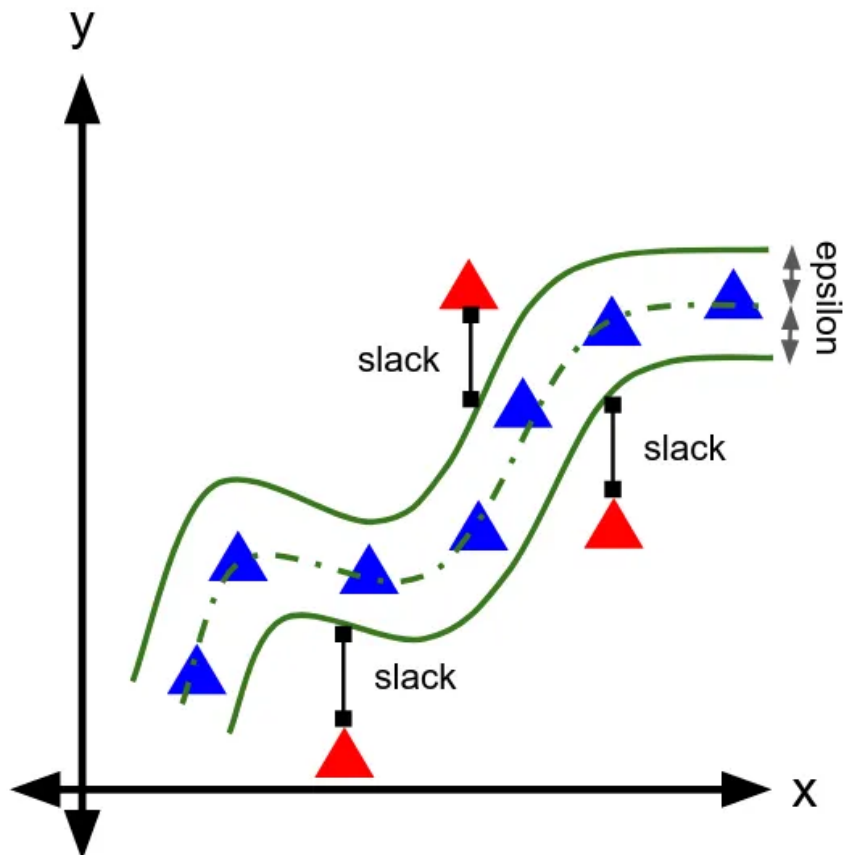


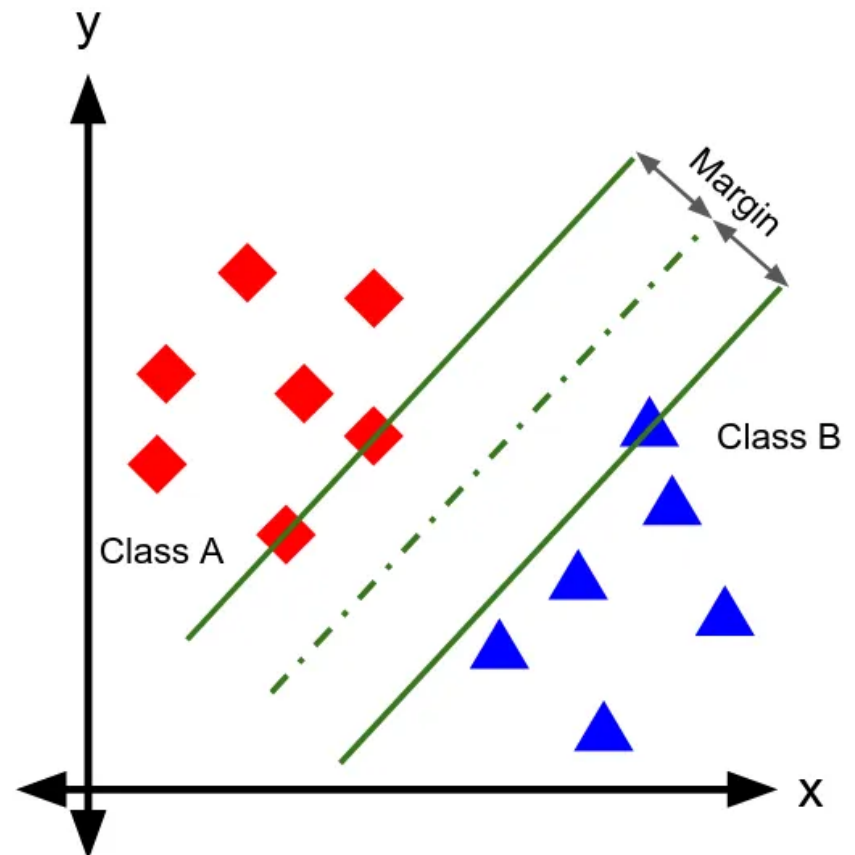
Machine Learning

SVM Regression

Jian Liu



Regression



Classification

Linear Regression

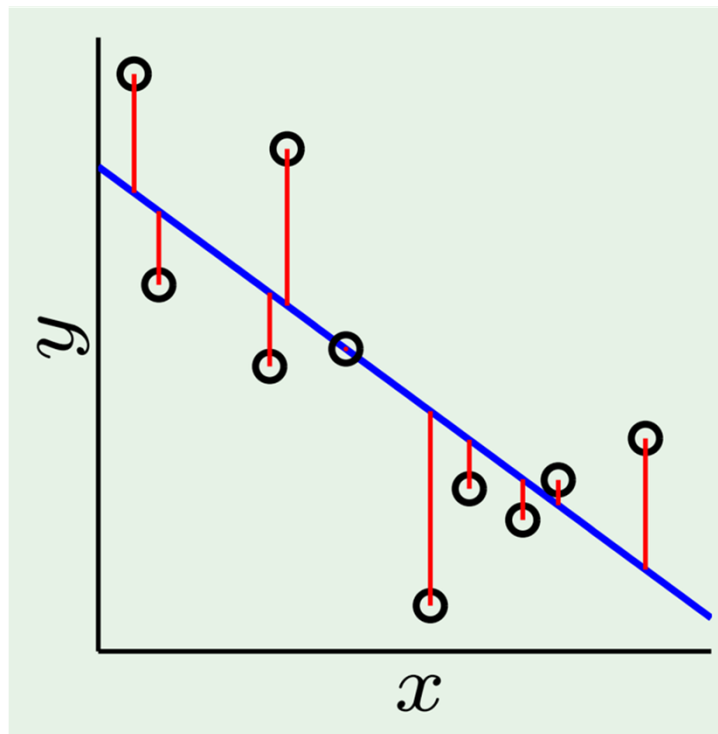
- Given data **X** and target **Y**
- The objective: Find a function that returns the best fit.
- Assume that the relationship between **X** and **Y** is approximately linear. The model can be represented as (**W** represents coefficients and **b** is an intercept)

$$f(w_1, \dots, w_n, b) = y = \mathbf{w} \cdot \mathbf{x} + b + \varepsilon$$

Linear Regression

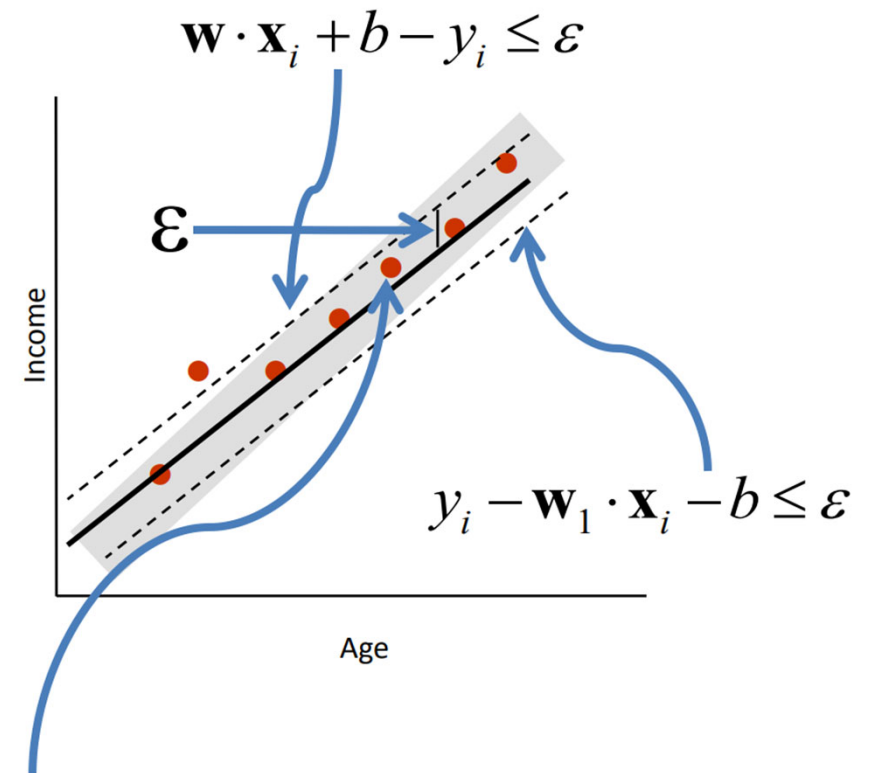
- To find the best fit, we minimize the sum of squared errors
-> Least square estimation

