Margin-Based Models

Qing-Yuan Jiang

The school of Computer Science and Engineering, NJUST

2026.10.15

Review of Previous Class

- What is the basic idea of a decision tree?
- How does a decision tree partition the feature space and make predictions?
- How do entropy, information gain, and Gini impurity help us choose a split?
- Why do deep trees overfit, and how can we control tree complexity?
- Why can combining multiple learners improve generalization?
- How does bagging use bootstrap sampling and aggregate predictions?
- How does boosting differ from bagging?
- How do AdaBoost and gradient boosting improve the ensemble at each step?

Support Vector Machine

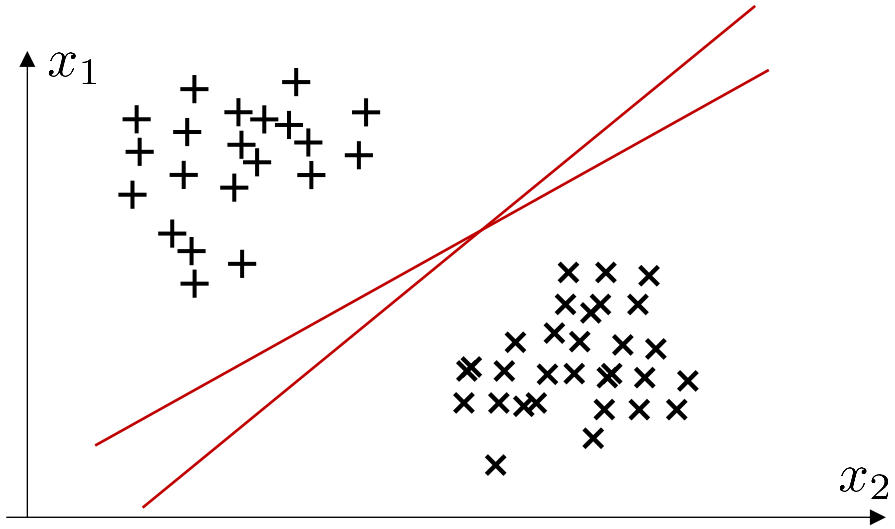
Classification Problem

- Suppose that we are given a training set:

$$\mathcal{D} = \{(x_i, y_i) | y_i \in \{-1, +1\}\}_{i=1}^{n}$$

- A classifier:

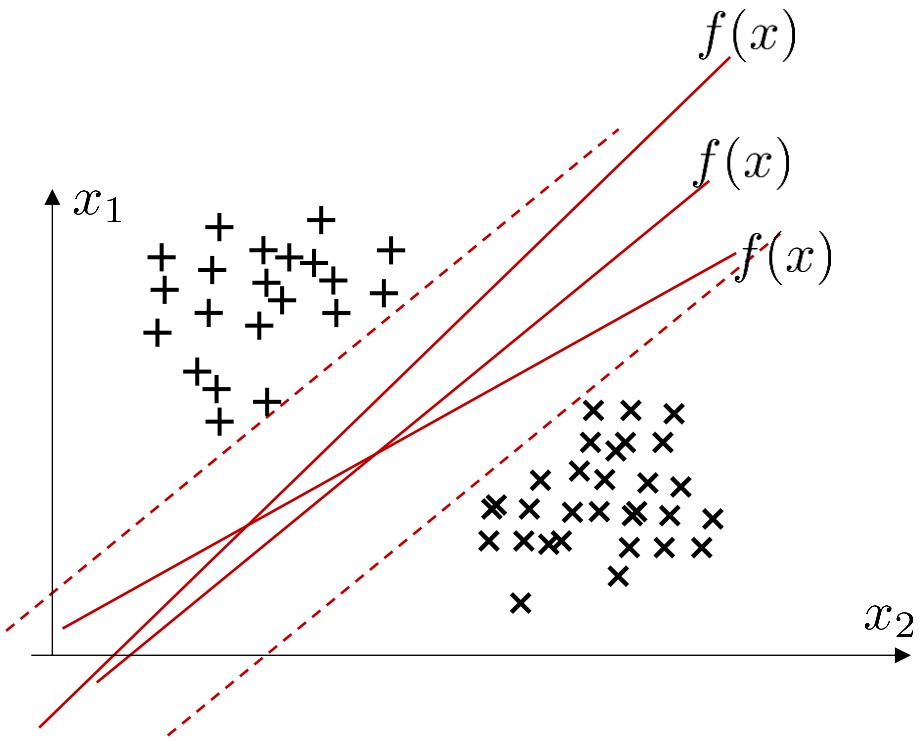
$$f(x) = w^\top x + b$$
$$\hat{y} = \text{sign}\left(f(x)\right)$$



