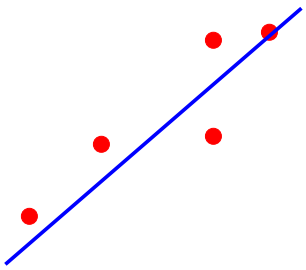
Support Vector Regression

Contents

- What is regression?
- Primal Linear Regression
- Ridge Regression
- Nonlinear Regression
- Support Vector Regression
- Geometric interpretation
- SVR Applications

Regression

- Find a function that will fit the data with minimal error.
- For example, linear regression finds the best line that fits the data.

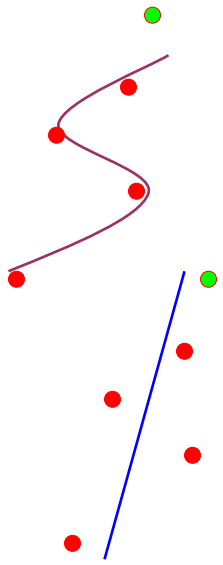


Regression optimization problem

- Define the set of INPUTS (experimental features) as $$X = \{x_1, \ldots x_N\}$$
- Define the set of OUTPUTS as $$Y = \{y_1, \ldots y_N\}$$
- Find *f(x)* that minimizes $$\sum_{i}^{N} |f(x_i) - y_i|$$

Robust regression models

- Overfitting → Poor predictions
- A robust regression model
  - does not overfit the data
  - can predict well on future data



Linear Regression – simple 1 dim

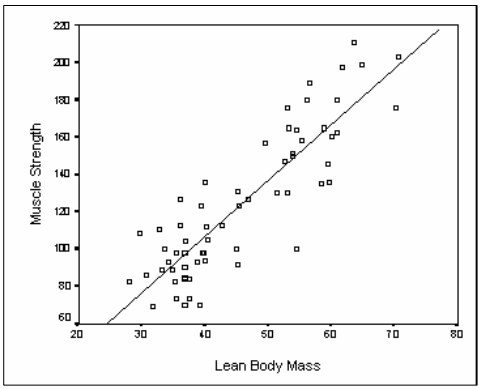
Example: X=height, Y=weight

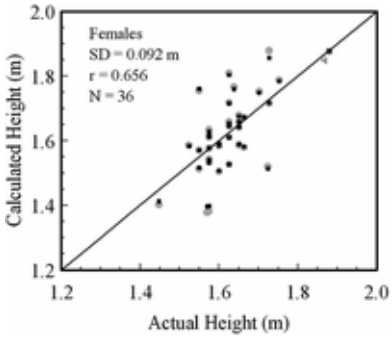
X=predictors, Y=dependent variable

Want to find a linear function that is the best predictor of Y.

Let linear function be aX+b.

Want a and b such that the difference between the predicted values and the real values Y is minimized.





Minimize squared error

Let $X = \{x_1, \ldots x_n\}$ and $Y = \{y_1, \ldots y_n\}$.

Find a and b that minimizes
$R^2(a, b) = \sum_i (y_i - (ax_i + b))^2$.

Take derivatives w.r.t. a and $b \ldots$

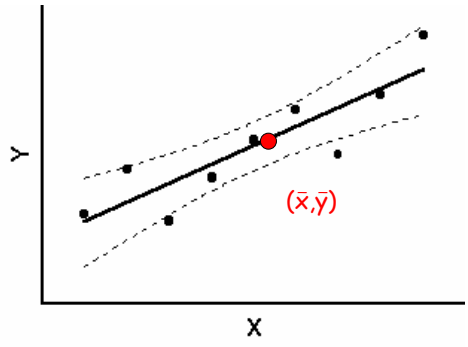
$$\frac{\partial R^2}{\partial a} = \sum_{i=1}^{n} -2x_i[y_i - (ax_i + b)] = 0$$