$$\min \sum_{i=1}^m (y_i - \hat{y}_i)^2 = \sum_{i=1}^m (y_i - (\mathbf{w} \cdot \mathbf{x}_i + b))^2$$



Support Vector Regression

- Find a function, $f(x)$,
with at most ε -deviation from the target y



We do not care about errors as long as they are less than ε

Support Vector Regression

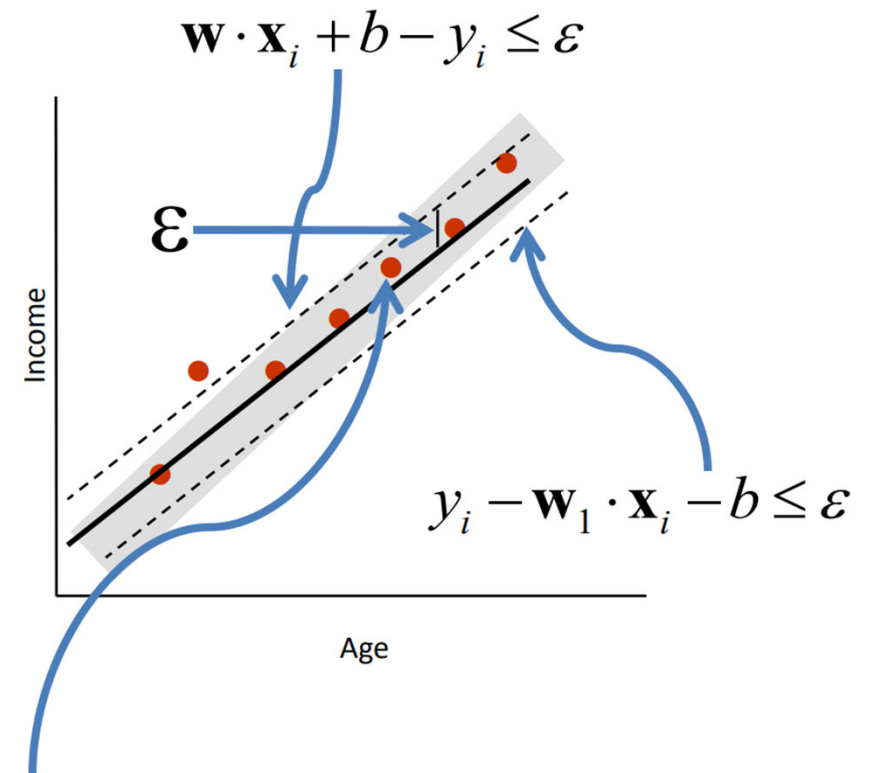
- Find a function, $f(x)$,
with at most ε -deviation from the target y

The problem can be written as
a convex optimization problem

$$\min \frac{1}{2} \|\mathbf{w}\|^2$$

$$s.t. \ y_i - \mathbf{w}_1 \cdot \mathbf{x}_i - b \leq \varepsilon;$$

$$\mathbf{w}_1 \cdot \mathbf{x}_i + b - y_i \leq \varepsilon;$$



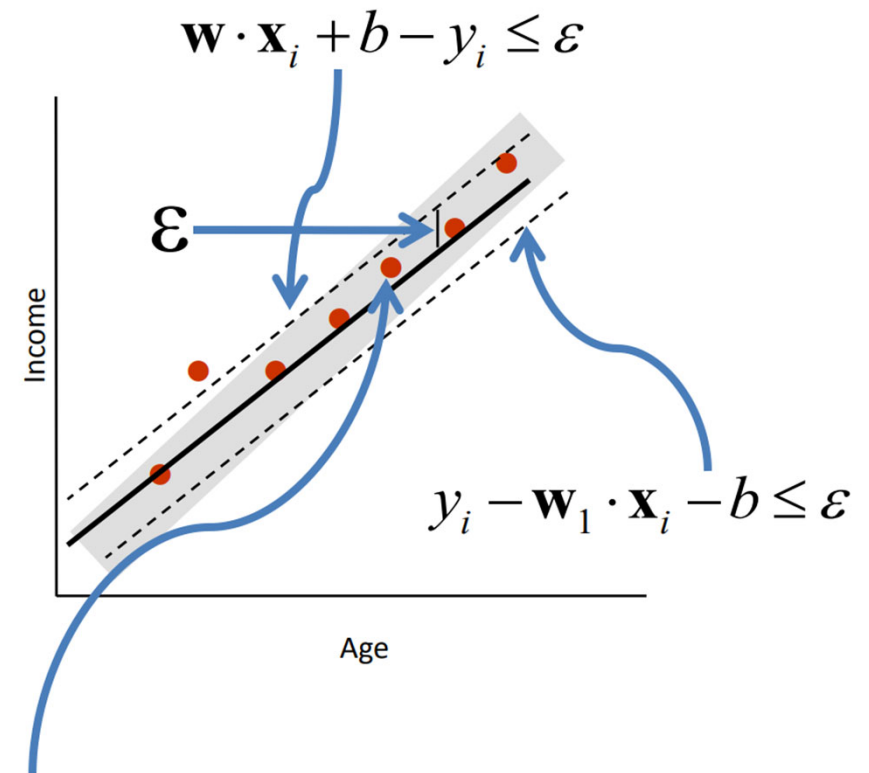
We do not care about errors as long as
they are less than ε

Support Vector Regression

- Find a function, $f(x)$,
with at most ε -deviation from the target y

What if the problem is not feasible?

