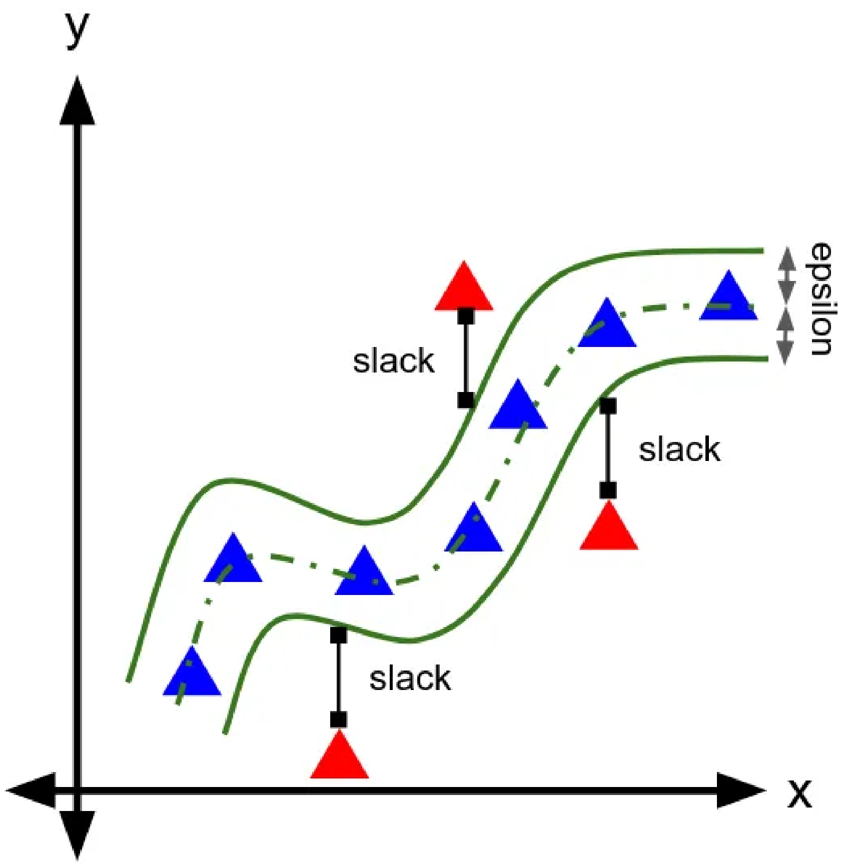
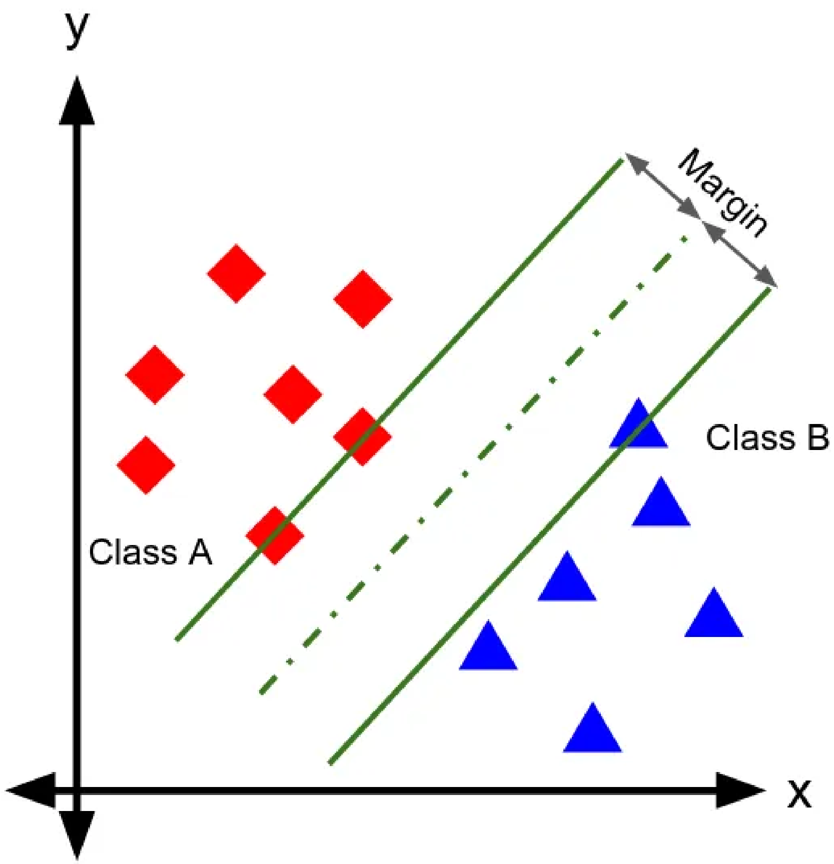
Machine Learning
SVM Regression

Jian Liu



Regression



Classification

Linear Regression

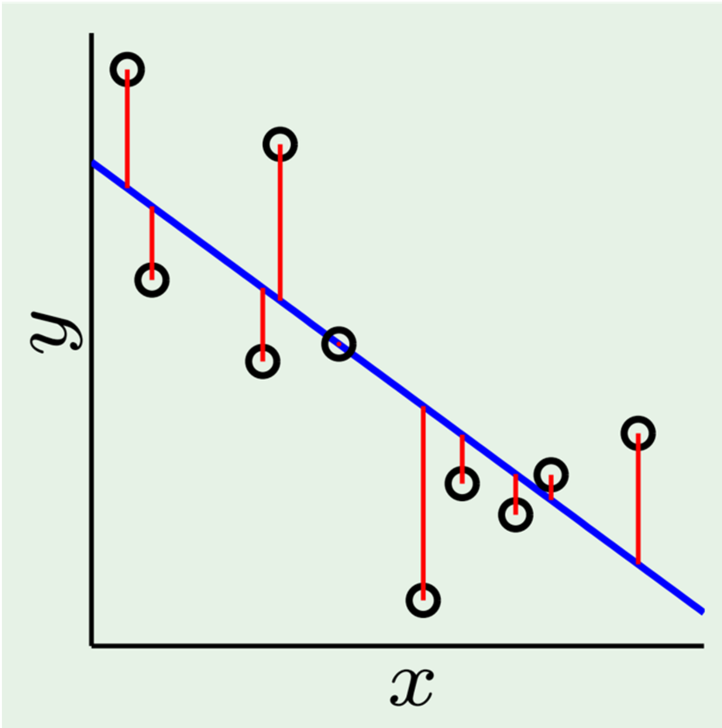
- Given data **X** and target **Y**
- The objective: Find a function that returns the best fit.
- Assume that the relationship between **X** and **Y** is approximately linear. The model can be represented as (**W** represents coefficients and **b** is an intercept)