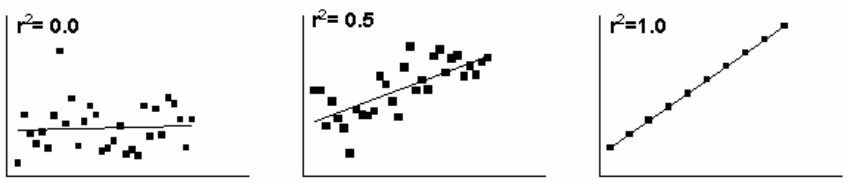
$$\frac{\partial R^2}{\partial b} = \sum_{i=1}^{n} -2[y_i - (ax_i + b)] = 0$$

Solve for a and $b \ldots$

$$a = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$b = \bar{y} - a\bar{x}$$





Linear Regression – D dimensions

Input variables: $\{x_1, \ldots x_n\} \in X$
where each x_i is a real-valued vector

Labels: $y_1, \ldots y_n$ where $y_i \in \Re$

Want to find w such that $|y = \langle w, x \rangle|$ is minimized.
Similar to the 1-dim case, with different notation.

Loss function: $L = \| y - Xw \|^2$
$\frac{\partial L}{\partial w} = -2X'y + 2X'Xw = 0$
$w = (X'X)^{-1}X'y$

Matrix notation

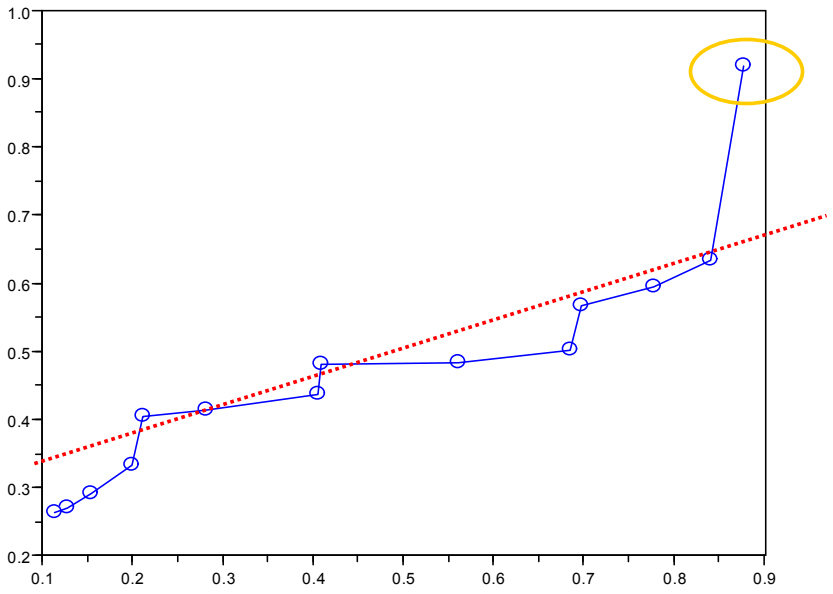
Nonlinear Regression

The ridge regression method we saw earlier finds a linear relationship between X and Y.

For nonlinear relationships, use an embedding to convert nonlinear relations into linear ones … kernel trick!

Regression redefined:
Find a function w such that
$|y - \langle w, \phi(x) \rangle|$ is minimal.

But now we face the problem of overfitting!!



Ridge Regression

$$\min_w \gamma \parallel w \parallel^2 + \sum_i (y_i - \langle w, x_i \rangle)^2$$

The parameter Υ is tunable parameter and represents the trade off between complexity and loss.
Another way is as follows:

$$\begin{array}{ll} \min_w & \sum_i (y_i - \langle w, \phi(x_i) \rangle)^2 \\ \text{s.t.} & \parallel w \parallel \leq B \end{array}$$