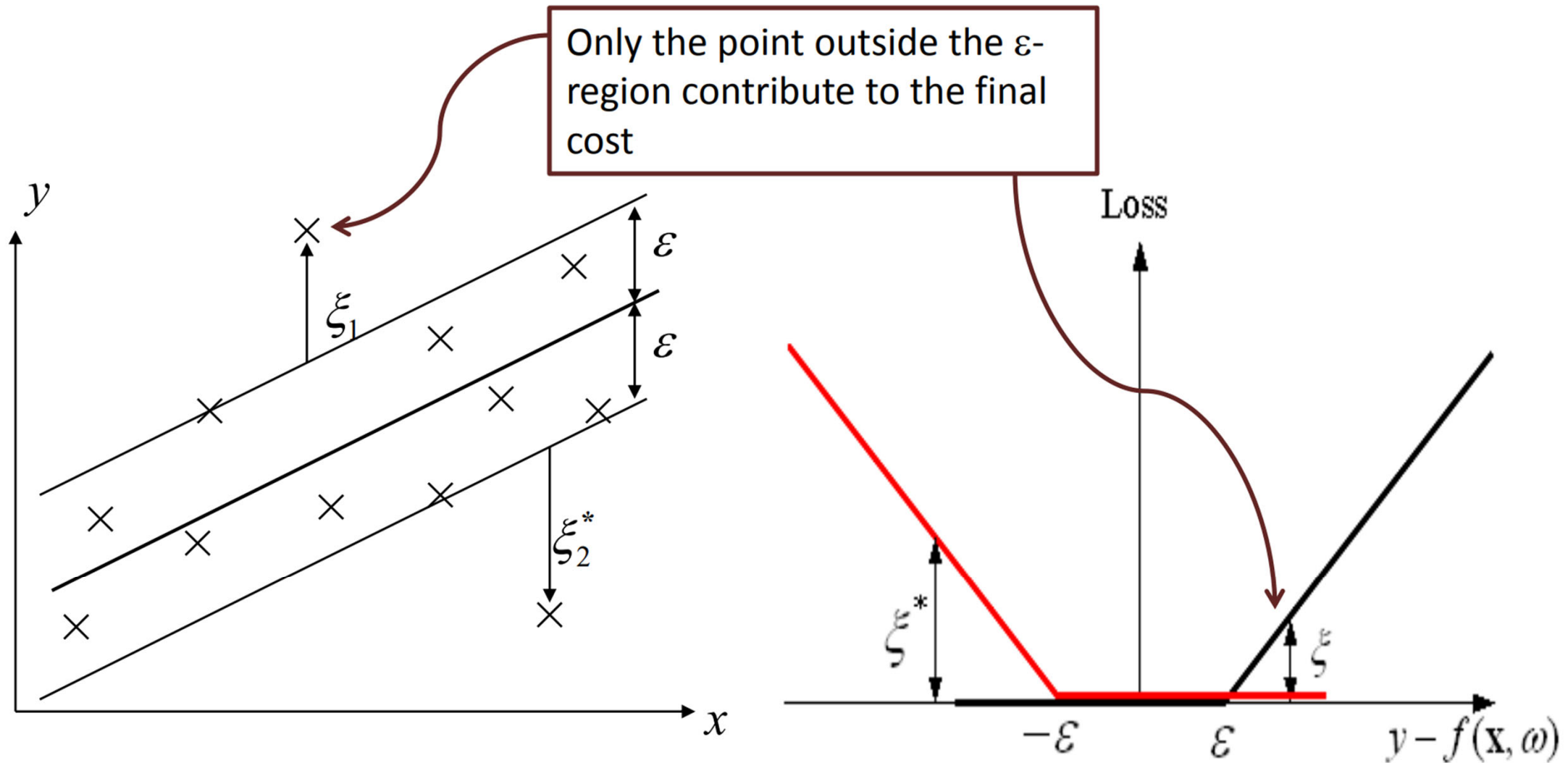
We can introduce slack variables
(similar to soft margin loss function).



We do not care about errors as long as
they are less than ε

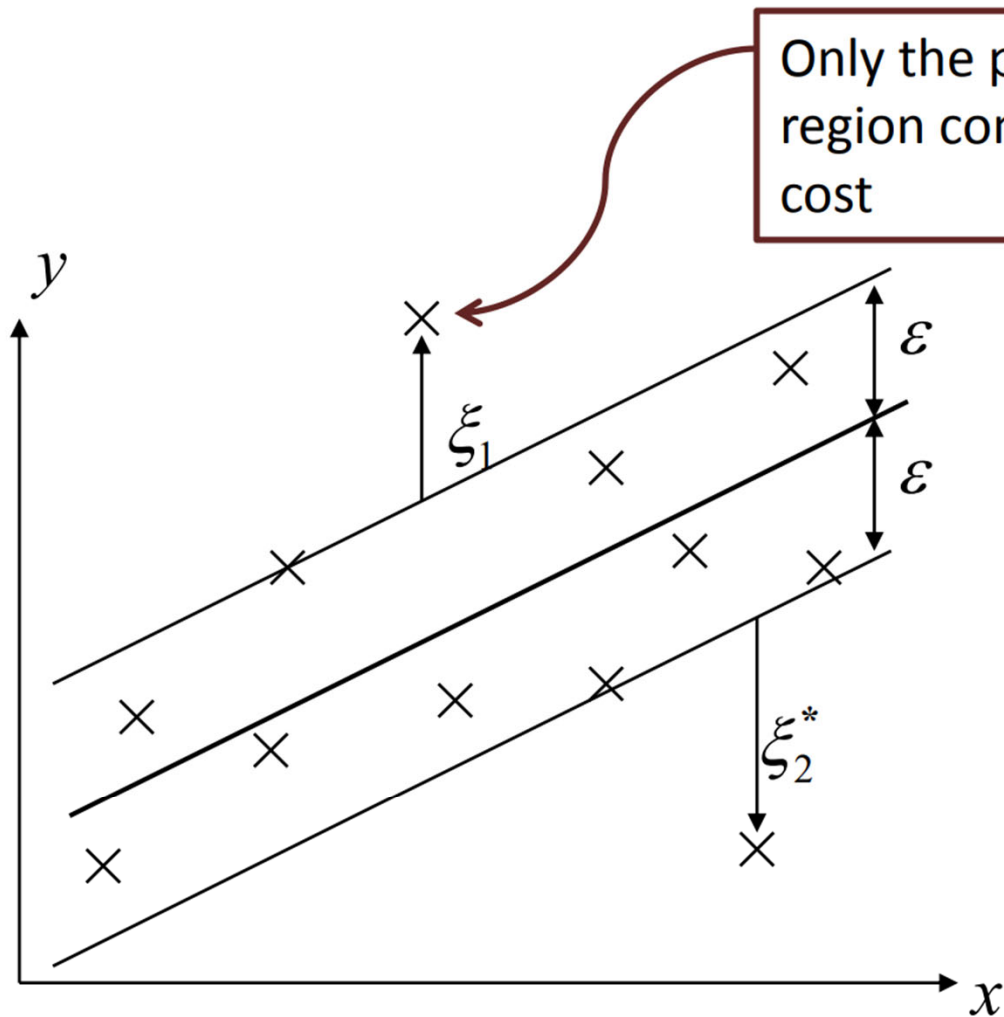
Support Vector Regression

We can introduce slack variables (similar to soft margin loss function).



Support Vector Regression

We can introduce slack variables (similar to soft margin loss function).



$$J(w) = \frac{1}{2}w'w + C \sum_1^N (\xi + \xi^*);$$

$$y_i - (x_i w + b) \leq \epsilon + \xi_i$$
$$(x_i w + b) - y_i \leq \epsilon + \xi_i^*$$

$$\xi^* \geq 0$$

$$\xi_i \geq 0$$

Support Vector Regression

We can introduce slack variables (similar to soft margin loss function).

Minimize:

- Minimize the sum of the slack variables ξ_i and ξ_i^* for all data points i .
- Minimize the squared norm of the weight vector w .

$$J(w) = \frac{1}{2}w'w + C \sum_{i=1}^N (\xi_i + \xi_i^*);$$

$$y_i - (x_i w + b) \leq \epsilon + \xi_i$$

$$(x_i w + b) - y_i \leq \epsilon + \xi_i^*$$

$$\xi_i^* \geq 0$$

$$\xi_i \geq 0$$

Optimizing the Lagrangian

$$\begin{aligned} L := & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) - \sum_{i=1}^{\ell} (\eta_i \xi_i + \eta_i^* \xi_i^*) \\ & - \sum_{i=1}^{\ell} \alpha_i (\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b) \\ & - \sum_{i=1}^{\ell} \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b) \end{aligned}$$

Lagrange multipliers $\alpha_i^{(*)}, \eta_i^{(*)} \geq 0$.

