Regression as Machine Learning

Given

- A data universe X.
- A sample set S where $S \subset X$.
- Some target function $f : X \to \mathbb{R}$.
- A training set D, where $D = \{(x, y) \mid x \in S \text{ and } y = f(x)\}$.

Compute a model $\hat{f} : X \to \mathbb{R}$ using D such that,

$$\hat{f}(x) \cong f(x),$$

for all $x \in X$.

Observation: Same as machine learning in classification except for the co-domains of the target function and the model.

Question: How do we compute the model?

Statistical Approaches

Assume we have a regression training set of the form,

$$D = \{(\overline{x}_1, y_1), (\overline{x}_2, y_2), \ldots, (\overline{x}_l, y_l)\} \subset \mathbb{R}^n \times \mathbb{R}.$$

Then let $\hat{f}(\overline{x})$ be a regression model on D, where the quantity

$$\rho_i = y_i - \hat{f}(\overline{x}_i)$$

for $(\overline{x}_i, y_i) \in D$ is called a *residual* and measures the difference between model output and the actual observation. Observe that the residual depends on the model we choose.

In linear regression we compute the minimum *sum of squared errors* in order to obtain an optimal model,

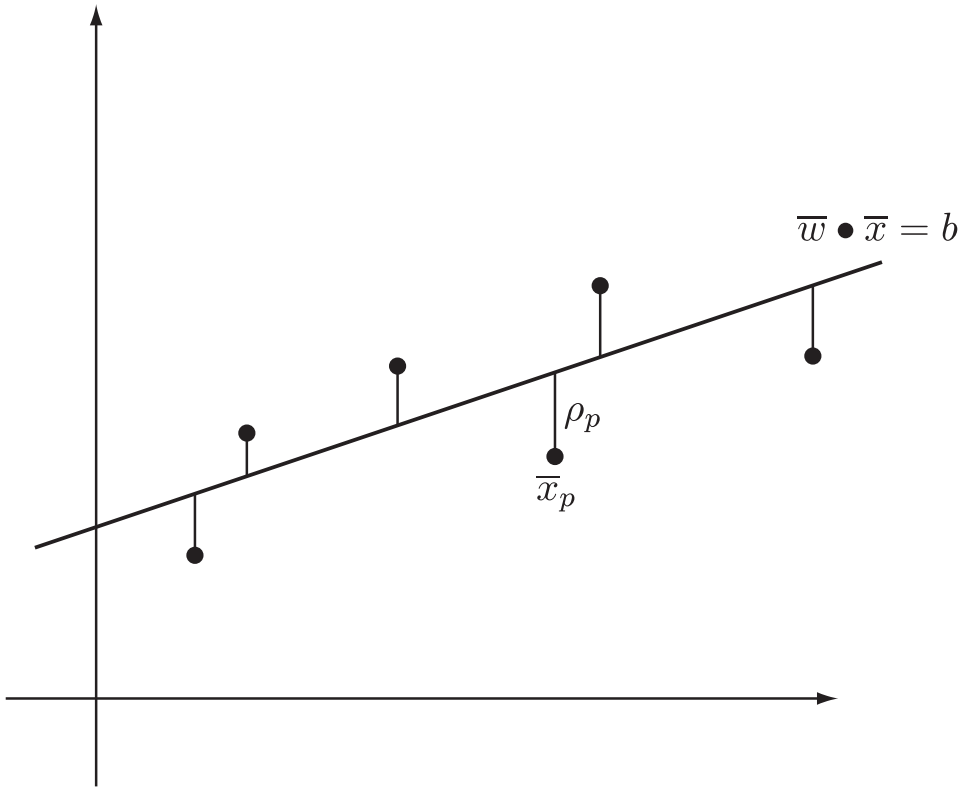
$$\min \sum_{i=1}^{l} \rho_i^2 = \min_{\hat{f}} \sum_{i=1}^{l} \left(y_i - \hat{f}(\overline{x}_i) \right)^2,$$

with $(\overline{x}_i, y_i) \in D$.

