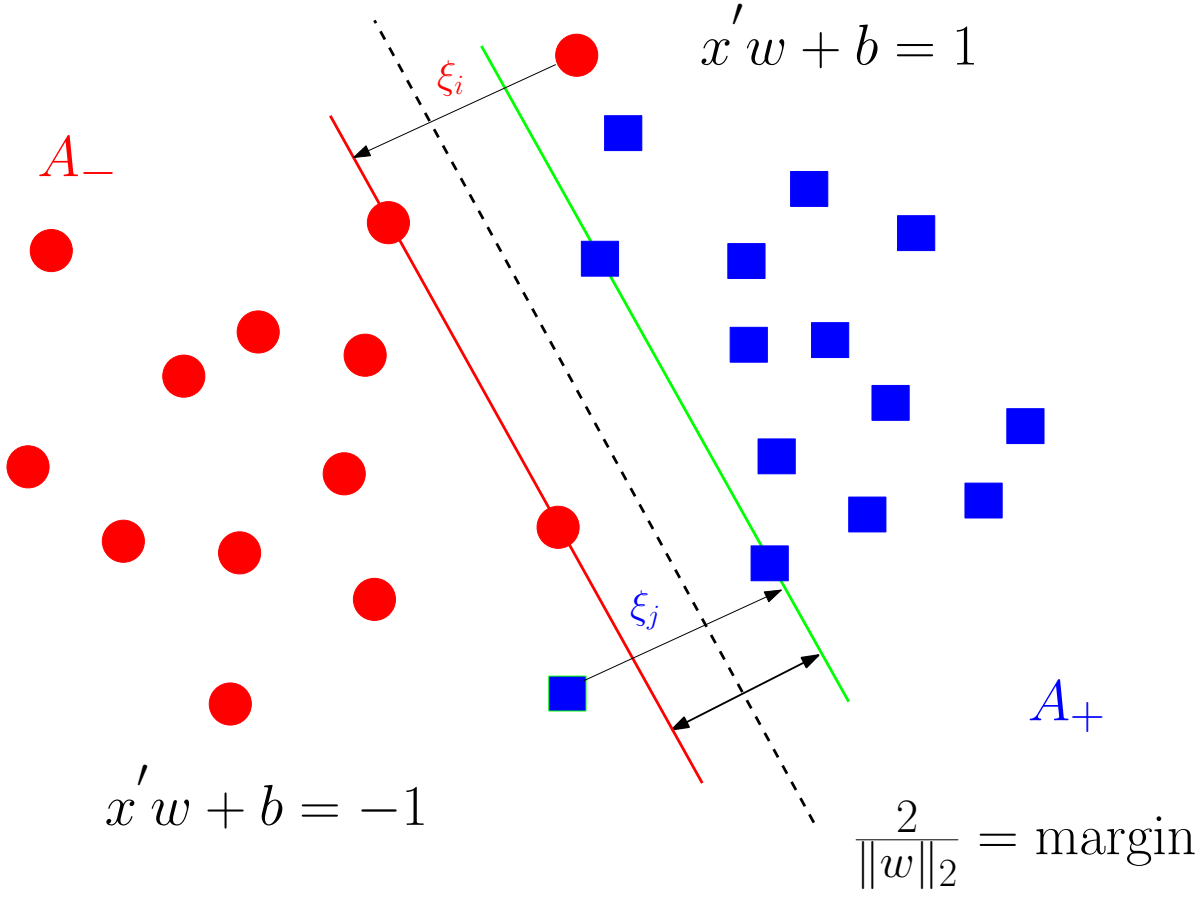
History of Support Vector Machine

Soften the margin by slack variables.

- The use of slack variables to overcome the problem of noise and non-separability was introduced by Smith (1968).
- Bennett and Mangasarian (1992) improved upon Smith's 1968 work on slack variables.

Support Vector Machine

Soft Margin: Introducing the slack variables.