Optimizing the Lagrangian

The partial derivatives of L with respect to the variables

$$\partial_b L = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) = 0$$

$$\partial_w L = w - \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i = 0$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0$$

Optimizing the Lagrangian

$$\partial_b L = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) = 0$$

$$\partial_w L = w - \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i = 0$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0$$

$$\begin{aligned} L := & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) - \sum_{i=1}^{\ell} (\eta_i \xi_i + \eta_i^* \xi_i^*) \\ & - \sum_{i=1}^{\ell} \alpha_i (\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b) \\ & - \sum_{i=1}^{\ell} \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b) \end{aligned}$$

Optimizing the Lagrangian

$$\partial_b L = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) = 0$$

$$\partial_w L = w - \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i = 0$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0$$

$$L := \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) - \sum_{i=1}^{\ell} (\eta_i \xi_i + \eta_i^* \xi_i^*)$$

$$- \sum_{i=1}^{\ell} \alpha_i (\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b)$$

$$- \sum_{i=1}^{\ell} \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b)$$

Optimizing the Lagrangian

$$\partial_b L = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) = 0$$

$$\partial_w L = w - \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i = 0$$

$$\partial_{\xi_i^{(*)}} L = \boxed{C - \alpha_i^{(*)}} - \eta_i^{(*)} = 0$$

$$L := \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) - \sum_{i=1}^{\ell} (\boxed{\eta_i} \xi_i + \eta_i^* \xi_i^*)$$

$$- \sum_{i=1}^{\ell} \alpha_i (\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b)$$

$$- \sum_{i=1}^{\ell} \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b)$$

Optimizing the Lagrangian

$$L := \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) - \sum_{i=1}^{\ell} (C - \alpha_i^{(*)}) (\eta_i \xi_i + \eta_i^* \xi_i^*) \\ - \sum_{i=1}^{\ell} \alpha_i (\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b) \\ - \sum_{i=1}^{\ell} \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b)$$

Optimizing the Lagrangian

$$\begin{aligned}
 L := & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) - \sum_{i=1}^{\ell} (\eta_i \xi_i + \eta_i^* \xi_i^*) \\
 & - \sum_{i=1}^{\ell} \alpha_i (\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b) \\
 & - \sum_{i=1}^{\ell} \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b)
 \end{aligned}$$

The diagram shows the Lagrangian function L with several terms highlighted in blue boxes and crossed out with red X's. The blue boxes are:

- $C - \alpha_i^{(*)}$ (above the second sum)
- η_i (in the third sum)
- ξ_i (in the second sum)
- ξ_i^* (in the second sum)
- ξ_i (in the fourth sum)
- ξ_i^* (in the fourth sum)

 The red X's are placed over the terms ξ_i and ξ_i^* in the second and fourth sums, and over the terms ξ_i and ξ_i^* in the fourth and second sums respectively.

Optimizing the Lagrangian

$$\partial_b L = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) = 0$$

$$\partial_w L = w - \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i = 0$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0$$

$$\begin{aligned}
 L := & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} (\cancel{\xi_i} + \cancel{\xi_i^*}) - \sum_{i=1}^{\ell} (\eta_i \cancel{\xi_i} + \eta_i^* \cancel{\xi_i^*}) \\
 & - \sum_{i=1}^{\ell} \alpha_i (\varepsilon + \cancel{\xi_i} - y_i + \langle w, x_i \rangle + b) \\
 & - \sum_{i=1}^{\ell} \alpha_i^* (\varepsilon + \cancel{\xi_i^*} + y_i - \langle w, x_i \rangle - b)
 \end{aligned}$$

Optimizing the Lagrangian

$$\partial_b L = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) = 0$$

$$\partial_w L = w - \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i = 0$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0$$

$$\begin{aligned}
 L := & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} (\cancel{\xi_i} + \cancel{\xi_i^*}) - \sum_{i=1}^{\ell} (\eta_i \cancel{\xi_i} + \eta_i^* \cancel{\xi_i^*}) \\
 & - \sum_{i=1}^{\ell} \alpha_i (\varepsilon + \cancel{\xi_i} - y_i + \langle w, x_i \rangle + b) \\
 & - \sum_{i=1}^{\ell} \alpha_i^* (\varepsilon + \cancel{\xi_i^*} + y_i - \langle w, x_i \rangle - b)
 \end{aligned}$$

Optimizing the Lagrangian

$$\partial_b L = \sum_{i=1}^{\ell} (\alpha_i^* - \alpha_i) = 0$$

$$\partial_w L = w - \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i = 0$$

$$\partial_{\xi_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0$$

$$L := \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} (\cancel{\xi_i} + \cancel{\xi_i^*}) - \sum_{i=1}^{\ell} (\eta_i \cancel{\xi_i} + \eta_i^* \cancel{\xi_i^*})$$

$$- \sum_{i=1}^{\ell} \alpha_i (\varepsilon + \cancel{\xi_i} - y_i + \langle w, x_i \rangle + b)$$

$$- \sum_{i=1}^{\ell} \alpha_i^* (\varepsilon + \cancel{\xi_i^*} + y_i - \langle w, x_i \rangle - b)$$

Optimizing the Lagrangian

$$\partial_w L = w - \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i = 0$$

$$L := \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) - \sum_{i=1}^{\ell} (\eta_i \xi_i + \eta_i^* \xi_i^*)$$

$$- \sum_{i=1}^{\ell} \alpha_i (\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b)$$

$$- \sum_{i=1}^{\ell} \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b)$$

Optimizing the Lagrangian

$$\partial_w L = w - \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i = 0 \quad \rightarrow \quad w = \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i$$

$$\begin{aligned}
 L := & \boxed{\frac{1}{2} \|w\|^2} + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) - \sum_{i=1}^{\ell} (\eta_i \xi_i + \eta_i^* \xi_i^*) \\
 & - \sum_{i=1}^{\ell} \alpha_i (\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b) \\
 & - \sum_{i=1}^{\ell} \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b)
 \end{aligned}$$

Optimizing the Lagrangian

$$\partial_w L = w - \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i = 0 \quad \rightarrow \quad w = \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i$$

$$\begin{aligned}
 L := & \boxed{\frac{1}{2} \|w\|^2} + C \sum_{i=1}^{\ell} (\cancel{\xi_i} + \cancel{\xi_i^*}) - \sum_{i=1}^{\ell} (\cancel{\eta_i \xi_i} + \cancel{\eta_i^* \xi_i^*}) \\
 & - \sum_{i=1}^{\ell} \boxed{\alpha_i (\varepsilon + \cancel{\xi_i} - y_i + \langle w, x_i \rangle + b)} \\
 & - \sum_{i=1}^{\ell} \boxed{\alpha_i^* (\varepsilon + \cancel{\xi_i^*} + y_i - \langle w, x_i \rangle - b)}
 \end{aligned}$$

Optimizing the Lagrangian

$$\partial_w L = w - \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i = 0 \quad \rightarrow \quad w = \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) x_i$$

maximize

$$\begin{cases} -\frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle \\ -\varepsilon \sum_{i=1}^{\ell} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{\ell} y_i (\alpha_i - \alpha_i^*) \end{cases}$$

subject to

$$\sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i, \alpha_i^* \in [0, C]$$

$$\begin{aligned} L := & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} (\cancel{\xi_i} + \cancel{\xi_i^*}) - \sum_{i=1}^{\ell} (\cancel{\eta_i \xi_i} + \cancel{\eta_i^* \xi_i^*}) \\ & - \sum_{i=1}^{\ell} \left[\alpha_i (\varepsilon + \cancel{\xi_i} - y_i + \langle w, x_i \rangle + b) \right. \\ & \left. - \sum_{i=1}^{\ell} \left[\alpha_i^* (\varepsilon + \cancel{\xi_i^*} + y_i - \langle w, x_i \rangle - b) \right] \right] \end{aligned}$$

Optimizing the Lagrangian

maximize

$$L := \boxed{\frac{1}{2} \|w\|^2} + C \sum_{i=1}^{\ell} (\cancel{\xi_i} + \cancel{\xi_i^*}) - \sum_{i=1}^{\ell} (\cancel{\eta_i \xi_i} + \cancel{\eta_i^* \xi_i^*})$$

subject to

$$\sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i, \alpha_i^* \in [0, C]$$

$$- \sum_{i=1}^{\ell} \boxed{\alpha_i (\varepsilon + \cancel{\xi_i} - y_i + \langle w, x_i \rangle + b)} - \sum_{i=1}^{\ell} \boxed{\alpha_i^* (\varepsilon + \cancel{\xi_i^*} + y_i - \langle w, x_i \rangle - b)}$$

The diagram illustrates the optimization of the Lagrangian L . The Lagrangian is defined as the sum of the squared norm of w (highlighted in a pink box), a regularization term $C \sum (\xi_i + \xi_i^*)$, and a hinge loss term $-\sum (\eta_i \xi_i + \eta_i^* \xi_i^*)$. The constraints are given as $\sum (\alpha_i - \alpha_i^*) = 0$ and $\alpha_i, \alpha_i^* \in [0, C]$. The constraints are also written as $-\sum \alpha_i (\varepsilon + \xi_i - y_i + \langle w, x_i \rangle + b) - \sum \alpha_i^* (\varepsilon + \xi_i^* + y_i - \langle w, x_i \rangle - b)$. The terms ξ_i , ξ_i^* , $\eta_i \xi_i$, $\eta_i^* \xi_i^*$, and the entire constraint expression are crossed out with red X's, indicating they are not part of the final optimization problem.