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

Linear Regression

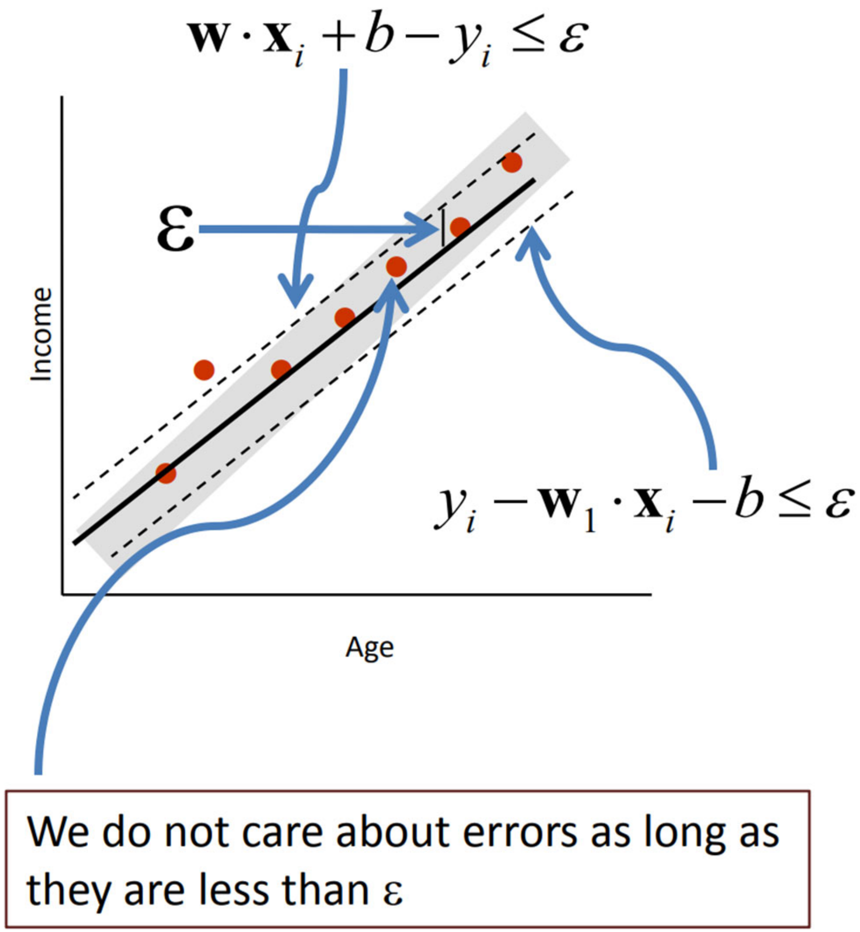
- To find the best fit, we minimize the sum of squared errors -> Least square estimation

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



Support Vector Regression

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



Support Vector Regression

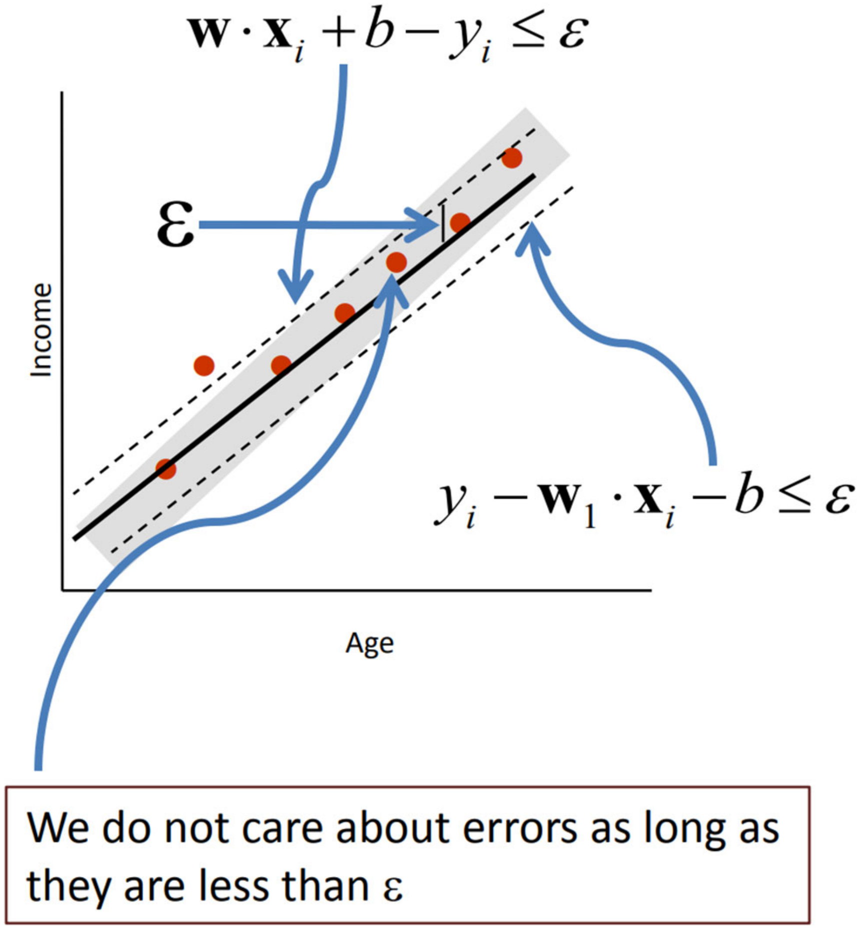
- Find a function, f(x), with at most ε-deviation from the target y