Multiple Decision Boundaries

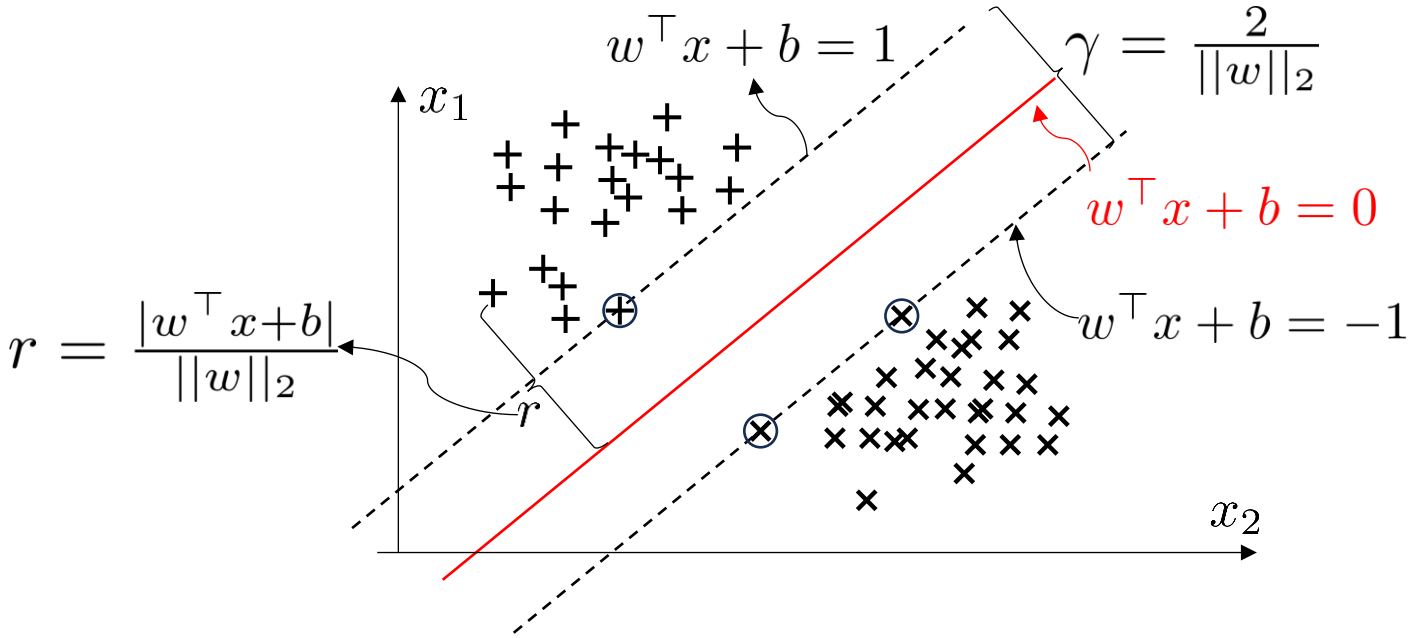
- Several lines can achieve zero training errors.
- Training accuracy alone does not select one boundary.

The separation from the nearest samples differs across lines.



What Is a Margin?

- The margin measures separation from the closest samples.



$$\text{margin-width} = \frac{2}{||w||_2}$$

Functional Margin vs. Geometric Margin

Functional margin	Geometric margin
$\gamma_i = y_i(w^\top x + b)$	$\gamma_i = y_i(w^\top x + b)/\|\|w\|\|_2$
Changes under positive rescaling	Signed distance to the boundary

- The decision boundary stays the same.
- Functional margin triples. Geometric margin is unchanged.