Rewriting the above optimization problem slightly we obtain,

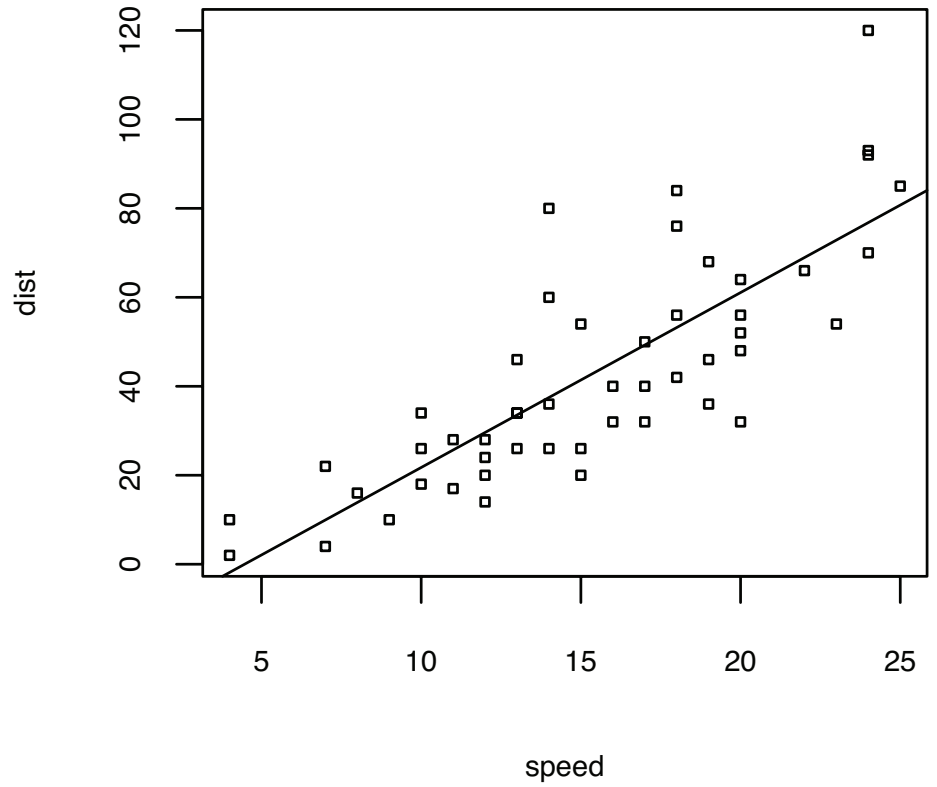
$$\hat{f}^* = \operatorname*{argmin}_{\hat{f}} \sum_{i=1}^{l} \left(y_i - \hat{f}(\overline{x}_i) \right)^2.$$

Statistical Approaches



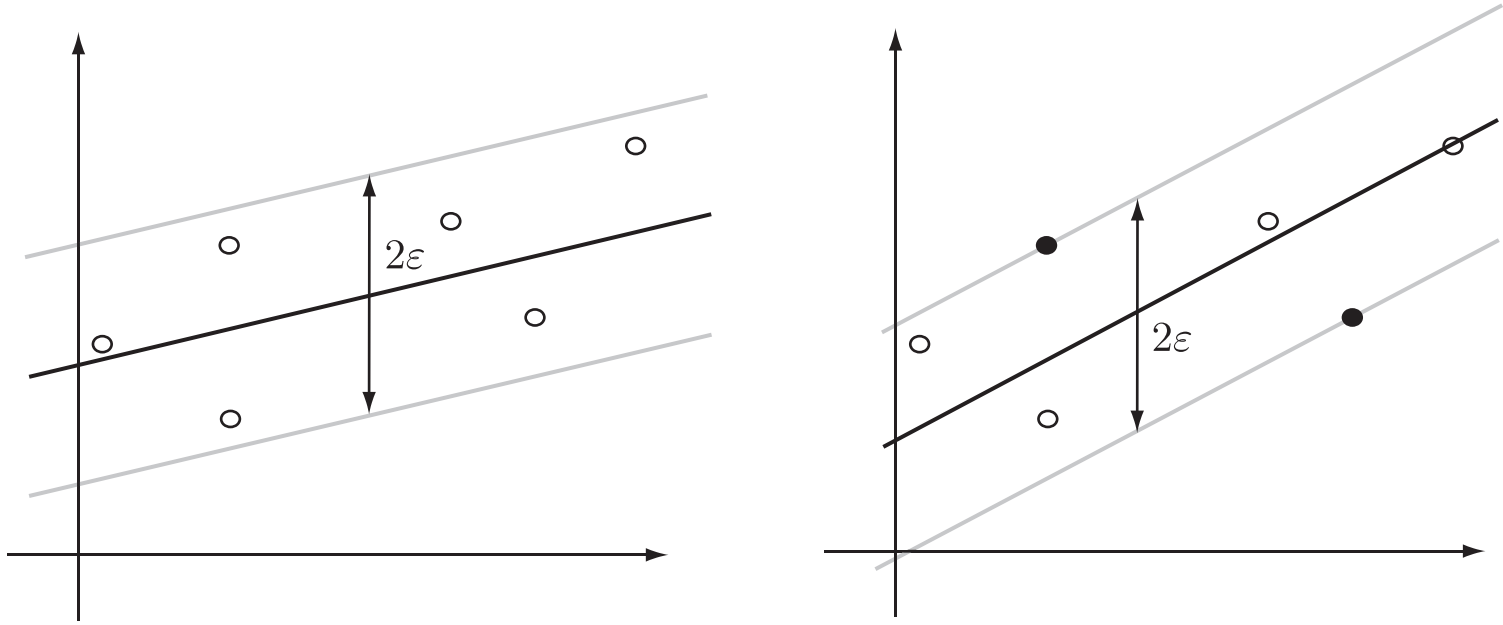
Linear regression with residuals, here the point $\overline{x}_p$ is an observation and ρ_p is the residual at that observation given the model $\overline{w} \bullet \overline{x} = b$.