The problem can be written as a convex optimization problem

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

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

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

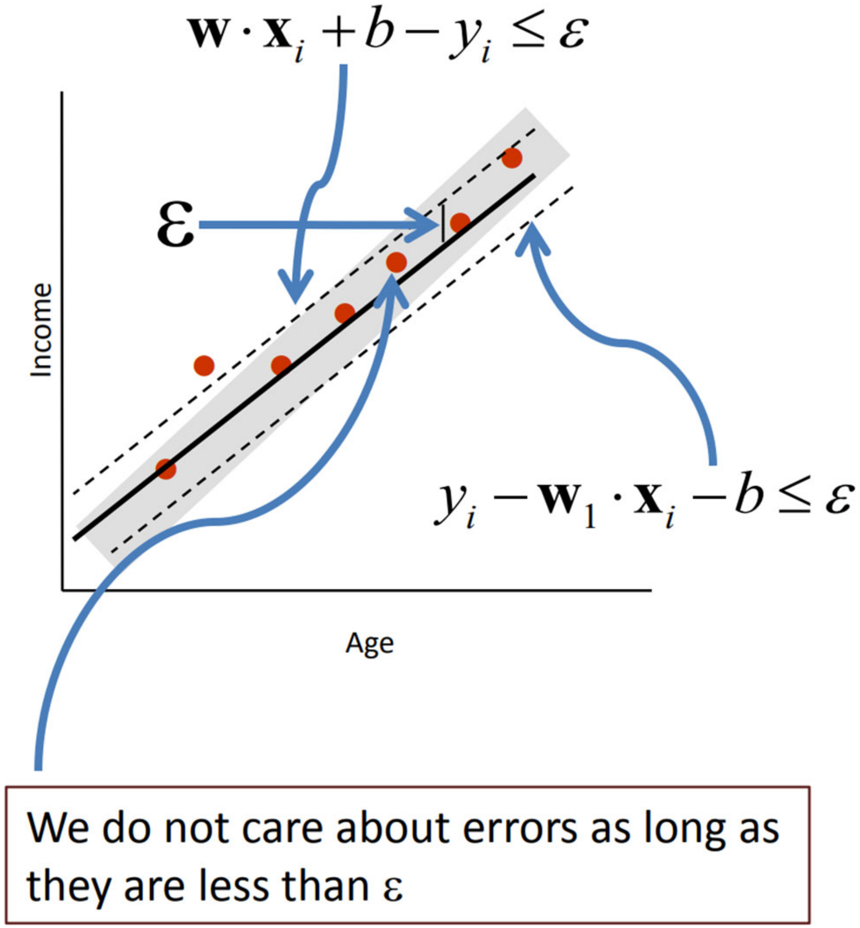


Support Vector Regression

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

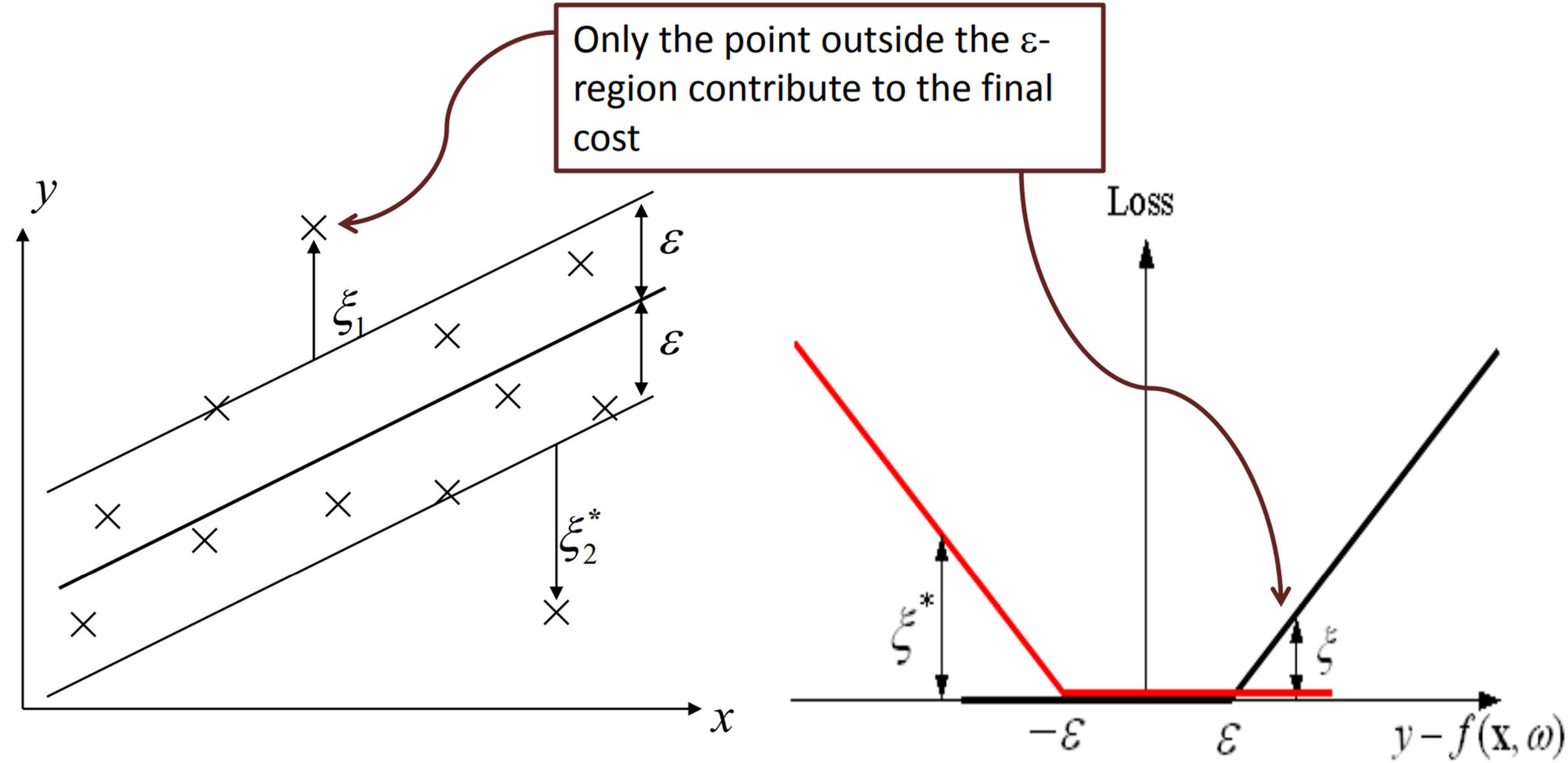
What if the problem is not feasible?