Why Maximize the Margin?

- A larger margin separates classes more widely.
- The closest training samples limit the achievable margin.
- Maximum margin selects among separating hyperplanes.
- Generalization also depends on the distribution and sample size.

The separation from the nearest samples differs across lines.

Maximizing the Margin

- Correct classification requires positive signed margins.
- Maximize the smallest geometric margin.

$$\max \quad \gamma$$

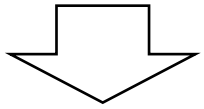
$$s.t.\ \forall i, y_i(w^\top x_i + b) \geq \gamma ||w||_2$$

Assume $w \neq 0$ **and linearly separable data.**

Support Vector Machine

- Maximize the margin:

$$\max \quad \frac{2}{||w||_2}$$
$$s.t.\ \forall i, y_i(w^\top x_i + b) \geq 1$$



$$\min \quad \frac{1}{2}||w||_2$$
$$s.t.\ \forall i, y_i(w^\top x_i + b) \geq 1$$

Convex Quadratic Programming Problem

Dual Problem

- Introduce Lagrange multipliers $\alpha_i \geq 0$ to obtain the Lagrangian function:

$$L(w, b, \alpha) = \frac{1}{2}||w||_2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(w^\top x_i + b))$$

- Let the partial derivatives of $L(w, b, \alpha)$ with respect to w and b be zero, we have:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i, \sum_{i=1}^{m} \alpha_i y_i = 0$$

- The we have:

$$\max_{\alpha} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^\top x_j$$

$$s.t. \quad \sum_{i=1}^{m} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = -1, 2, \cdots, m.$$

Primal SVM vs. Dual Form

- Both are quadratic convex programming problems, solvable to global optimum via standard optimization software.
- Each has pros and cons; choice depends on efficiency and effectiveness.
- **Primal form suits**:
  - Low-dimensional data;
  - High-dimensional sparse data (mostly zero attributes).
- **Dual form suits**:
  - High-dimensional dense data, and easily integrates kernels for non-linear classification.

Nonseparable Data

- Overlapping classes can make hard-margin constraints infeasible.
- Noise and outliers can force an unhelpful boundary.
- Allow controlled margin violations with slack variables.

Soft-margin SVM balances a wide margin against violations.

Soft-Margin SVM

$$\begin{aligned} \min \quad & \frac{1}{2}||w||_2^2 + C\sum_i \xi_i \\ s.t. \quad & y_i(w^\top x_i + b) \geq 1 - \xi_i, \xi \geq 0 \end{aligned}$$

- The norm term encourages a wider margin.
- The slack term penalizes margin violations.
- C > 0 controls the penalty for each violation.

Optimal Slack and Sample Position

$$y_i f(x_i) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

Slack	Geometric interpretation
$\xi_i = 0$	On or outside the margin
$0 < \xi_i < 1$	Correct class, but inside the margin
$\xi_i = 1$	On the decision boundary
$\xi_i > 1$	Misclassified

Optimal Slack: $\xi_i = \max(0, 1 - y_i f(x_i))$

Loss Function for Soft-Margin SVM

$$\min \quad \frac{1}{2}||w||_2^2 + C\sum_i \ell_{0/1}(y_i(w^\top x_i + b) - 1)$$