Example




```
> data(cars)
> model <- lm(cars$dist ~ ., data = cars)
> plot(cars)
> abline(model)
```

Maximum Margin Machines



a) b)

Solving regression problems with linear models using a ε hyper-tube. In part a) we show a regression model where all observations are within the hyper-tube depicted with the light gray lines, part b) depicts the optimal regression model with a maximum margin.

Maximum Margin Machines

Proposition: Given the regression training set,

$$D = \{(\overline{x}_1, y_1), (\overline{x}_2, y_2), \ldots, (\overline{x}_l, y_l)\} \subseteq \mathbb{R}^n \times \mathbb{R},$$

where the optimal maximum margin regression model can be computed as the optimization,

$$\min \phi(\overline{w}, b) = \min_{\overline{w}, b} \frac{1}{2} \overline{w} \bullet \overline{w}$$

such that the constraints,

$$y_i - \hat{f}(\overline{x}_i) \leq \varepsilon,$$

$$\hat{f}(\overline{x}_i) - y_i \leq \varepsilon,$$

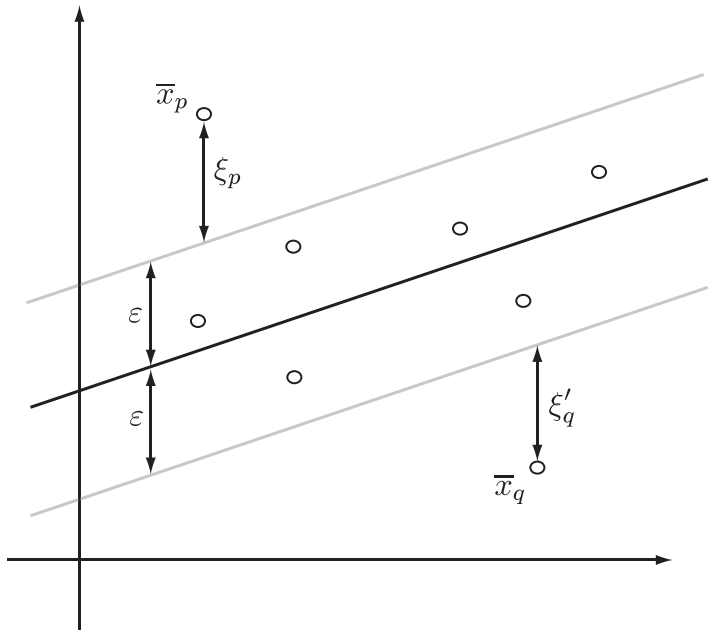
are satisfied for $i = 1, \ldots, l$ and where $\hat{f}(\overline{x}) = \overline{w} \bullet \overline{x} - b$.

The constraints specify the the solution must be a model such that the observations are contained within the ε-tube, $|y_i - \hat{f}(\overline{x}_i)| \leq \varepsilon$.

Maximum Margin Machines

In real-world settings it is unrealistic to assume that all observations will fall into a reasonable ε-tube, (e.g. cars data set).