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



Support Vector Regression

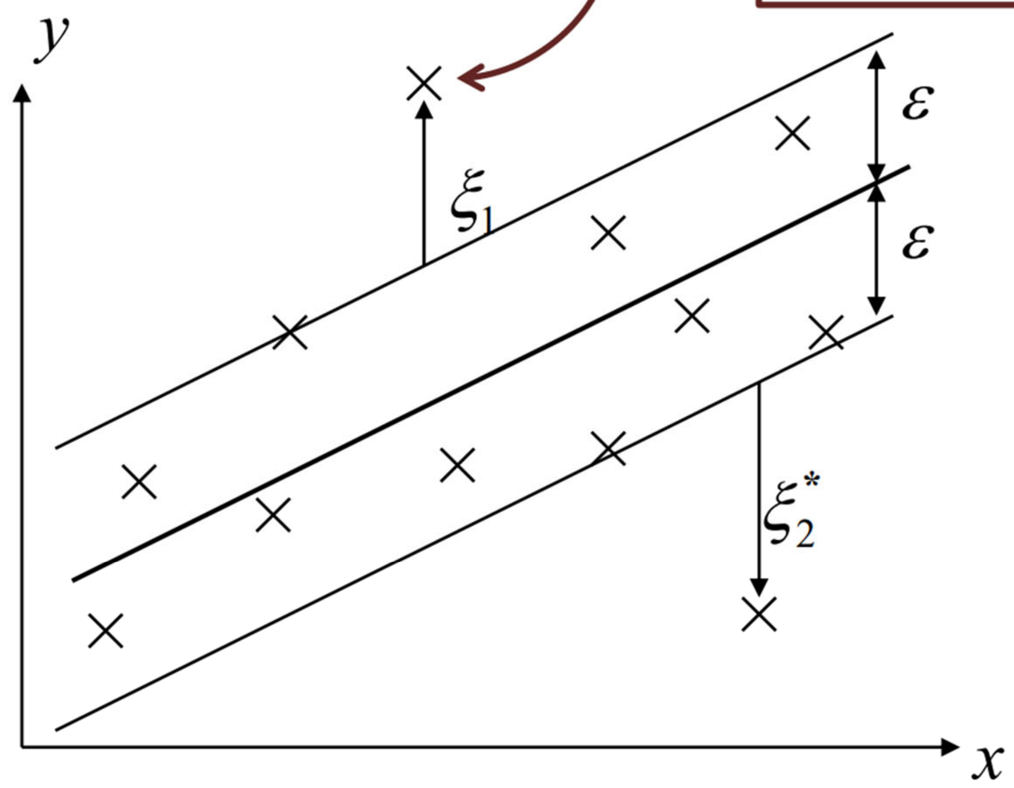
We can introduce slack variables (similar to soft margin loss function).



Support Vector Regression

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

Only the point outside the ε-region contribute to the final cost



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

$$y_i - (x_i w + b) \leq \epsilon + \xi_i$$
$$(x_i w + b) - y_i \leq \epsilon + \xi_i^*$$
$$\xi^* \geq 0$$
$$\xi_i \geq 0$$

Support Vector Regression

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

M~ujwufwfr jyjw C:

- Fx C nshwjfxjx1tzwytqjwfshj ktw utnsyx tzyxnij tk ϵ fqxt nshwjfxjx3
- Fx C fuuwtfhmjx 51ymj ytqjwfshj fuuwtfhmjx 5 fsi ymj jvzfynts htqqfuxjx nsyt ymj xnr uqnknji -fqymtzlm xtr jynr jx nskjfxngqj.tsj3

