Lecture 4: Regression Introduction
DD2431

Atsuto Maki

Autumn, 2015

Part I: we will visit

- Function approximation
- Linear Regression / Least Squares
  - RANSAC (handling outliers)
- KNN Regression

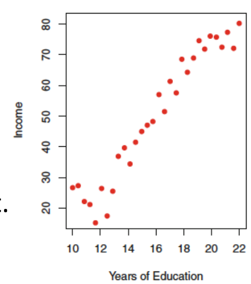
Regression => Real-valued output

Function approximation

- How do we fit this dataset D?

$$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$$

of N pairs of inputs x_i and targets $y_i \in R$.
D can be measurements in an experiment.



- Task of regression:
  to predict target associated to any arbitrary new input

Note: Here we have a single *input feature*, but inputs to regression tasks are often vectors $\mathbf{x}$ of *multiple input features*.

Linear Regression (parametric)

Linear regression tries to estimate the function f and predict the output by

$$\hat{f}(x) = \sum_{i=0}^{d} w_i x_i = w^T x$$

How to measure the error:

- To see how well $\hat{f}(x)$ approximates $f(x)$, square error is used: $(\hat{f}(x) - f(x))^2$
- Mean Square Error: $E_{in}(\hat{f}) = \frac{1}{N} \sum_{n=1}^{N} (\hat{f}(x_n) - y_n)^2$
  (in-sample)

Minimizing in-sample MSE

E_{in} can be expressed as:

$$E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N}(w^T x_n - y_n)^2 = \frac{1}{N}\|Xw - Y\|^2$$

where

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

We want to compute the parameters w that minimize E_{in}.

Residual sum of squares (RSS)

The sum of squared errors is a convex function of w

$$E_{in}(w) = \|Xw - Y\|^2$$

The gradient with respect to the weights is:

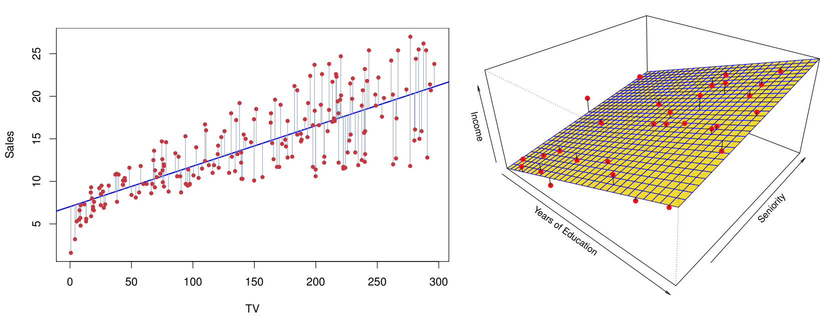
$$\frac{\partial}{\partial w} E_{in}(w) = 2X^T\left(Xw - Y\right)$$