For observations that fall outside the hyper-tube with a fixed value of ε we introduce correction terms or *slack variables* that tell us how much of a correction is needed in order for these observations to be moved into the hyper-tube.



Linear maximum margin regression with slack variables.

Maximum Margin Machines

We define the slack variables formally as,

$$\xi_i = \begin{cases} 0 & \text{if } y_i - \hat{f}(\overline{x}_i) \leq \varepsilon, \\ |y_i - \hat{f}(\overline{x}_i)| - \varepsilon & \text{otherwise,} \end{cases}$$

$$\xi_i' = \begin{cases} 0 & \text{if } \hat{f}(\overline{x}_i) - y_i \leq \varepsilon, \\ |y_i - \hat{f}(\overline{x}_i)| - \varepsilon & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, l$ with $(\overline{x}_i, y_i) \in D$.

Here the slack variables ξ_i are zero except for observations that lie above the hyper-tube. Conversely, the slack variables ξ_i' are zero except for observations that lie below the hyper-tube.

Maximum Margin Machines

We can now state regression with maximum margin machines as follows,

Proposition: Given a regression training set,

$$D = \{(\overline{x}_1, y_1), (\overline{x}_2, y_2), \ldots, (\overline{x}_l, y_l)\} \subseteq \mathbb{R}^n \times \mathbb{R},$$

we can compute the optimal regression model $\hat{f}^*(\overline{x}) = \overline{w}^* \bullet \overline{x} - b^*$ as the optimization

$$\min \phi(\overline{w}, b, \overline{\xi}, \overline{\xi}') = \min_{\overline{w}, b, \overline{\xi}, \overline{\xi}'} \frac{1}{2}\overline{w} \bullet \overline{w} + C \sum_{i=1}^{l} (\xi_i + \xi_i'),$$

such that the constraints,

$$\begin{aligned} y_i - \hat{f}(\overline{x}_i) &\leq \xi_i + \varepsilon, \\ \hat{f}(\overline{x}_i) - y_i &\leq \xi_i' + \varepsilon, \\ 0 &\leq \xi_i, \xi_i', \end{aligned}$$

for $i = 1, \ldots, l$ hold with $\hat{f}(\overline{x}) = \overline{w} \bullet \overline{x} - b$.

In the above proposition the penalty constant C modulates the trade-off between margin maximization and the minimization of the slack variables.

Regression with SVMs

Recall that SVMs are the dual to maximum margin machines. Also recall that we can derive the dual to maximum margin optimization by constructing the Lagrangian optimization,

$$\max_{\overline{\alpha}} \min_{\overline{x}} L(\overline{\alpha}, \overline{x}) = \max_{\overline{\alpha}} \min_{\overline{x}} \left(\phi(\overline{x}) - \sum_{i=1}^{l} \alpha_i g_i(\overline{x}) \right),$$

subject to the constraints,

$$\alpha_i \geq 0,$$

for $i = 1, \ldots, l$. Here $g_i(\overline{x}) \geq 0$ are inequality constraints and the variables $\overline{\alpha}$ and $\overline{x}$ are called the dual and primal variables of the optimization problem, respectively.

Regression with SVMs

As a first step in constructing the Lagrangian optimization we derive our inequality constraints. This is easily done by slightly rewriting the constraints appearing in the primal optimization problem,

$$\begin{aligned} \xi_i + \varepsilon - y_i + \hat{f}(\overline{x}_i) &\geq 0, \\ \xi_i' + \varepsilon - \hat{f}(\overline{x}_i) + y_i &\geq 0, \\ \xi_i &\geq 0, \\ \xi_i' &\geq 0. \end{aligned}$$

The four sets of inequality constraints imply that we have to introduce *four sets of dual variables* into our Lagrangian optimization.