Optimizing the Lagrangian

$$\begin{aligned} &\text{maximize} && \begin{cases} -\frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle \\ -\varepsilon \sum_{i=1}^{\ell} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{\ell} y_i (\alpha_i - \alpha_i^*) \end{cases} \\ &\text{subject to} && \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i, \alpha_i^* \in [0, C] \end{aligned}$$

Optimizing the Lagrangian

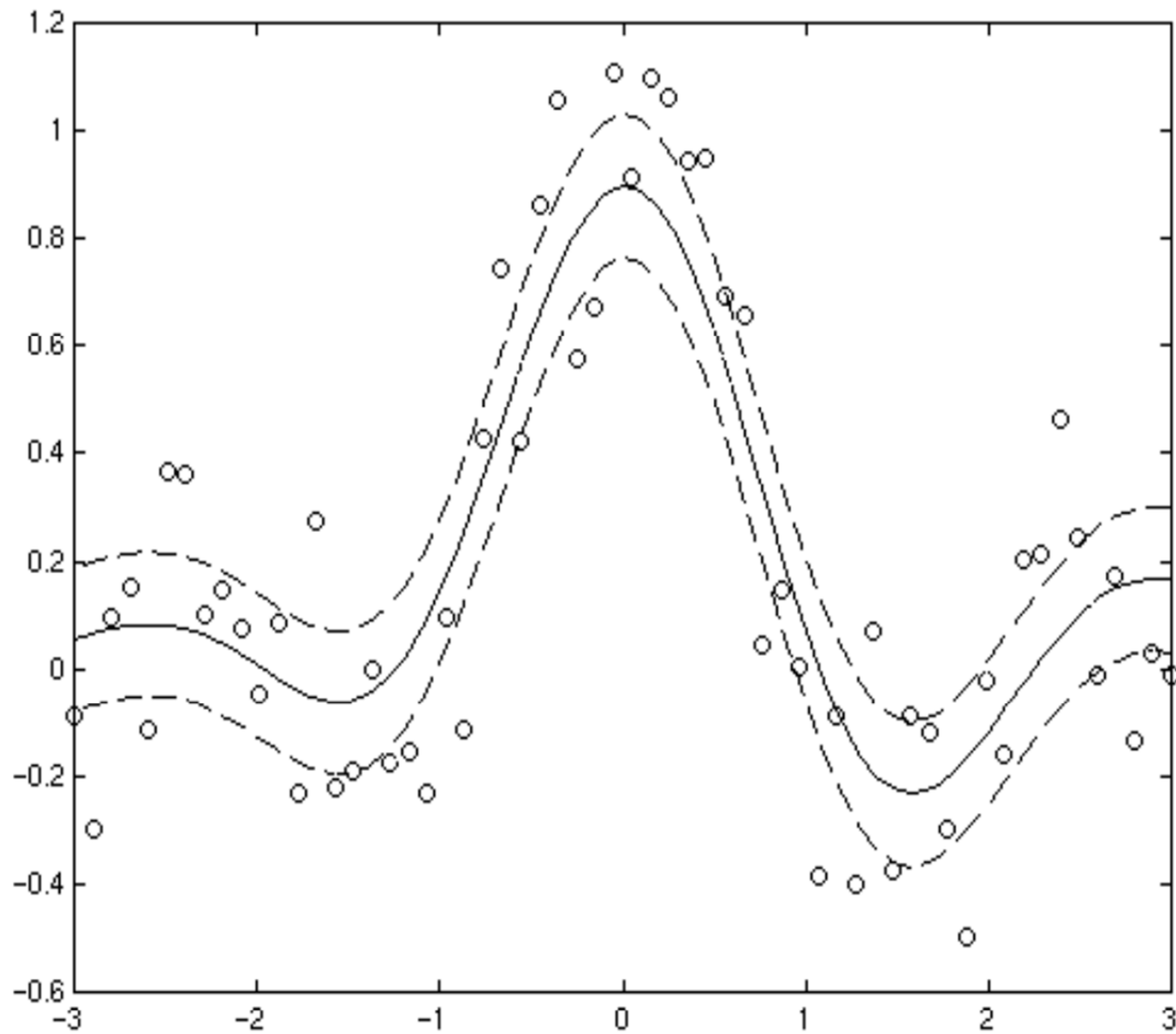


Dual optimization

$$\begin{aligned} &\text{maximize} \quad \begin{cases} -\frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle \\ -\varepsilon \sum_{i=1}^{\ell} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{\ell} y_i (\alpha_i - \alpha_i^*) \end{cases} \\ &\text{subject to} \quad \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i, \alpha_i^* \in [0, C] \end{aligned}$$

Now we can use the similar tricks as in SVM

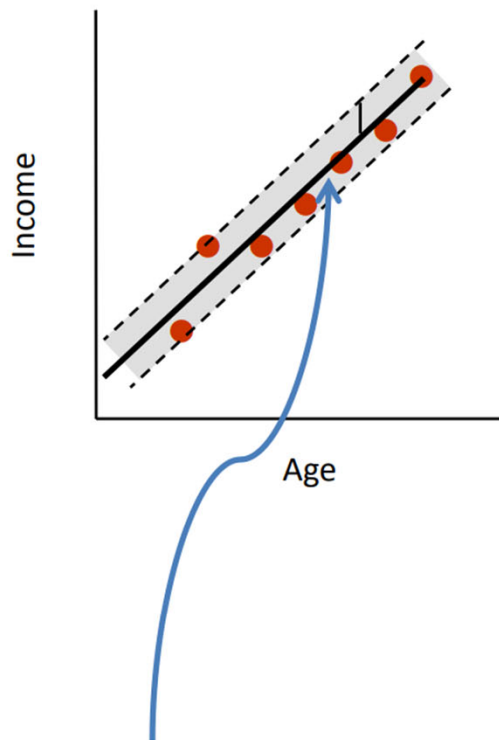
How about a non-linear case?



How about a non-linear case?