The weight vector that sets the gradient to zero minimizes the errors

$$X^T X w = X^T Y$$

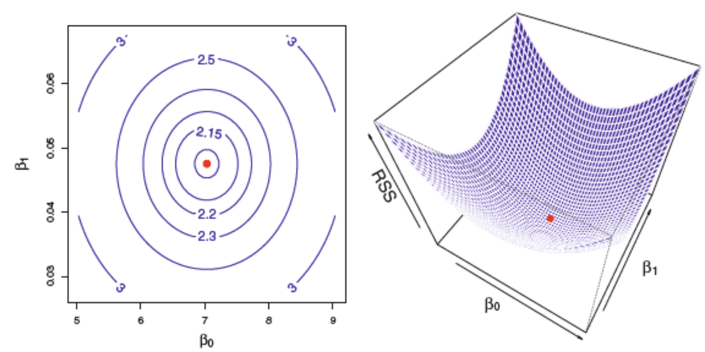
$$w = \left(X^T X\right)^{-1} X^T Y$$

Examples of least squares fit



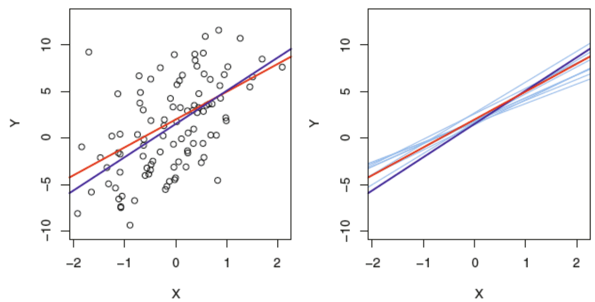
Figures from An Introduction to Statistical Learning (G. James et al.)

Examples of plots of RSS



Figures from An Introduction to Statistical Learning (G. James et al.)

Least squares line



- Red: the true relationship f(x) = 2 + 3x, the population regression line
- Blue: the least squares line, estimate based on the observed data
- Light blue (in right): least squares lines, each based on a separate random set of observations

Figures from An Introduction to Statistical Learning (G. James et al.)

RANSAC: RANdom SAmpling Consensus

<u>Objective</u>

Robust fit of model to data set *S* which contains outliers

<u>Algorithm</u>

(i) Randomly select a sample of *s* data points from *S* and instantiate the model from this subset.

(ii) Determine the set of data points S_i which are within a distance threshold *t* of the model. The set S_i is the consensus set of samples and defines the inliers of *S*.

(iii) If the subset of S_i is greater than some threshold *T*, re-estimate the model using all the points in S_i and terminate

(iv) If the size of S_i is less than *T*, select a new subset and repeat the above.

(v) After *N* trials the largest consensus set S_i is selected, and the model is re-estimated using all the points in the subset S_i

(in Hartley and Zisserman, adapted from Fischler '81)

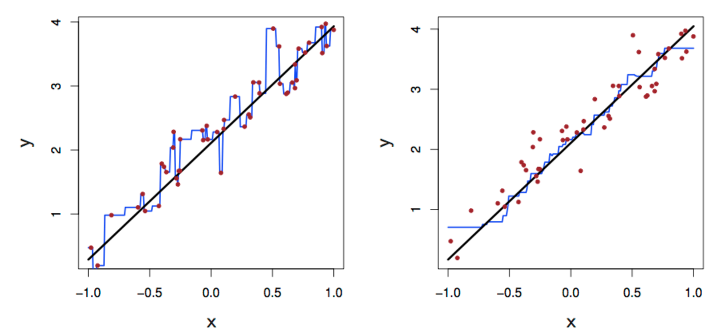
k-NN Regression (non-parametric)

- Similar to the k-NN classifier
- To regress Y for a given value of X, consider k closest points to X in training data and take the average of the responses.

$$f(x) = \frac{1}{k} \sum_{x_i \in N_i} y_i$$

- Larger values of k provide a smoother and less variable fit (lower variance!)

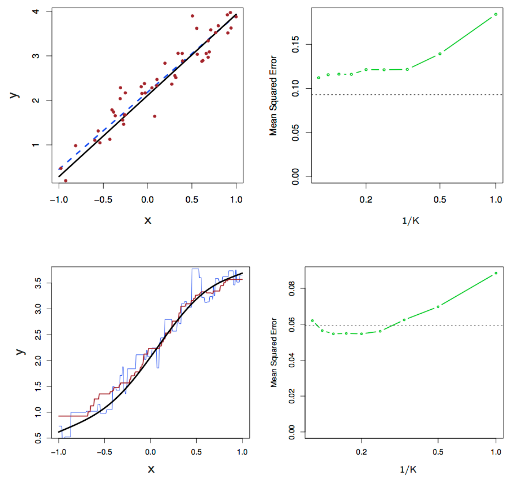
Example plots of $\hat{f}(x)$ with k-NN regression (1d)



k = 1 k = 9

Figures from An Introduction to Statistical Learning (G. James et al.)

