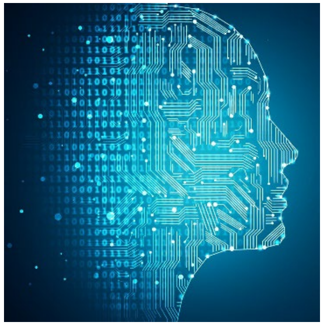
Machine Learning

Linear Regression

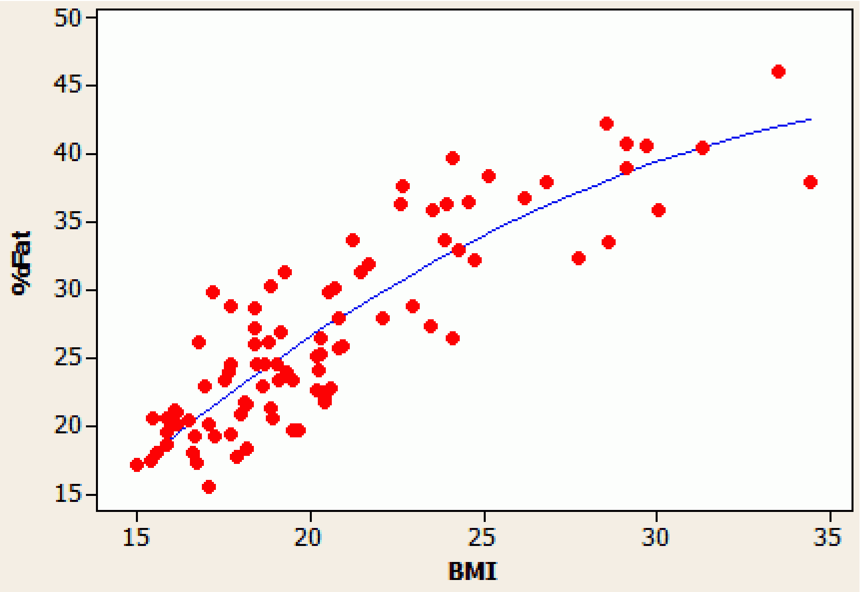
Kevin Moon (kevin.moon@usu.edu)

STAT/CS 5810/6655



Regression

- Regression is the other main supervised learning problem besides classification.
- Every feature vector $\boldsymbol{X}$ is associated to a variable Y, and the goal is to predict Y from $\boldsymbol{X}$.
- As in classification, this prediction function must be learned from training data



Mean Squared Error

- Main difference between regression and classification?
  - Regression $\Rightarrow Y$ continuous; classification $\Rightarrow Y$ discrete
- Motivates different performance measures
- Probabilistic setting: jointly distributed variables $(\boldsymbol{X}, Y)$ where

$$\boldsymbol{X} \in \mathbb{R}^d, \qquad \mathrm{Y} \in \mathbb{R}$$

  and the goal is to predict Y from $\boldsymbol{X}$ using a *regression function*

$$f: \mathbb{R}^d \to \mathbb{R}$$

- The *mean squared error* of a regression function f is

$$R(f) \coloneqq \mathbb{E}_{\boldsymbol{X},Y}\left[\left(Y - f(\boldsymbol{X})\right)^2\right]$$

Conditional Mean

- Just like classification, there is a regression function f^* that achieves the minimum value R^* of the mean squared error

- **Theorem:** The function

$$f^*(\boldsymbol{x}) := E_{Y|\boldsymbol{X}}[Y|\boldsymbol{X} = \boldsymbol{x}]$$

  minimizes the mean squared error.

- This function is called the *conditional mean* predictor.