- Linear case

$$f : \text{age} \rightarrow \text{income}$$

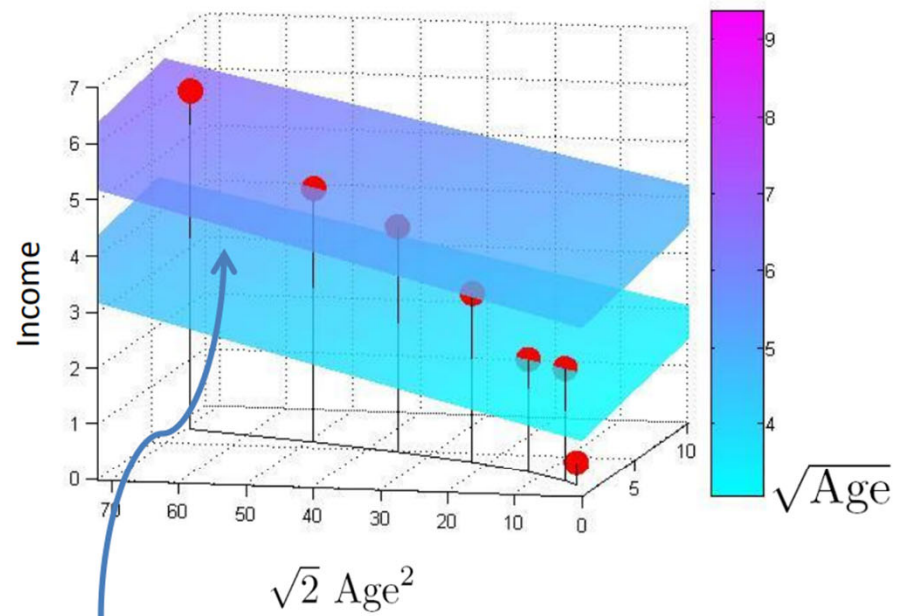


$$y_i = \mathbf{w}_1 \cdot \mathbf{x}_i + b$$

- Non-linear case

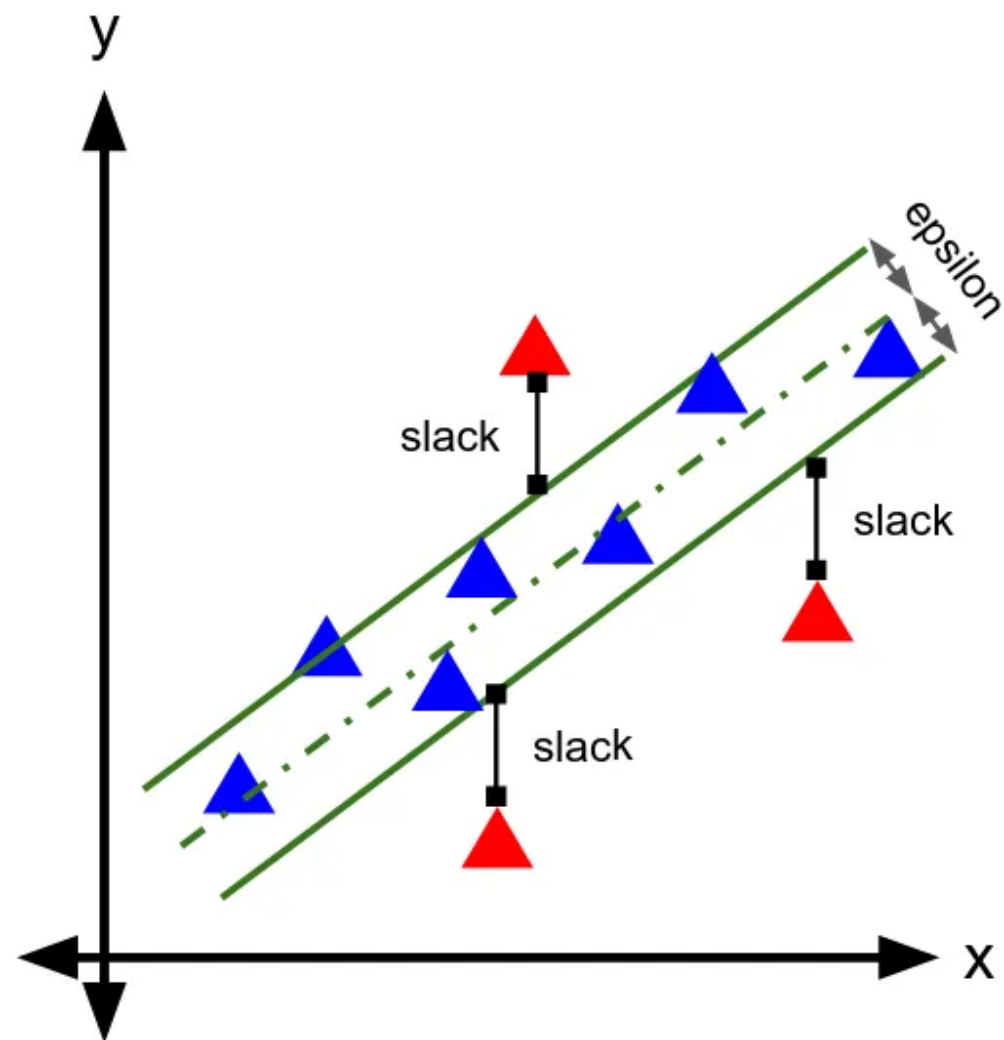
- Map data into a higher dimensional space, e.g.,

$$f : (\sqrt{\text{age}}, \sqrt{2\text{age}^2}) \rightarrow \text{income}$$

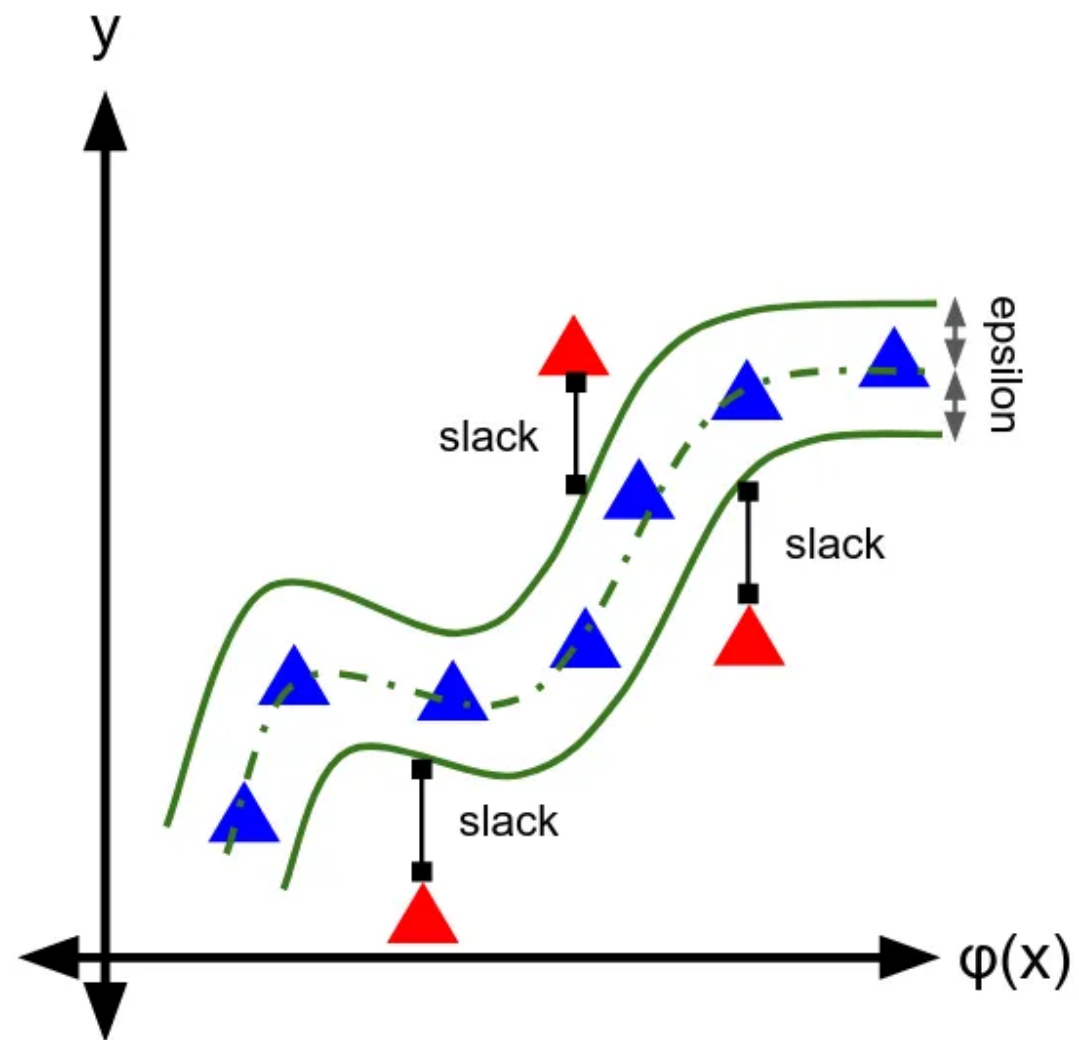


$$y_i = \mathbf{w}_1 \sqrt{\mathbf{x}_i} + \mathbf{w}_2 \sqrt{2\mathbf{x}_i^2} + b$$