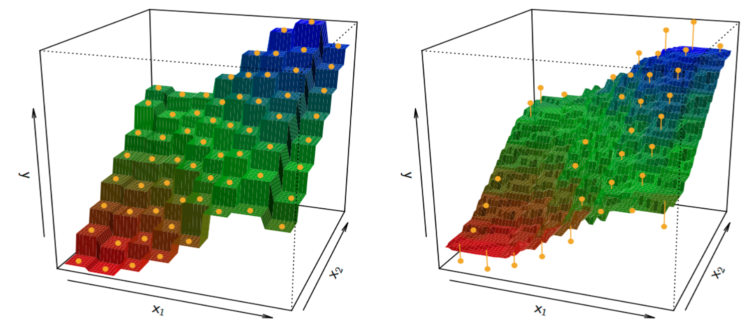
k-NN vs. Linear Regression



k = 1 (blue)
k = 9 (red)

Figures from An Introduction to Statistical Learning (G. James et al.)

Example plots of $\hat{f}(x)$ with k-NN regression (2d)



k = 1 k = 9

In higher dimensions k-NN often preforms worse than linear regression.

Figures from An Introduction to Statistical Learning (G. James et al.)

Part II: we will visit

- Linear regression + regularization
  - Ridge regression
  - The Lasso (a more recent alternative)

Ridge regression

Similar to least squares but minimizes different quantity:

$$RSS + \lambda \sum_{i=1}^{d} w_i^2$$

The second term is called shrinkage penalty

- Shrinkage penalty: small when w_i are close to zero
- The parameter λ : controls the relative impact of the two terms, the selection is critical!

Sample problem: The Credit dataset

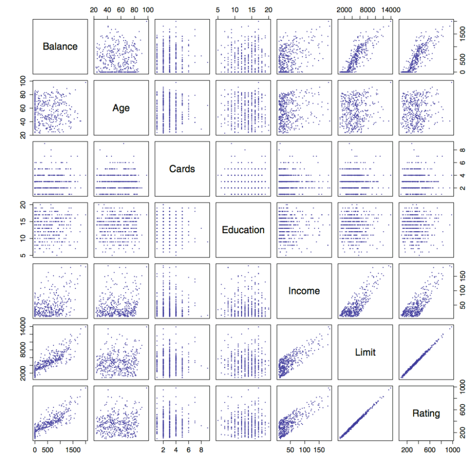
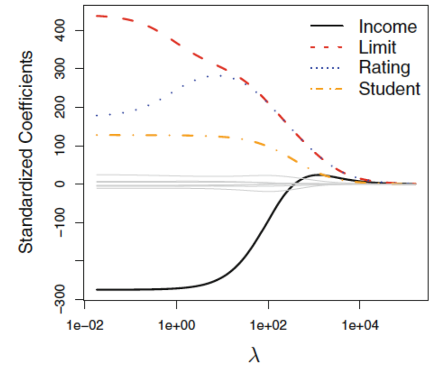


FIGURE 3.6. *The* Credit *data set contains information about* balance, age, cards, education, income, limit, *and* rating *for a number of potential customers.*

Figures from An Introduction to Statistical Learning (G. James et al.)

Ridge regression coefficients

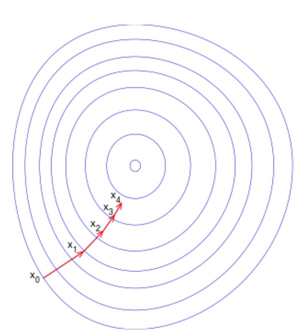


As λ increases, the standardized coefficients shrinks towards zero (but not exactly forced to zero).

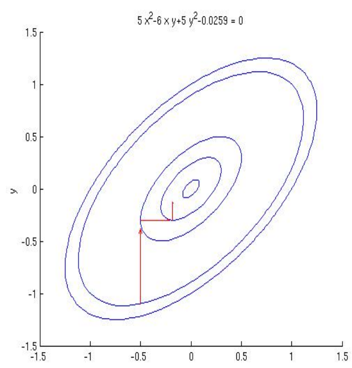
Figures from An Introduction to Statistical Learning (G. James et al.)