Conditional Mean

Proof of Theorem: Let f be any regression function.

$$
\begin{aligned}
R(f) &= E_{\boldsymbol{X}Y}[(f(\boldsymbol{X}) - Y)^2] \\
&= E_{\boldsymbol{X}} E_{Y|\boldsymbol{X}}[(f(\boldsymbol{X}) - Y)^2 | \boldsymbol{X}] \\
&= E_{\boldsymbol{X}} E_{Y|\boldsymbol{X}}[(f(\boldsymbol{X}) - E[Y|\boldsymbol{X}] + E[Y|\boldsymbol{X}] - Y)^2 | \boldsymbol{X}] \\
&= E_{\boldsymbol{X}} E_{Y|\boldsymbol{X}}[(f(\boldsymbol{X}) - E[Y|\boldsymbol{X}])^2] + (E[Y|\boldsymbol{X}] - Y)^2 \\
&\qquad - 2(f(\boldsymbol{X}) - E[Y|\boldsymbol{X}])(E[Y|\boldsymbol{X}] - Y) | \boldsymbol{X}]
\end{aligned}
$$

The second term is independent of f, and the third term is zero. The first term can be made to equal 0 by taking f to be the conditional mean, so this minimizes the MSE.

Linear Regression

- In practice we don't have access to the joint distribution and must estimate f^* using training data $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)$.

- Choose f to minimize the *empirical MSE*

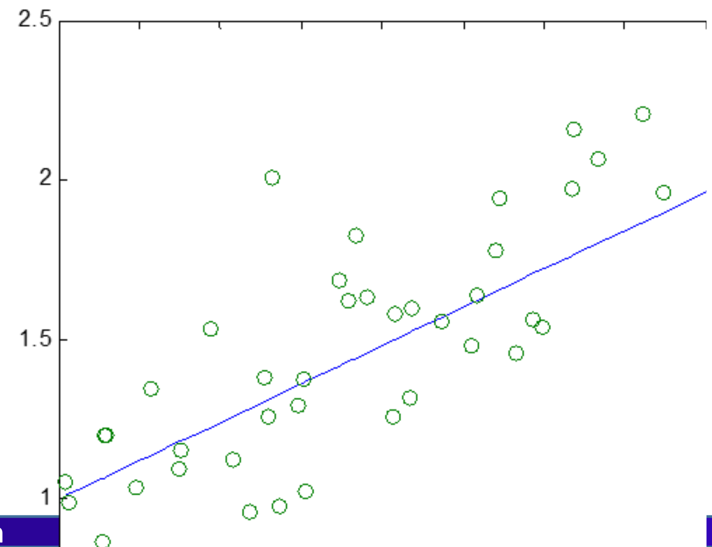
$$\widehat{R}(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2$$

- To make this optimization tractable, we need to restrict f to belong to a *regression model*, i.e., a class of candidates for f.

- We'll initially focus on the *linear model*

$$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$$

where $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$.

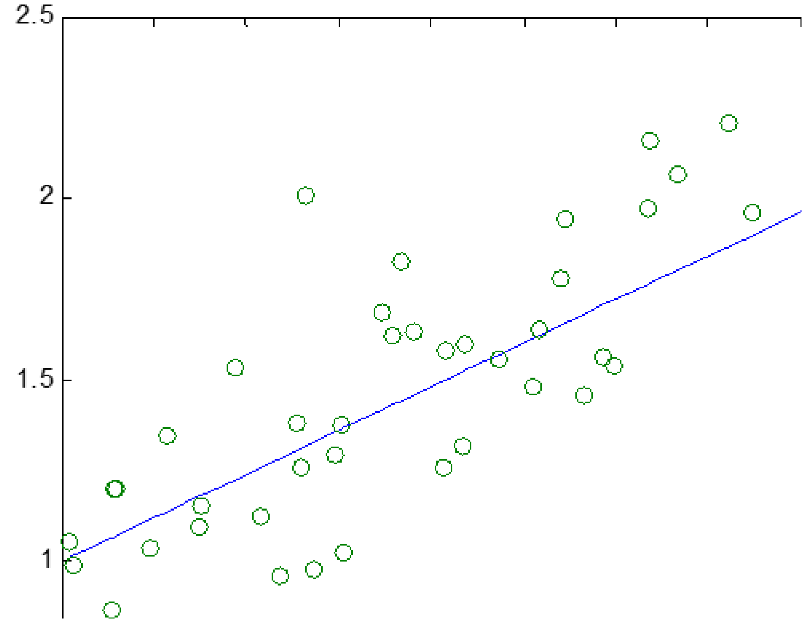


Least Squares Linear Regression

- *Least squares linear regression* solves

$$\min_{\boldsymbol{w},b} \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b)^2$$

The method is also known as *ordinary least squares.*



- For greater generality, we can add a *regularization term*

$$\min_{\boldsymbol{w},b} \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b)^2 + \lambda\|\boldsymbol{w}\|^2$$

This method is known as *ridge regression*, and the term $\lambda\|\boldsymbol{w}\|^2$ is called the *ridge penalty*. $\lambda \geq 0$ is the *regularization parameter*.