Linear vs Non-linear



Linear



Non-linear

Dual problem

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*)$$

$$s.t. \begin{cases} y_i - (\mathbf{w} \cdot \mathbf{x}_i) - b \leq \varepsilon + \xi_i \\ (\mathbf{w} \cdot \mathbf{x}_i) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, i = 1, \dots, m \end{cases}$$

Primal variables: \mathbf{w} for each feature dim

Complexity: the dim of the input space

Primal

$$\max \begin{cases} \frac{1}{2} \sum_{i,j=1}^m (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle \\ -\varepsilon \sum_{i=1}^m (\alpha_i + \alpha_i^*) + \sum_{i=1}^m y_i (\alpha_i - \alpha_i^*) \end{cases}$$

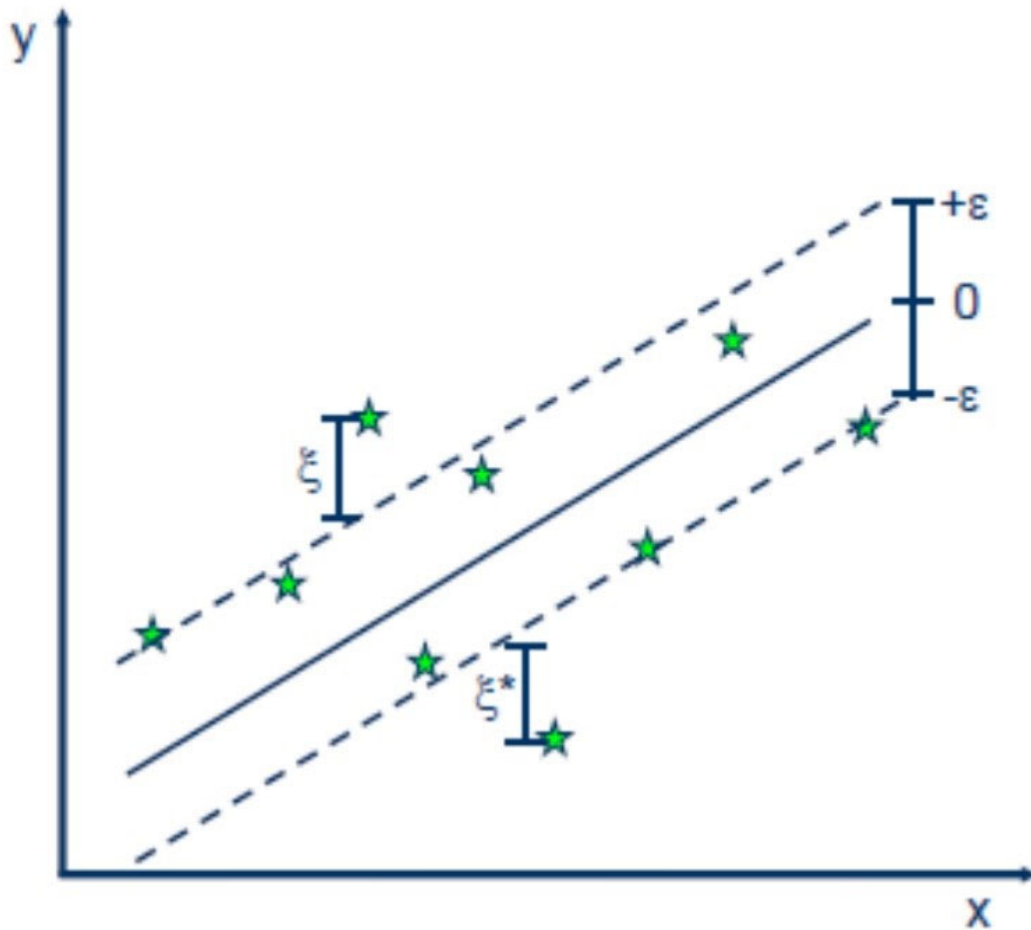
$$s.t. \sum_{i=1}^m (\alpha_i - \alpha_i^*) = 0; \quad 0 \leq \alpha_i, \alpha_i^* \leq C$$

Dual variables: α, α^* for each data point

Complexity: Number of support vectors

Dual

Dual problem



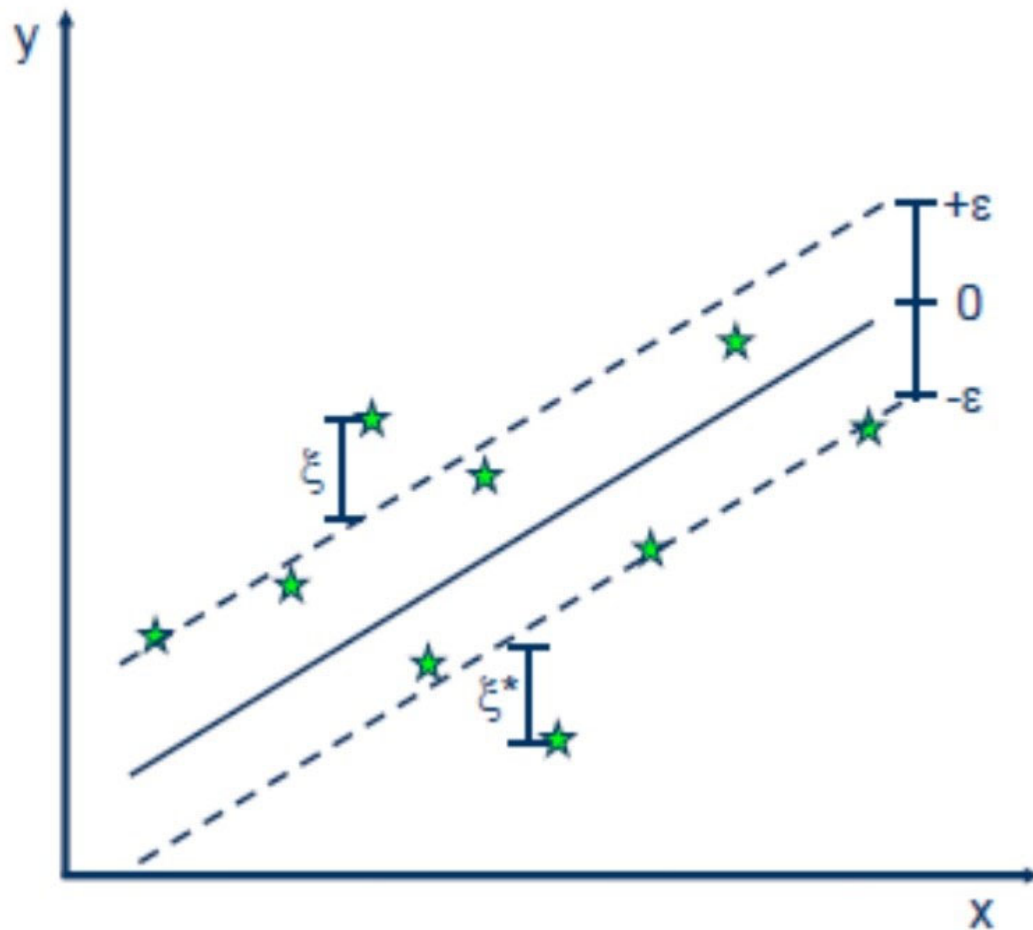
$$\begin{aligned} \max \quad & \begin{cases} \frac{1}{2} \sum_{i,j=1}^m (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle x_i, x_j \rangle \\ -\varepsilon \sum_{i=1}^m (\alpha_i + \alpha_i^*) + \sum_{i=1}^m y_i (\alpha_i - \alpha_i^*) \end{cases} \\ s.t. \quad & \sum_{i=1}^m (\alpha_i - \alpha_i^*) = 0; \quad 0 \leq \alpha_i, \alpha_i^* \leq C \end{aligned}$$

Dual variables: α, α^* for each data point

Complexity: Number of support vectors

Dual

Dual problem



Kernel trick

$$\max \begin{cases} \frac{1}{2} \sum_{i,j=1}^m (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle \varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j) \rangle \\ - \varepsilon \sum_{i=1}^m (\alpha_i + \alpha_i^*) + \sum_{i=1}^m y_i (\alpha_i - \alpha_i^*) \end{cases}$$