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

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

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

$$\xi^* \geq 0$$

$$\xi_i \geq 0$$

Optimizing the Lagrangian

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

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

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

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

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

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

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

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

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

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

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

$$C - \alpha_i^{(*)}.$$

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

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

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

Optimizing the Lagrangian

$$\boxed{C - \alpha_i^{(*)}}$$

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

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

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

Optimizing the Lagrangian

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

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

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

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

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

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

Optimizing the Lagrangian

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

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

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

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

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

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

Optimizing the Lagrangian

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

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

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

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

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

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

Optimizing the Lagrangian

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

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

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

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

Optimizing the Lagrangian

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

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

Optimizing the Lagrangian

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

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

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

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

Optimizing the Lagrangian

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

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

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

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

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

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

Optimizing the Lagrangian

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

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

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

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

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

Optimizing the Lagrangian

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

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

Optimizing the Lagrangian

↓

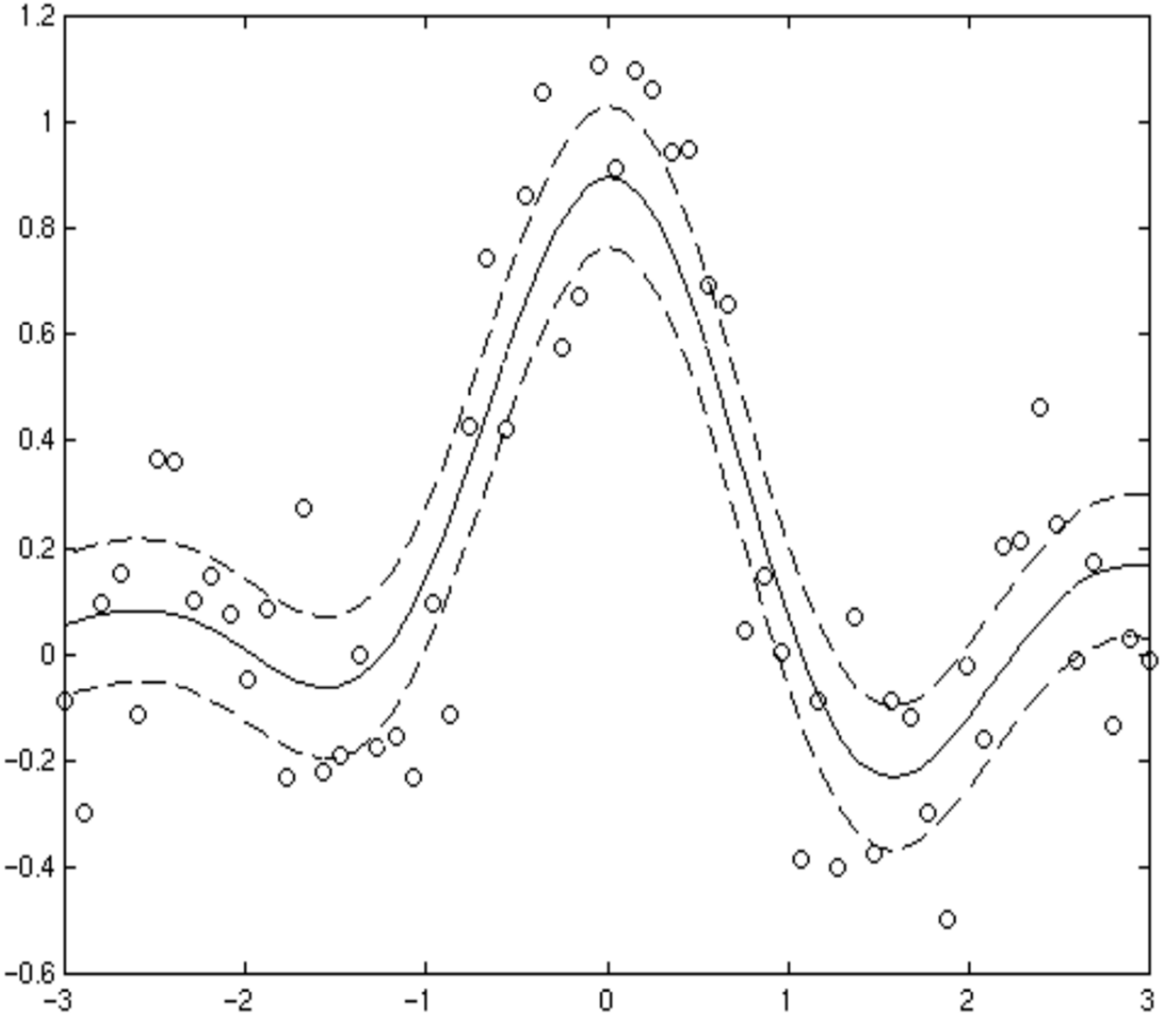
Dual optimization

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

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

Now we can use the similar tricks as in SVM

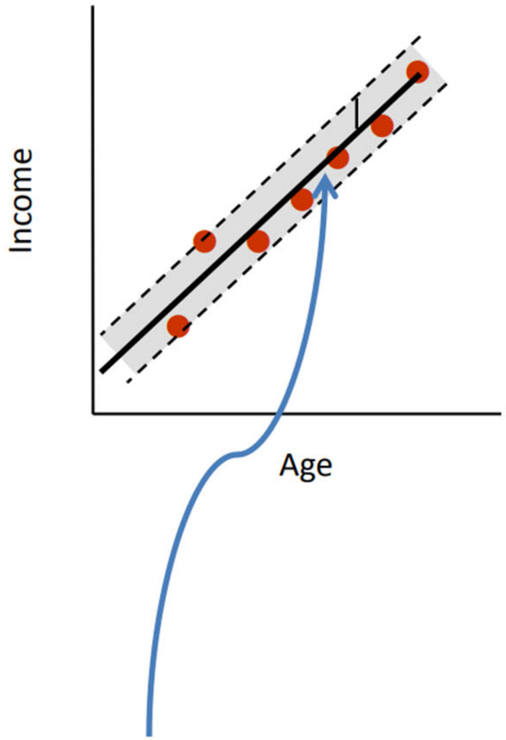
How about a non-linear case?



How about a non-linear case?