Support Vector Machine

The hyperplane is adjusted for the slack variables.

$$w'x_i + b \geq +1 - \xi_i \quad \text{for all} \quad y_i = +1,$$

$$w'x_i + b \leq -1 + \xi_i \quad \text{for all} \quad y_i = -1,$$

$$\xi_i \geq 0 \qquad \forall i.$$

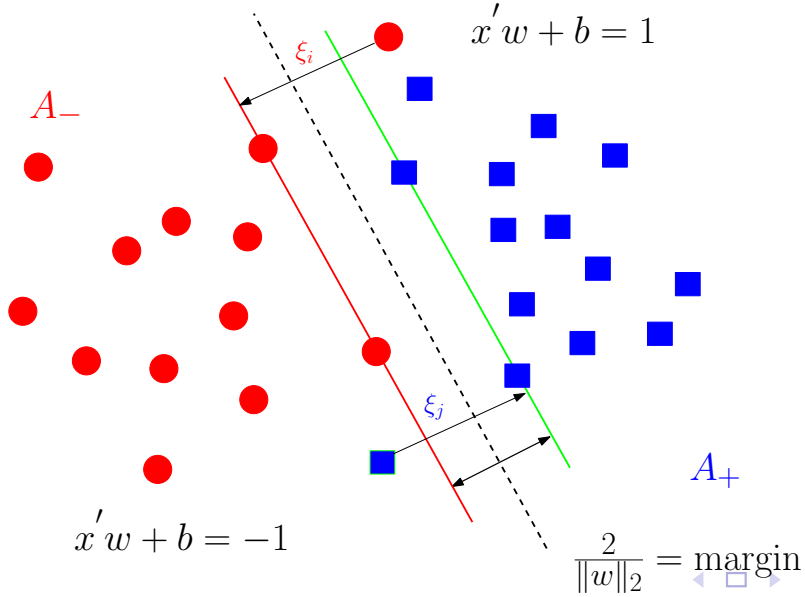
Two become One!

$$1 - y_i(w'x_i + b) - \xi_i \leq 0 \quad \text{and} \quad \xi_i \geq 0, \ \forall i.$$

The Optimization Criteria for SVM

$$\min_{w\in\Re^p, b\in\Re, \xi\in\Re^n} \frac{\|w\|_2^2}{2} + C\sum_{i=1}^{n}\xi_i \text{ subject to } 1 - y_i(w'x_i + b) - \xi_i \le 0, \ \ \xi_i \ge 0, \ \ \forall i.$$

C is a tuning parameter to balance the margin and the slack variable (error endurance).



The Optimization Criteria for SVM

With Lagrangian multipliers $\alpha_i \geq 0$ and $\beta_i \geq 0$ $1 \leq i \leq n$:

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|_2^2 + C\mathbf{1}'\xi + \alpha' \left[\mathbf{1} - D(Xw + b\mathbf{1}) - \xi\right] - \beta'\xi,$$

where D is a diagonal matrix with y_i in the diagonals. Variables w, b, ξ are called *primal variables*, and variables $\alpha, \beta \in \Re^n$ are called *dual variables*.

The Solution for SVM

$w = \sum_{j=1}^{n} \alpha_j y_j x_j$.

$b = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left(y_i - \sum_{j=1}^{n} \alpha_j y_j x_j' x_i \right)$, where $\mathcal{I} = \{i : 0 < \alpha_i < C\}$.

$$f(x) = w'x + b = \sum_{i=1}^{n} \alpha_i \, y_i \, x_i' \, x + b = \sum_{x_i : \mathrm{SV}} \alpha_i \, y_i \, \boxed{x_i' \, x} + b. \qquad (1)$$

- Points x_i associated with nonzero α_i are called support vectors (SVs).
- Only the SVs contribute to the decision function for classification.
- In addition, the contribution of an input instance as a SV is at most at a scale C in the final decision function.