Approaches to parameter estimations

- Gradient decent

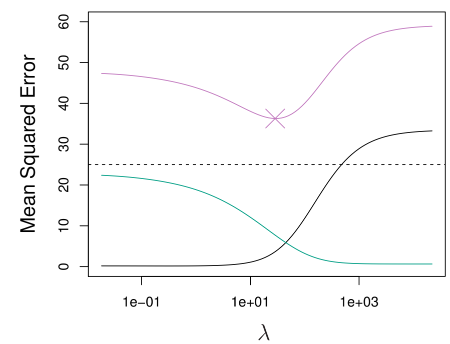


- Coordinate decent



Ridge Regression Bias/Variance

- Purple: MSE
- Black: Bias
- Green: Variance



Increase λ decreases variance while increasing bias

Figures from An Introduction to Statistical Learning (G. James et al.)

The Lasso
(Least Absolute Shrinkage and Selection Operator)

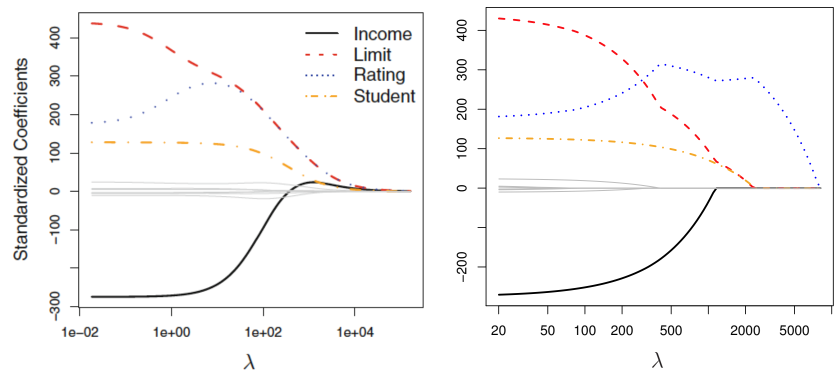
Similar to ridge regression but with slightly different term:

$$RSS + \lambda \sum_{i=1}^{d} |w_i|$$

The shrinkage penalty is now replaced by l_1 norm

- Ridge regression: it includes all features in the final model, making it harder to interpret – its drawback
- The lasso could be proven mathematically that some coefficients end up being set to exactly zero
  - variable selection
  - yielding sparse model

Comparison of estimated coefficients



Ridge regression The Lasso

Figures from An Introduction to Statistical Learning (G. James et al.)

Another formulations

For every value of λ there is some s such that the equations will give the same coefficient estimates:

- Ridge regression: Mimimizing $RSS + \lambda \sum_{i=1}^{d} w_i^2$

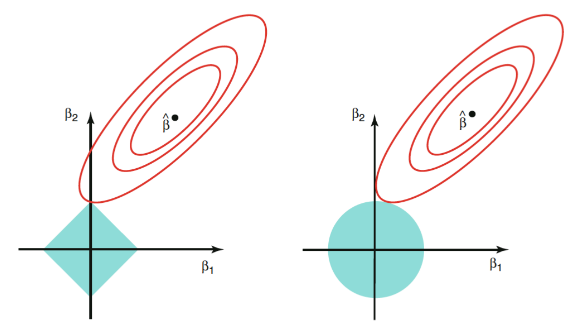
  Mimimizing $RSS, sub.to \sum_{i=1}^{d} w_i^2 \le s$

- Lasso: Mimimizing $RSS + \lambda \sum_{i=1}^{d} |w_i|$

  $RSS, sub.to \sum_{i=1}^{d} |w_i| \le s$

The variable selection property

The coefficient estimates: the first point where an ellipse contacts the constraint region as it expands.



The solid blue areas are the constraint regions for

Left: the Lasso Right: Ridge regression

Figures from An Introduction to Statistical Learning (G. James et al.)