- **Linear case**

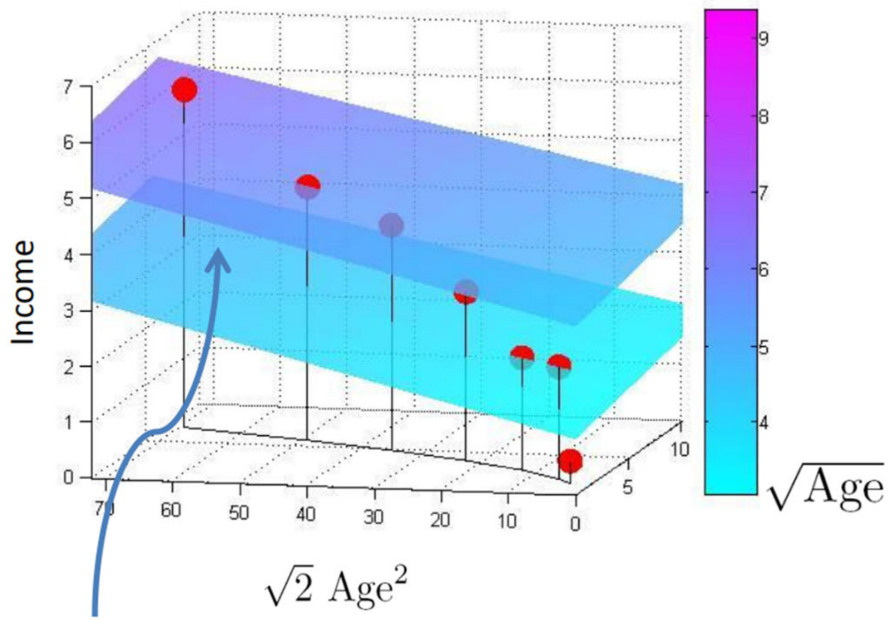
$$f: age \rightarrow income$$



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

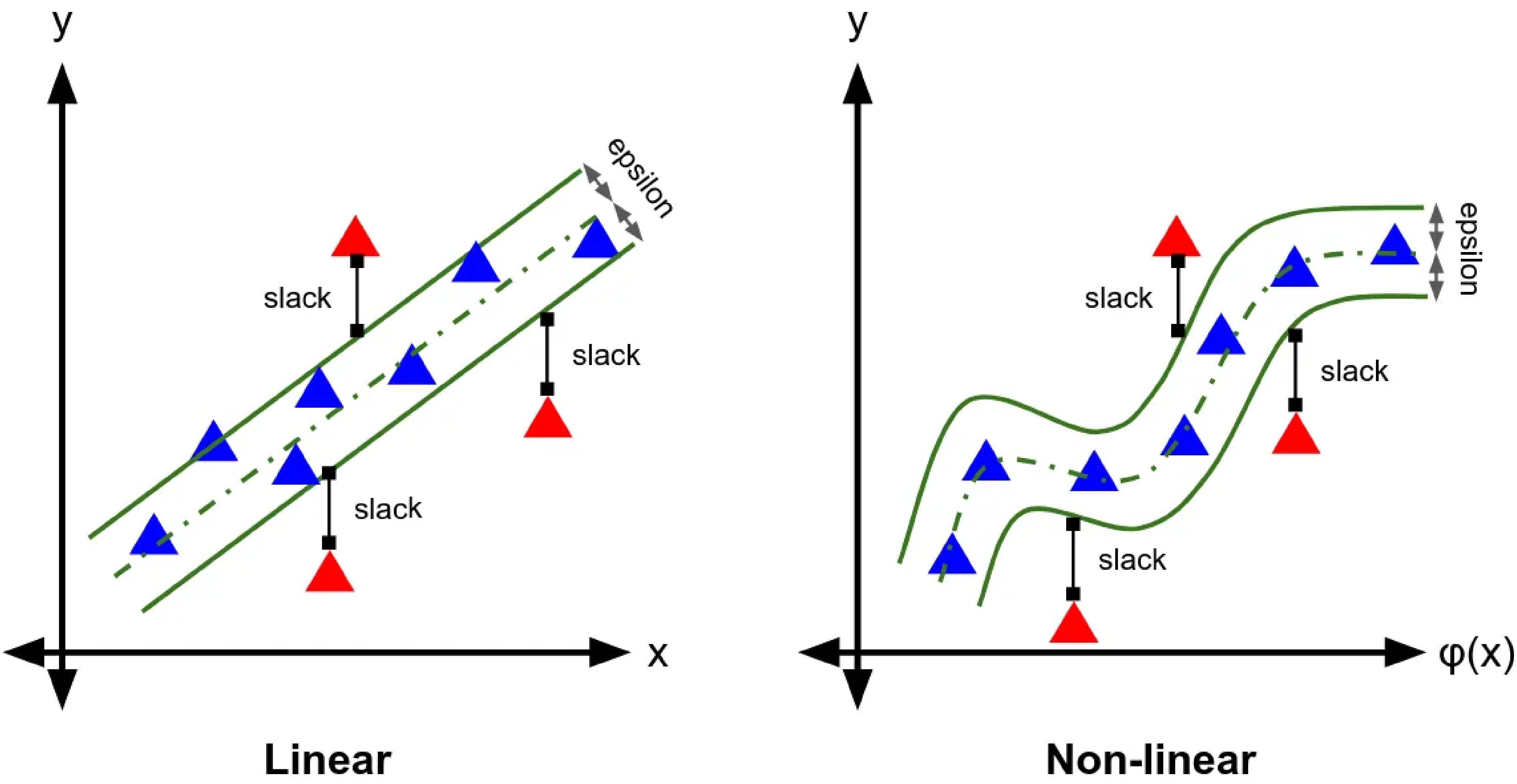
- **Non-linear case**
  - Map data into a higher dimensional space, e.g.,

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



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

Linear vs Non-linear



Dual problem

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

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

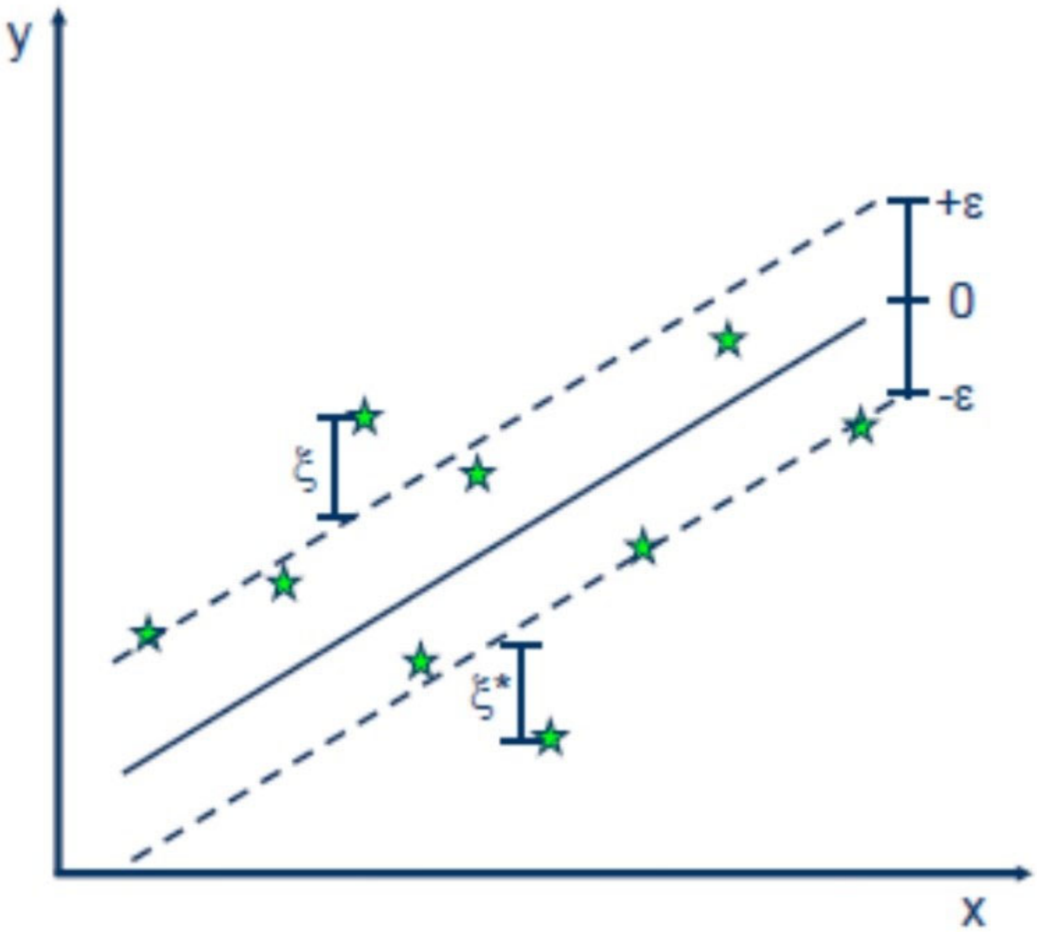
$$\max\begin{cases} \frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle \\ -\varepsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i(\alpha_i - \alpha_i^*) \end{cases}$$