Margin

Penalty

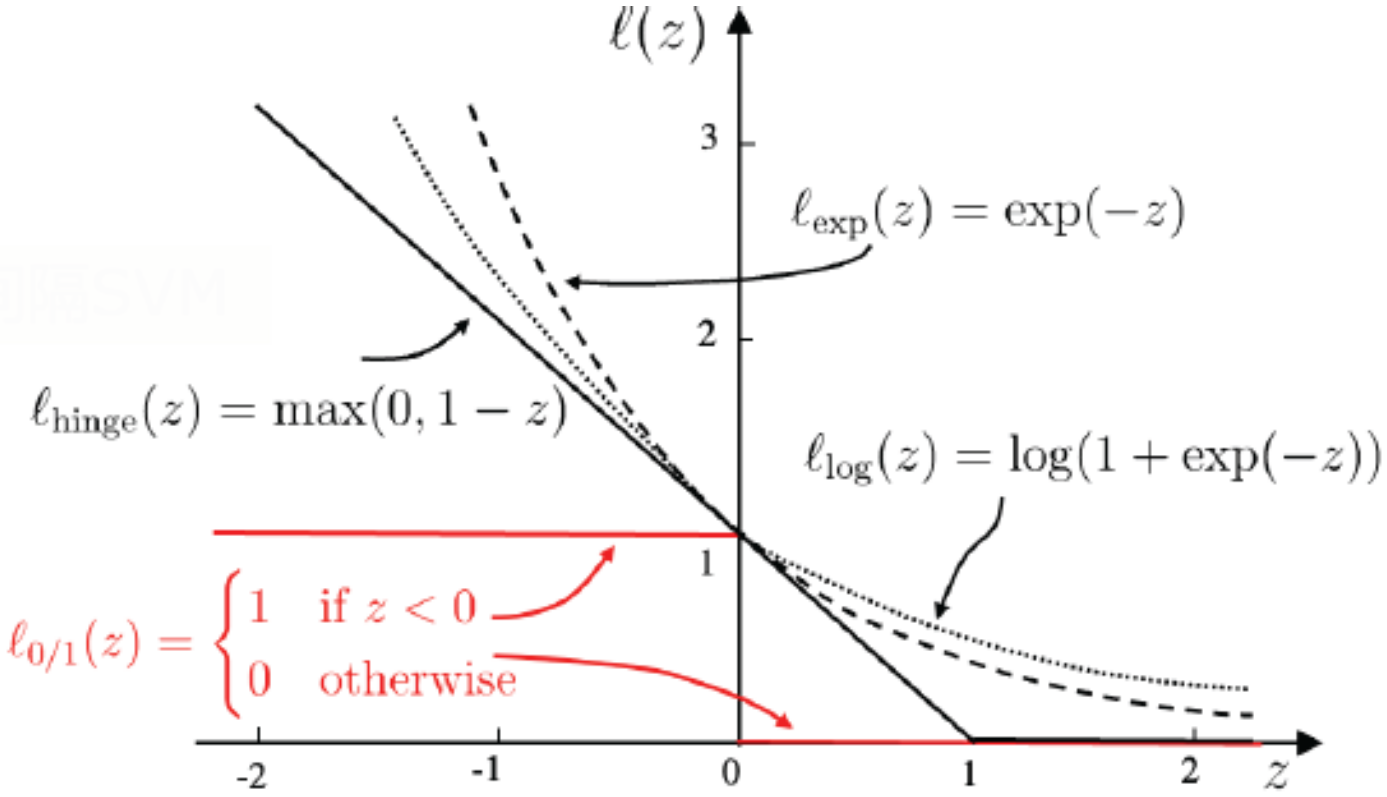
$$\ell_{0/1}(z) = \begin{cases} 1, & z < 0, \\ 0, & otherwise. \end{cases}$$

- Optimization is difficult:
  - Non-convex
  - Non-continuous
- Use the hinge loss as the replacement

Hinge Loss

- Hinge loss:

$$\ell(x) = \max(0, 1 - x)$$



Replacement loss functions have good mathematical properties and are generally the upper bounds of the 0/1 loss function.

Soft-Margin SVM as Regularized Risk Minimization

$$\min \quad \tfrac{1}{2}||w||_2^2 + C\sum_i \max(0, 1 - y_i f(x_i))$$

- The regularizer controls the weight norm.
- Hinge loss penalizes failure to achieve the unit margin.
- Eliminating optimal slack yields this equivalent objective.

Understanding the Regularization Parameter

Small C	Large C
Lower penalty for margin violations	Higher penalty for margin violations
Stronger regularization relative to loss	Stronger pressure to fit training samples

- Select C using validation data.
- The best boundary depends on the dataset.

A larger C does not always improve test accuracy.