Regression with SVMs

Our Lagrangian:

$$\max_{\overline{\alpha},\overline{\alpha}',\overline{\beta},\overline{\beta}'} \min_{\overline{w},b,\overline{\xi},\overline{\xi}'} L(\overline{\alpha},\overline{\alpha}',\overline{\beta},\overline{\beta}',\overline{w},b,\overline{\xi},\overline{\xi}') =$$

$$\max_{\overline{\alpha},\overline{\alpha}',\overline{\beta},\overline{\beta}'} \min_{\overline{w},b,\overline{\xi},\overline{\xi}'} \Bigg(\frac{1}{2}\overline{w} \bullet \overline{w} + C \sum_{i=1}^{l} (\xi_i + \xi_i')$$

$$- \sum_{i=1}^{l} \alpha_i \left(\xi_i + \varepsilon - y_i + \hat{f}(\overline{x}_i) \right)$$

$$- \sum_{i=1}^{l} \alpha_i' \left(\xi_i' + \varepsilon - \hat{f}(\overline{x}_i) + y_i \right)$$

$$- \sum_{i=1}^{l} \beta_i \xi_i$$

$$- \sum_{i=1}^{l} \beta_i' \xi_i' \Bigg),$$

subject to the constraints,

$$\alpha_i, \alpha_i', \beta_i, \beta_i' \geq 0,$$

for $i = 1, \ldots, l$ and where $\hat{f}(\overline{x}) = \overline{w} \bullet \overline{x} - b$.

Regression with SVMs

Given a solution to the Lagrangian optimization,

$$\max_{\overline{\alpha},\overline{\alpha}',\overline{\beta},\overline{\beta}'} \min_{\overline{w},b,\overline{\xi},\overline{\xi}'} L(\overline{\alpha},\overline{\alpha}',\overline{\beta},\overline{\beta}',\overline{w},b,\overline{\xi},\overline{\xi}') = L(\overline{\alpha}^*,\overline{\alpha}'^*,\overline{\beta}^*,\overline{\beta}'^*,\overline{w}^*,b^*,\overline{\xi}^*,\overline{\xi}'^*),$$

we know that the KKT conditions need to hold.

KKT Conditions

$$\frac{\partial L(\overline{\alpha}^*, \overline{\alpha}'^*, \overline{\beta}^*, \overline{\beta}'^*, \overline{w}^*, b^*, \overline{\xi}^*, \overline{\xi}'^*)}{\partial \overline{w}} = \overline{0},$$

$$\frac{\partial L(\overline{\alpha}^*, \overline{\alpha}'^*, \overline{\beta}^*, \overline{\beta}'^*, \overline{w}^*, b^*, \overline{\xi}^*, \overline{\xi}'^*)}{\partial b} = 0,$$

$$\frac{\partial L(\overline{\alpha}^*, \overline{\alpha}'^*, \overline{\beta}^*, \overline{\beta}'^*, \overline{w}^*, b^*, \overline{\xi}^*, \overline{\xi}'^*)}{\partial \xi_i} = 0,$$

$$\frac{\partial L(\overline{\alpha}^*, \overline{\alpha}'^*, \overline{\beta}^*, \overline{\beta}'^*, \overline{w}^*, b^*, \overline{\xi}^*, \overline{\xi}'^*)}{\partial \xi_i'} = 0,$$

$$\alpha_i^* \left(\xi_i^* + \varepsilon - y_i + \hat{f}^*(\overline{x}_i)\right) = 0,$$

$$\alpha_i'^* \left(\xi_i'^* + \varepsilon - \hat{f}^*(\overline{x}_i) + y_i\right) = 0,$$

$$\beta_i^* \, \xi_i^* = 0,$$

$$\beta_i'^* \, \xi_i'^* = 0,$$

$$\xi_i^* + \varepsilon - y_i + \hat{f}^*(\overline{x}_i) \geq 0,$$

$$\xi_i'^* + \varepsilon - \hat{f}^*(\overline{x}_i) + y_i \geq 0,$$

$$\xi_i^*, \xi_i'^* \geq 0,$$

$$\alpha_i, \alpha_i' \geq 0,$$

$$\beta_i^*, \beta_i'^* \geq 0,$$

where $i = 1, \ldots, l$ and $\hat{f}^*(\overline{x}) = \overline{w}^* \bullet \overline{x} - b^*$ is the optimal regression function.