$K(\mathbf{x}_i, \mathbf{x}_j)$

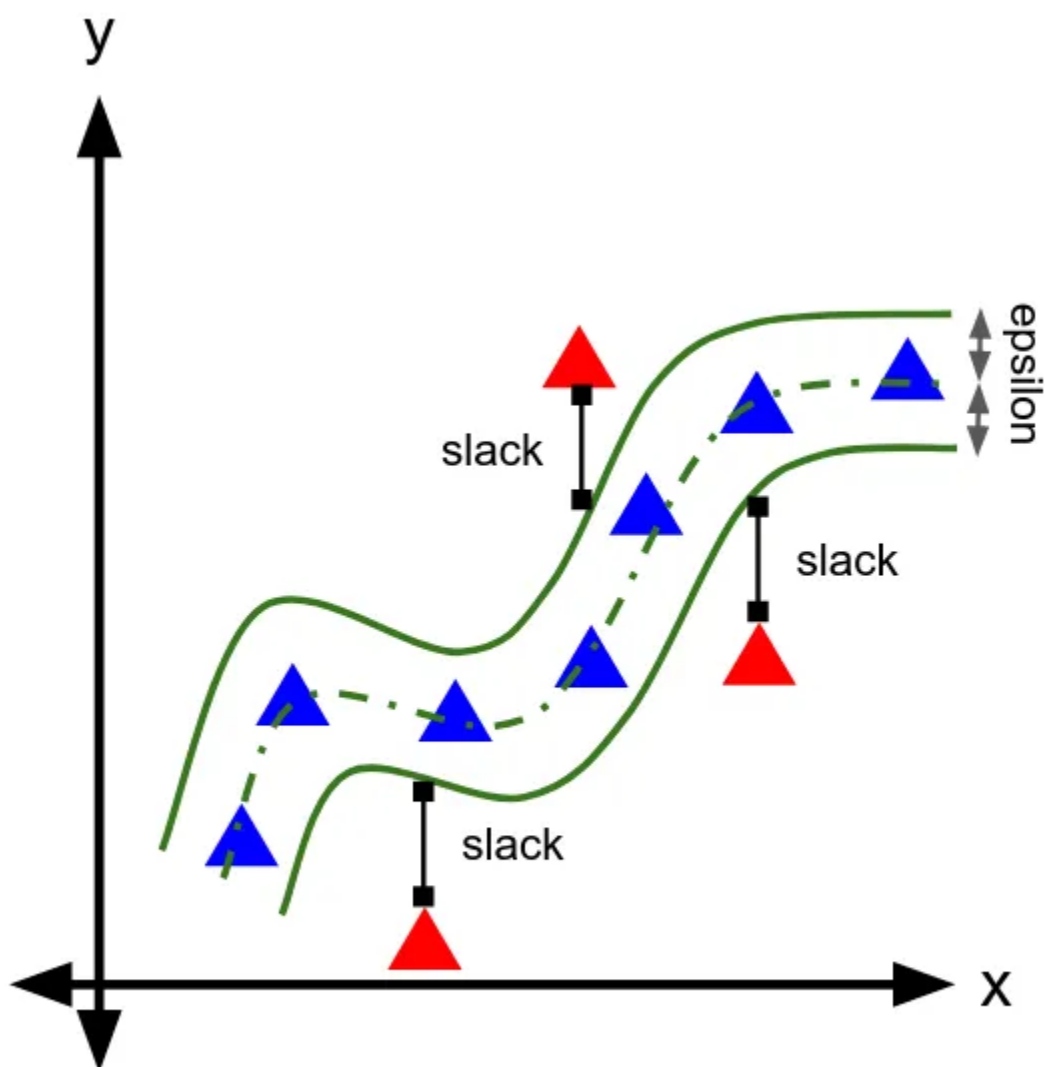
$$s.t. \sum_{i=1}^m (\alpha_i - \alpha_i^*) = 0; \quad 0 \leq \alpha_i, \alpha_i^* \leq C$$

Dual variables: α, α^* for each data point

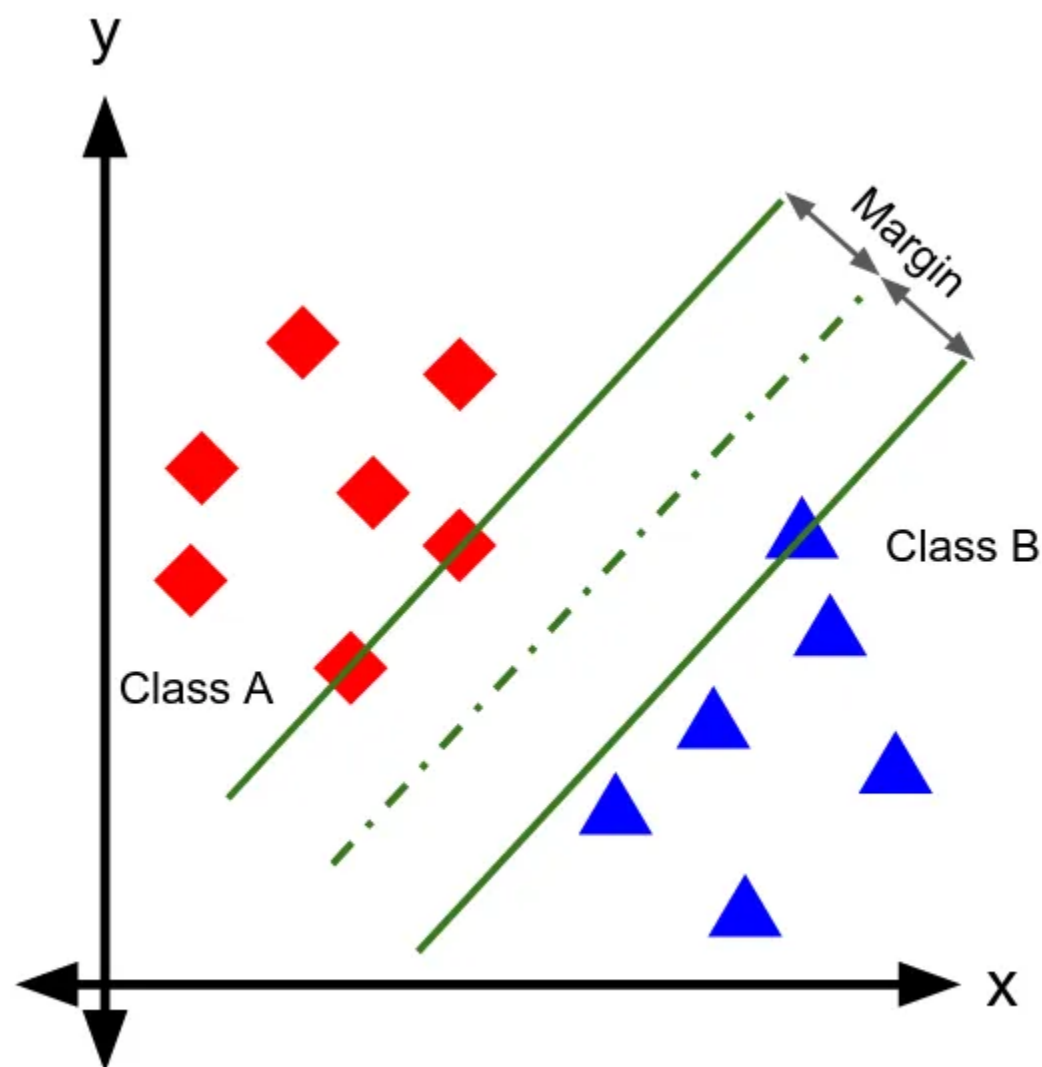
Complexity: Number of support vectors

Dual

SVM: Regression vs Classification



Regression



Classification

Summary

- Linear regression tries to minimize the error between the real and predicted value.
- SVR tries to fit the best line within a threshold value (a tube).
- The threshold value is the distance between the hyperplane and boundary line.
- Observations within the threshold of epsilon produce no error, only the observation outside of the epsilon range produce error – sparse kernel machines