Support Vectors

- Support vectors have nonzero dual coefficients α_i
- They can lie on the margin, inside it, or on the wrong side.
- Samples far outside the margin usually have $\alpha_i = 0$

$$\alpha_i > 0 \;\Rightarrow\; \text{sample } i \text{ contributes to } f(x)$$

Only support vectors contribute directly to the decision function.

Dual Problem of Soft-Margin SVM

- **Primal problem:**

$$\min \quad \frac{1}{2}||w||_2^2 + C\sum_i \max(0, 1 - y_i f(x_i))$$

- **Dual problem:**

$$\max_{\alpha} \quad \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^{\top} x_j$$

$$s.t. \quad \sum_{i=1}^{m}\alpha_i y_i = 0, \quad 0 \le \alpha_i \le C, \quad i = -1, 2, \cdots, m.$$

It is still a quadratic convex programming problem. According to convex optimization theory, mature numerical optimization software packages can be used to obtain the global optimal solution.

SVM Extension - Regularization

- A more general form of statistical learning models

$$\min_f \quad \Omega(f) + C\sum_{i=1}^{m} \ell(f(x_i), y_i)$$

Structural risk
Describes certain properties of the model itself.

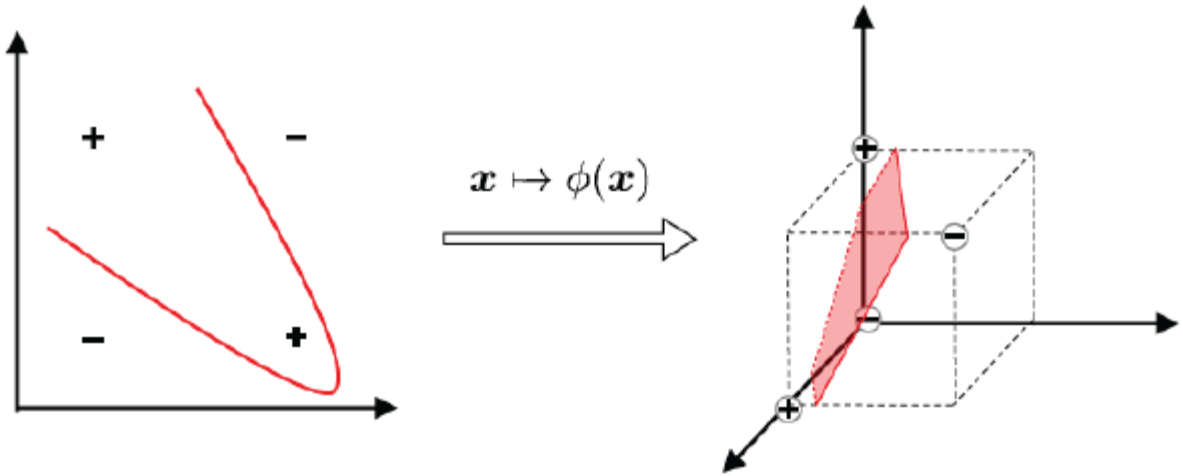
Empirical risk
Describes the degree of fit between the model and the training data.

- Regularization can be understood as a "penalty function method". By imposing penalties on undesired results, the optimization process tends towards the desired goal.
- From the perspective of Bayesian estimation, it can be considered as providing the prior probability of the model.