History of Support Vector Machine

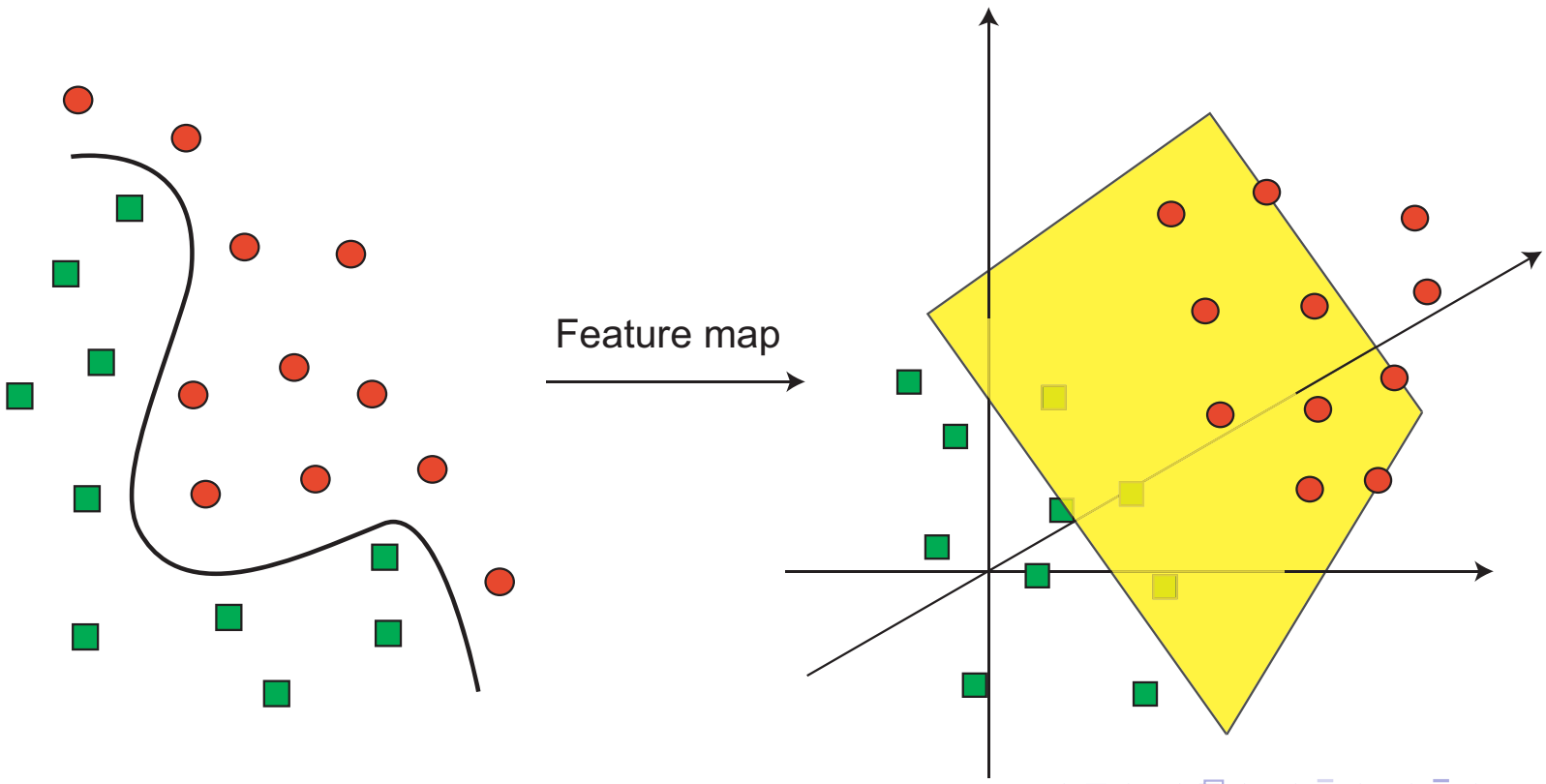
Nonlinear generalization

- Aronszajn (1950) introduced the **Theory of Reproducing Kernels**.
- Aizerman, Braverman and Rozonoer (1964) introduced the geometrical interpretation of the kernels as inner products in a feature space.
- Poggio and Girosi (1990) and Wahba (1990) discussed the use of kernels.

Support Vector Machine

Kernel Method

- Linear vs Non-linear is like Apple to Non-Apple. I



nonlinear pattern in data space

approximate linear pattern in feature space

Support Vector Machine

Kernel Method

- Main idea: feature mapping to a high dimensional space, $\Phi : \mathcal{X} \mapsto \mathcal{Z}$

$$x \mapsto \Phi(x) = (\phi_1(x), \phi_2(x), \phi_3(x), \dots) =: z \in \mathcal{Z}. \tag{2}$$

Inner product in $\mathcal{Z}$: $\Phi(x)'\Phi(u) = \sum_q \lambda_q \phi_q(x)\phi_q(u) = K(x, u)$.

- Kernel Trick: We do not really need to know $\Phi(x)$. Instead, we work on the Kernel $K(x, u)$.