Ridge Regression Solution (1)

First, eliminate b

(b is called the "offset" or "bias")

$$\frac{\partial}{\partial b}(\text{obj. fun}) = -\cancel{\frac{2}{n}} \sum_i (y_i - w^T x_i - b) = 0$$

$$nb = \sum (y_i - w^T x_i)$$

$$b = \frac{1}{n} \sum (y_i - w^T x_i) = \bar{y} - w^T \bar{x}$$

where $\bar{y} = \frac{1}{n} \sum y_i$, $\bar{x} = \frac{1}{n} \sum x_i$

Ridge Regression Solution (2)

Eliminating b, the objective function becomes

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - \bar{y} - \boldsymbol{w}^T(\boldsymbol{x}_i - \bar{\boldsymbol{x}})]^2 + \lambda\|\boldsymbol{w}\|^2$$

So let's denote $\widetilde{y}_i = y_i - \bar{y}$, $\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \bar{\boldsymbol{x}}$.

$$\frac{1}{n}\sum(\tilde{y}_i - w^T\tilde{x}_i)^2 + \lambda\|w\|^2$$

$$= \frac{1}{n}\|\tilde{y} - \tilde{X}w\|^2 + \lambda\|w\|^2$$

$$\tilde{y} = \begin{bmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_n \end{bmatrix}, \quad \tilde{X} = \begin{bmatrix} \tilde{x}_1^{(1)} & \cdots & \tilde{x}_1^{(d)} \\ \vdots & & \vdots \\ \tilde{x}_n^{(1)} & \cdots & \tilde{x}_n^{(d)} \end{bmatrix}$$

Ridge Regression Solution (3)

After further simplification

$$\text{obj. fun.} \propto \|\tilde{y} - \tilde{X}w\|^2 + n\lambda\|w\|^2$$

$$= (\tilde{y} - \tilde{X}w)^T(\tilde{y} - \tilde{X}w) + n\lambda w^T w$$

$$= \tilde{y}^T\tilde{y} \underbrace{- \tilde{y}^T(\tilde{X}w) - (\tilde{X}w)^T\tilde{y}}_{-2\tilde{y}^T\tilde{X}w} + w^T(\tilde{X}^T\tilde{X} + n\lambda I)w$$

$$-2\tilde{y}^T\tilde{X}w = -2(\tilde{X}^T\tilde{y})^T w$$

Group Exercise

We have shown that the regularized least squares (i.e. ridge regression) objective function can be written (after eliminating b) as

$$J(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T A \boldsymbol{w} + \boldsymbol{r}^T \boldsymbol{w} + c,$$

where $A = 2\left(\tilde{X}^T \tilde{X} + n\lambda I\right)$, $r = -2\tilde{X}^T \tilde{\boldsymbol{y}}$, and $c = \tilde{\boldsymbol{y}}^T \tilde{\boldsymbol{y}}$.

1. Verify that A is PSD if $\lambda \geq 0$ and PD if $\lambda > 0$
2. Determine a minimizer
3. Explain why regularization is necessary when $d > n$

OLS Alternate Solution

- When $\lambda = 0$ (OLS), there is an alternate, but equivalent, solution
- Set $\boldsymbol{\theta} = \begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix}$
- Rewrite the objective function:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i - b)^2 = \frac{1}{n}||\boldsymbol{y} - X\boldsymbol{\theta}||^2$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_1^{(d)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^{(1)} & \dots & x_n^{(d)} \end{bmatrix}$$

Group Exercise

1. Determine a formula for the minimizer in the alternate form of OLS.
2. What is a drawback of the two least squares solutions we have discussed today?

Further Reading

- ISL Chapter 3
- ESL Chapter 3