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

Primal variables: w for each feature dim	Dual variables: α, α* for each data point
Complexity: the dim of the input space	Complexity: Number of support vectors

Primal **Dual**

Dual problem



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

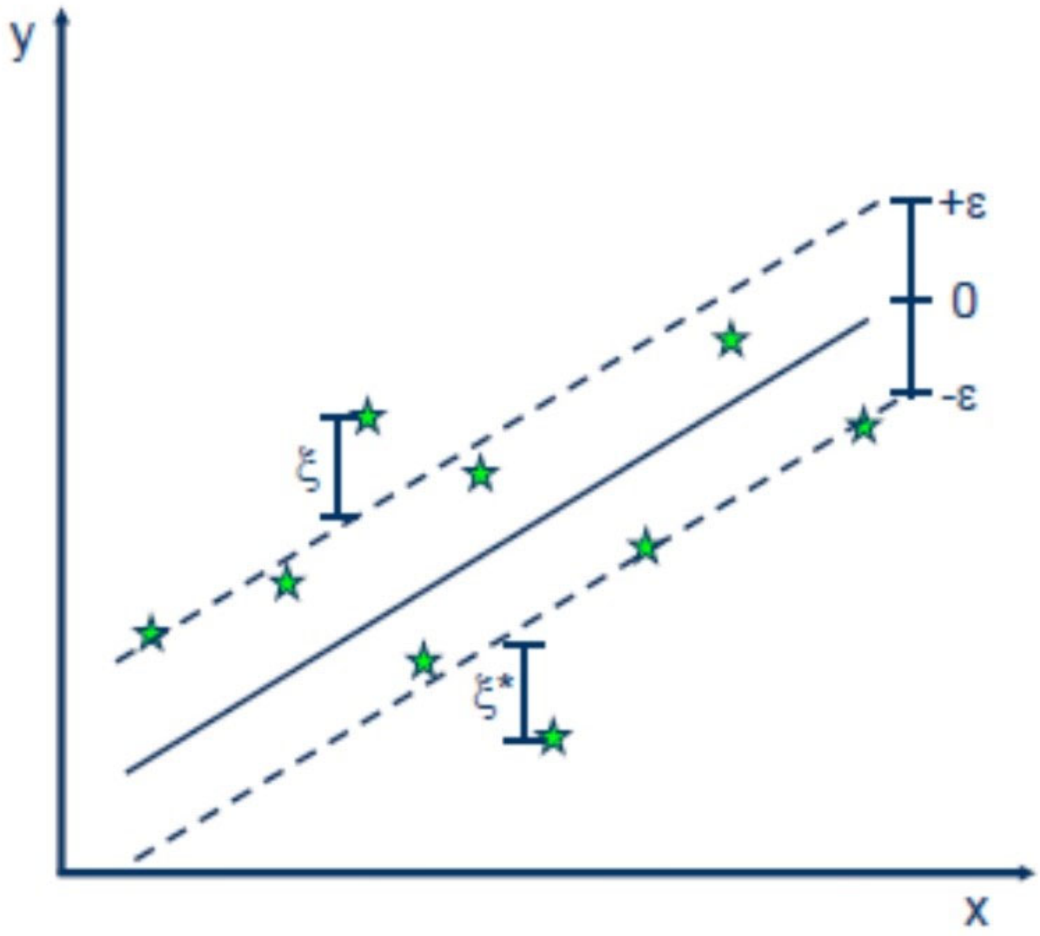
$$s.t.\sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0;\ \ 0 \le \alpha_i, \alpha_i^* \le C$$

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

Complexity: Number of support vectors

Dual

Dual problem



Kernel trick

K(xi, xj)