Logistic Regression

- If we let $y \in \{1, -1\}$, $P(y, \theta = x^T\beta) = \frac{1}{1+\exp(-yx^T\beta)}$.
- For training data $(x_i, y_i), 1 \leq i \leq n$, we want to find MLE of β which minimizes $\sum_{i=1}^{n} \log(1 + \exp(-y_i x_i^T \beta))$
- Regularized Logistic Regression:

$$\min_{\beta} \left(\frac{1}{2}\beta^T\beta + C\sum_{i=1}^{n} \log(1 + \exp(-y_i x_i^T \beta)) \right)$$

- SVM:

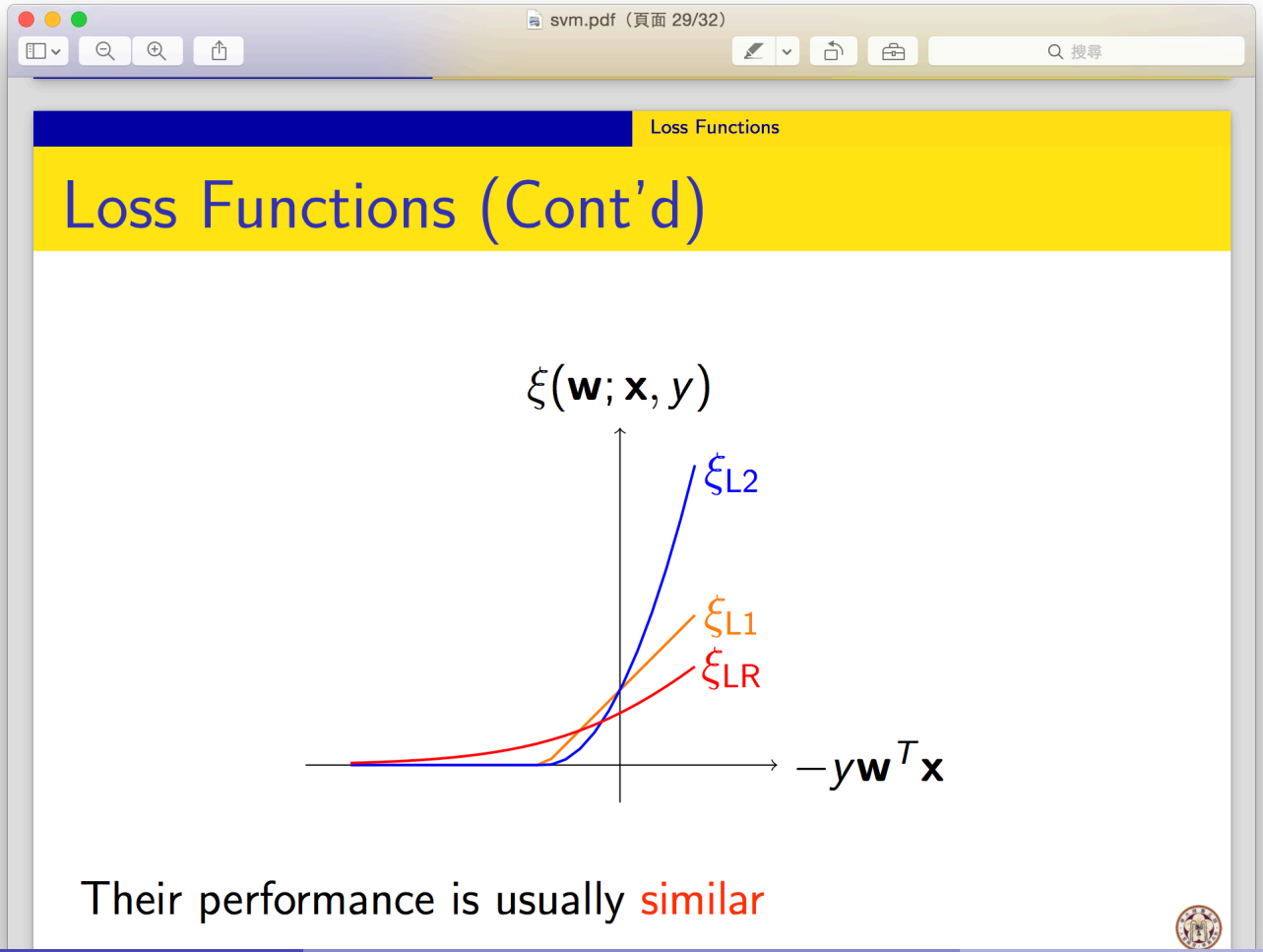
$$\min_{\mathbf{w}} \left(\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} \max\{0, 1 - y_i x_i^T \mathbf{w}\} \right)$$

- Remember that

$$1 - y_i(w'x_i + b) < \xi_i \quad \text{and} \quad \xi_i > 0, \forall i.$$

Support Vector Machine and Logistic Regression

- Loss functions from CJ Lin's SVM lecture Slides.



Summary

1. Who is SVM?
2. What is LDA? A model-base classification method, far before SVM.
3. Where did SVM come from?
4. What are the characters of SVM?
5. SVM and its sibling–Logistic Regression