Kernel Method

Feature Space Mapping

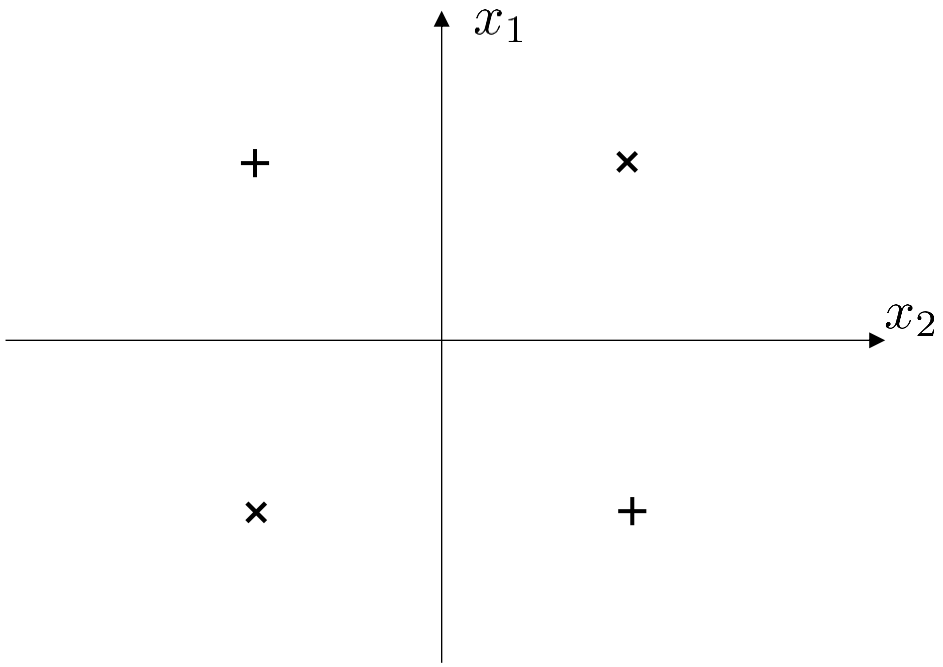
- The Case of Linearly Inseparable Data:
- What if there is no hyperplane that can correctly separate the two classes of samples?
- Map the samples from the original space to a higher-dimensional feature space, so that the samples are linearly separable in this feature space.



If the original space is finite-dimensional, then there must exist a higher-dimensional feature space that makes the samples separable.

Example: XOR

- XOR labels depend on the interaction between x_1 and x_2.



No single straight line separates the two classes.

A Simple Idea: Transform the Feature Space

$$\phi(\boldsymbol{x}) = x_1 x_2$$

Input x	Label y	New feature $\phi(x)$
(1, 1), (−1, −1)	+1	+1
(1, −1), (−1, 1)	−1	−1

$f(\boldsymbol{x}) = \phi(\boldsymbol{x})$ **gives the correct XOR signs**

An Explicit Feature Mapping Example

$$\psi(x) = [x_1^2, \sqrt{x_1 x_2}, x_2^2]^\top$$

$$\psi(x)^\top \psi(z) = (x^\top z)^2$$

- The three features represent quadratic terms.
- A linear function of $\psi(x)$ is quadratic in the input.
- The factor $\sqrt{2}$ matches the cross-term coefficient.

Kernel Trick

$$K(x, z) = \psi(x)^\top \psi(z)$$

Explicit mapping	Kernel evaluation
Construct $\psi(x)$ and $\psi(z)$	Evaluate $K(x, z)$ directly
Take their inner product	Obtain the same inner product

Replace $x^\top z$ in the dual with $K(x, z)$.

Review the SVM and the Dual Problem of SVM

- SVM problem:

$$\min_{w,b} \quad \frac{1}{2}||w||_2^2$$
$$s.t. \quad y_i(w^\top x_i + b) \geq 1, i = 1, 2, \cdots, m$$

- The dual problem of SVM:

$$\max_{\alpha} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^\top x_j$$
$$s.t. \quad \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, \cdots, m$$

- Prediction: $f(x) = w^\top x + b = \sum_{i=1}^{m} \alpha_i y_i x_i^\top x + b$

In Feature Space

- Introduce $\phi(x)$, the hyperplane is defined as: $f(x) = w^\top \phi(x) + b$
- Dual problem of SVM:

$$\max_{\alpha} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^\top x_j$$

$$s.t. \quad \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, \cdots, m$$



$$\max_{\alpha} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(x_i)^\top \phi(x_j)$$

$$s.t. \quad \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, \cdots, m$$

In Feature Space

- Prediction:

$$f(x) = w^\top x + b = \sum_{i=1}^{m} \alpha_i y_i x_i^\top x + b$$



$$f(x) = w^\top x + b = \sum_{i=1}^{m} \alpha_i y_i \phi(x_i)^\top \phi(x) + b$$

Appears only in inner product form

Thanks