Dual details in KMPA, Chapters 2 & 7

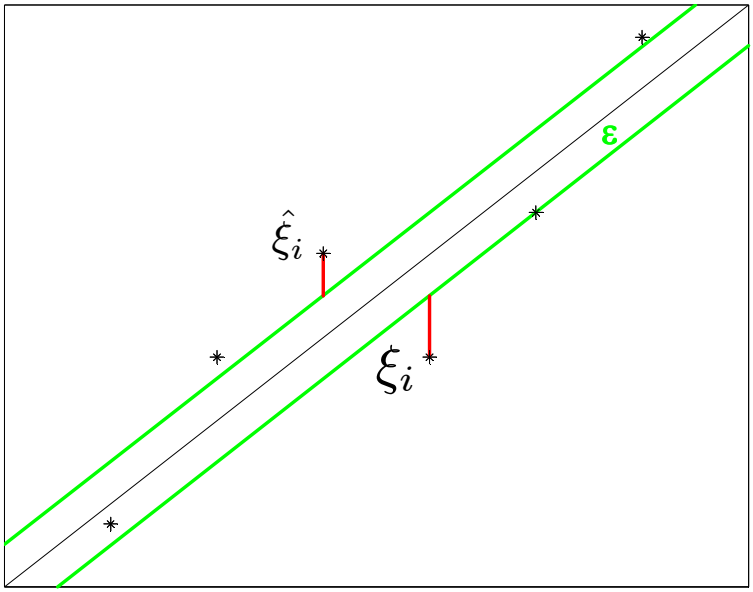
ε-Insensitive SVR

$$\begin{array}{ll} \min_{w,b,\xi,\hat{\xi}} & \frac{1}{2} \parallel w \parallel^2 + C \sum_i (\xi_i^2 + \hat{\xi}_i^2) \\ \text{s.t.} & (\langle w, \phi(x_i) \rangle + b) - y_i \leq \epsilon + \xi_i \\ & y_i - (\langle w, \phi(x_i) \rangle + b) \leq \epsilon + \hat{\xi}_i \\ & \xi_i, \hat{\xi}_i \geq 0 \end{array}$$

Slack variables

Use KKT conditions and Lagrangian to derive the dual.
Dual derivation details in KMPA, Chapter 7

Support Vector Regression



ε-Insensitive SVR - Dual

Primal

$$\min_{w,b,\xi,\hat{\xi}} \quad \frac{1}{2} \parallel w \parallel^2 + C \sum_i (\xi_i^2 + \hat{\xi}_i^2)$$

$$\text{s.t.} \quad (\langle w, \phi(x_i) \rangle + b) - y_i \leq \epsilon + \xi_i$$
$$y_i - (\langle w, \phi(x_i) \rangle + b) \leq \epsilon + \hat{\xi}_i$$
$$\xi_i, \hat{\xi}_i \geq 0$$

Lagrangian

$$L(w, b, \alpha, \hat{\alpha}) = \frac{1}{2} \parallel w \parallel^2 + C \sum_i (\xi_i^2 + \hat{\xi}^2)$$
$$+ \sum_i \alpha_i [(\langle w, \phi(x_i) \rangle + b) y_i - \epsilon - \xi_i]$$
$$+ \sum_i \hat{\alpha}_i [y_i - (\langle w, \phi(x_i) \rangle + b) - \epsilon - \hat{\xi}_i]$$

Dual

$$\max \quad \sum_i (\hat{\alpha}_i - \alpha_i) y_i - \epsilon \sum_i (\hat{\alpha}_i + \alpha_i)$$
$$-0.5 \sum_{i,j} (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j) \kappa(x_i, x_j)$$

$$\text{s.t.} \quad \sum_i (\hat{\alpha}_i - \alpha_i) = 0$$
$$0 \leq \alpha_i, \hat{\alpha}_i \leq C$$

Applications and recent work

- *U* SVR (more later)
- Semi-supervised regression (reading pres.)
- ✓ Compressed Regression

Shuheng Zhou, John Lafferty, Larry Wasserman
http://books.nips.cc/papers/files/nips20/NIPS2007_0195.pdf

Idea behind compressed regression is to **compress the input variables** and develop a **sparse linear model**.
Primary compression motivation: preserve privacy
Secondary motivations: storage and manipulation
Results: predicts as well as regular methods