Regression with SVMs

Proposition: Given a regression training set,

$$D = \{(\overline{x}_1, y_1), (\overline{x}_2, y_2), \ldots, (\overline{x}_l, y_l)\} \subseteq \mathbb{R}^n \times \mathbb{R},$$

then we can compute the optimal support vector regression model $\hat{f}^*(\overline{x}) = \overline{w}^* \bullet \overline{x} - b^*$ with

$$\max_{\overline{\alpha}, \overline{\alpha}'} \phi'(\overline{\alpha}, \overline{\alpha}') = \max_{\overline{\alpha}, \overline{\alpha}'} \left(\frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} (\alpha_i - \alpha_i')(\alpha_j - \alpha_j') \overline{x}_i \bullet \overline{x}_j + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i') - \varepsilon \sum_{i=1}^{l} (\alpha_i + \alpha_i') \right),$$

subject to the constraints, $\sum_{i=1}^{l} (\alpha_i - \alpha_i') = 0$ and $C \geq \alpha_i, \alpha_i' \geq 0$, for $i = 1, \ldots, l$ where

$$\overline{w}^* = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i'^*) \overline{x}_i,$$

$$b^* = \frac{1}{l} \sum_{i=1}^{l} \overline{w}^* \bullet \overline{x}_i - y_i.$$

Regression with SVMs

As in the case of classification it is perhaps interesting to look at a solution to the optimization in terms of the complementarity conditions,

$$\alpha_i^* \left(\xi_i^* + \varepsilon - y_i + \hat{f}^*(\overline{x}_i) \right) = 0,$$

$$\alpha_i'^* \left(\xi_i'^* + \varepsilon - \hat{f}^*(\overline{x}_i) + y_i \right) = 0,$$

$$\beta_i^* \xi_i^* = 0,$$

$$\beta_i'^* \xi_i'^* = 0,$$

and the fact that a support vector is now described by the coefficient $(\alpha_i - \alpha_i') \neq 0$.

Regression with SVMs

First we show that if some point $\overline{x}_i$ is strictly contained in the ε-tube then it is not a support vector. If the point is strictly within the ε-tube then the following is true

$$\varepsilon > y_i - \hat{f}(\overline{x}_i),$$
$$\varepsilon > \hat{f}(\overline{x}_i) - y_i.$$

and $\xi_i, \xi_i' = 0$. This means that the following holds,

$$\xi_i + \varepsilon - y_i + \hat{f}(\overline{x}_i) > 0,$$
$$\xi_i' + \varepsilon - \hat{f}(\overline{x}_i) + y_i > 0.$$

This implies that $\alpha_i = 0$ and $\alpha_i' = 0$ in order for the complimentarity conditions to be satisfied.

It follows that the coefficient $(\alpha_i - \alpha_i') = 0$, thus a point $\overline{x}_i$ contained in the ε-tube is *not* a support vector.

Regression with SVMs

Now consider a point $\overline{x}_i$ that sits right on the ε-tube boundary, then either

$$\xi_i + \varepsilon - y_i + \hat{f}(\overline{x}_i) = 0,$$
$$\xi_i' + \varepsilon - \hat{f}(\overline{x}_i) + y_i > 0,$$

or

$$\xi_i + \varepsilon - y_i + \hat{f}(\overline{x}_i) > 0,$$
$$\xi_i' + \varepsilon - \hat{f}(\overline{x}_i) + y_i = 0,$$

with $\xi_i, \xi_i' = 0$.

(the point $\overline{x}_i$ cannot be on both boundaries at the same time)

This in turn implies that either $0 < \alpha_i < C$ and $\alpha_i' = 0$ or $0 < \alpha_i' < C$ and $\alpha_i = 0$.

Note that in this case the coefficient $(\alpha_i - \alpha_i') \neq 0$ and therefore the point $\overline{x}_i$ is considered a support vector.

Regression with SVMs

Finally, consider the point $\overline{x}_i$ outside of the ε-tube, then either

$$\xi_i + \varepsilon - y_i + \hat{f}(\overline{x}_i) = 0,$$
$$\xi_i' + \varepsilon - \hat{f}(\overline{x}_i) + y_i > 0,$$

or

$$\xi_i + \varepsilon - y_i + \hat{f}(\overline{x}_i) > 0,$$
$$\xi_i' + \varepsilon - \hat{f}(\overline{x}_i) + y_i = 0,$$

(the point $\overline{x}_i$ cannot be above both boundaries at the same time)

This in turn implies that either $\alpha_i = C$ and $\alpha_i' = 0$ with $\xi_i > 0$ and $\xi_i' = 0$ or $\alpha_i' = C$ and $\alpha_i = 0$ with $\xi_i = 0$ and $\xi_i' > 0$.

Note that in this case the coefficient $(\alpha_i - \alpha_i') \neq 0$ and therefore the point $\overline{x}_i$ is considered a support vector.