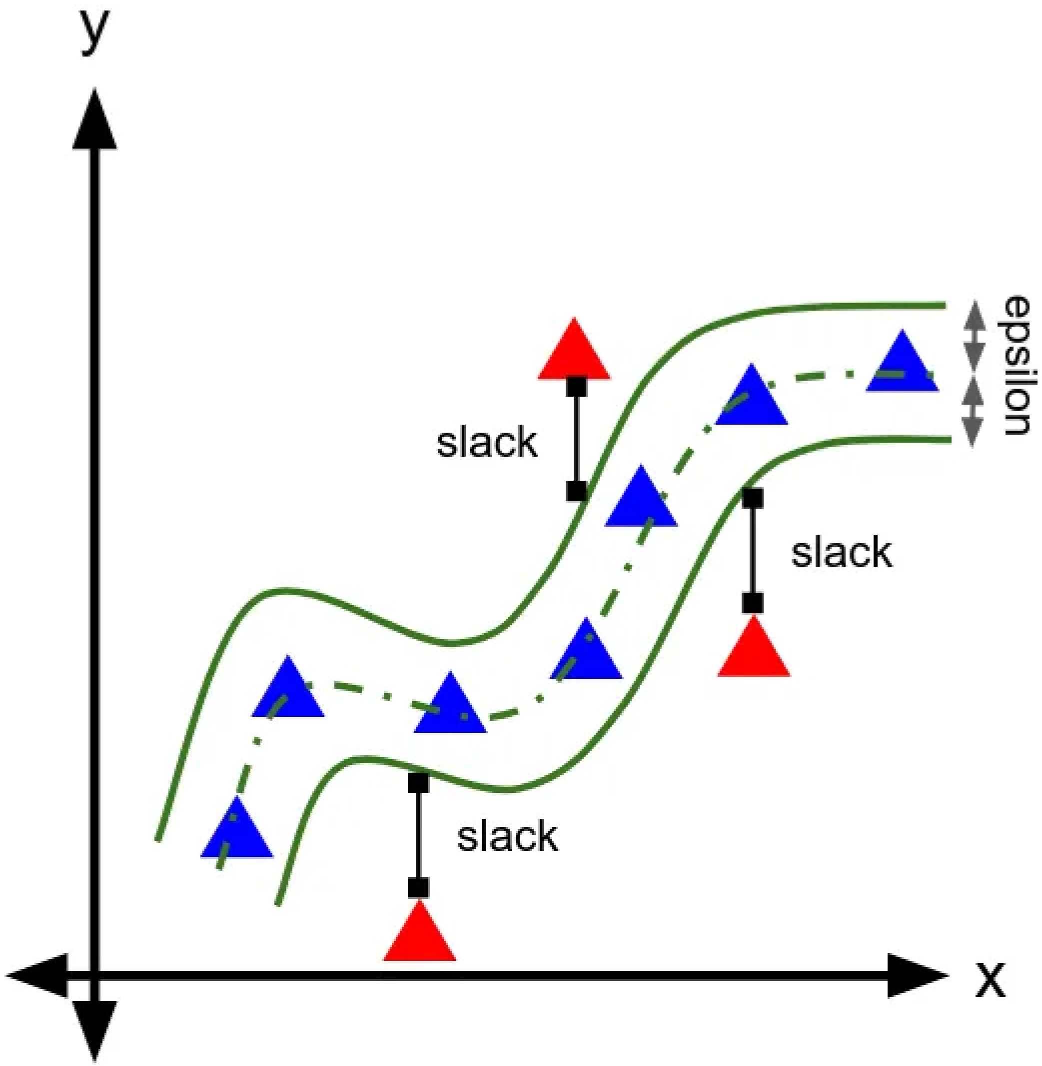
$$\max \begin{cases} \frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle\varphi(\mathbf{x}_i),\varphi(\mathbf{x}_j)\rangle \\ -\varepsilon\sum_{i=1}^{m}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i(\alpha_i - \alpha_i^*) \end{cases}$$

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

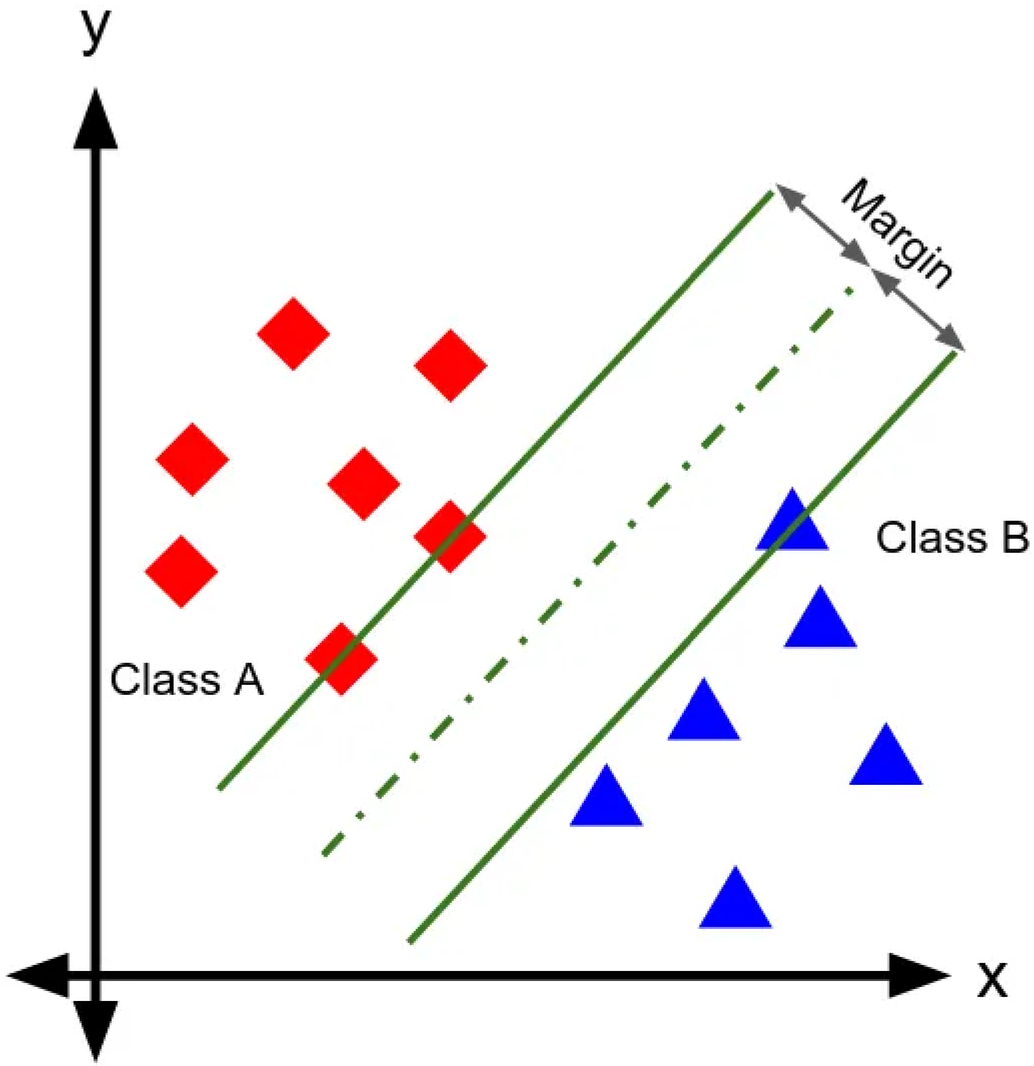
	Dual variables: α, α* for each data point
	Complexity: Number of support vectors

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