Linear and Non-linear Regression

by *Ramesh Hariharan*

created: 2015-Jul-17

updated: 2018-Aug-09

pages: 22

Linear and Non-linear Regression

Data Analytics Course, Aug 2015

Ramesh Hariharan

The Regression Problem

- Given an (unknown) function $g(x)$ evaluated at n points $x_1 \ldots x_n$
- The goal is recover g
- Of course, g is not specified uniquely by these n evaluations
- So assume g is of a simple, given form, say $ax + b$ or $ax^2 + bx + c$, where a, b, c are unknown parameters
- Find the a, b, c that match the evaluations most closely

The Regression Problem More Precisely

- Given n pairs x_i, y_i
- $x_i \in R^d$, $y_i \in R$
- Given a model $f_\Theta(x)$, $x \in R^d$, Θ represents the unknown parameters
- Find Θ that minimizes $h(\Theta) = \sum_i |f_\Theta(x_i) - y_i|^2$

Linear Regression

- Assume $f_{\Theta}(x) = x^T\Theta$, $\Theta, x \in R^d$
- Find Θ that minimizes $h(\Theta) = \sum_i |{x_i}^T\Theta - y_i|^2$
- Set $\frac{\partial h}{\partial \Theta}$ to 0
- Check that this is indeed a minimum, not a maximum, not a saddle etc

Determining Parameters in Linear Regression

- Rewriting more cleanly..
- $h(\Theta) = (X\Theta - y)^T(X\Theta - y)$, where X is $n \times d$, the ith row of X is x_i
- $\frac{dh}{d\Theta} = 2X^T(X\Theta - y)$, where $\frac{dh}{d\Theta}$ is a vector of partial derivatives w.r.t each of the entries in the vector Θ
- Setting these to 0, we get a system of d equalities $X^T X\Theta = X^T y$. Solution: $(X^T X)^{-1} X^T y$

Global Minimality

- Invertibility? One solution if invertible, many otherwise. Why?
- Each solution of the form $\Theta = (X^T X)^{-1} X^T Y$ is a global minimum. Why?
- By Taylor's expansion for multivariate functions, for any vector $\Delta = (\Delta_1, \Delta_2, \ldots, \Delta_d) \in R^d$,

$$h(\Theta + \Delta) = h(\Theta) + \Delta^T h'(\Theta) + \frac{1}{2}\Delta^T h''(\Theta)\Delta$$

- Higher order terms are 0 for linear regression

Multivariate Taylor's Theorem Outline

- $r(t) = s(x(t), y(t))$ Univariate Taylor: $r(t) = r(0) + tr'(0) + \frac{t^2}{2} r''(0) + \cdots$
- $r'(t) = \frac{\partial s(x,y)}{\partial x} \frac{\partial x}{\partial t} + \frac{\partial s(x,y)}{\partial y} \frac{\partial y}{\partial t}$
- $r''(t) = \frac{\partial^2 s(x,y)}{\partial x \partial x} (\frac{\partial x}{\partial t})^2 + 2 \frac{\partial^2 s(x,y)}{\partial y \partial x} \frac{\partial y}{\partial t} \frac{\partial x}{\partial t} + \frac{\partial^2 s(x,y)}{\partial y \partial y} (\frac{\partial y}{\partial t})^2$ plus terms which involve second derivative w.r.t t
- $x(t) = x_0 + \Delta_1 t$, $y(t) = y_0 + \Delta_2 t$
- $s(x_0 + \Delta_1 t, y_0 + \Delta_2 t) = s(x_0, y_0) + t\Delta^T s'(x_0, y_0) + \frac{t^2}{2} \Delta^T s''(x_0, y_0) \Delta + \cdots$

Global Minimality

- $h'(\Theta) = \begin{pmatrix} \frac{\partial h}{\partial \Theta_1} \\ \frac{\partial h}{\partial \Theta_2} \\ \vdots \\ \frac{\partial h}{\partial \Theta_d} \end{pmatrix}, h''(\Theta) = \begin{pmatrix} \frac{\partial^2 h}{\partial \Theta_1 \partial \Theta_1} & \frac{\partial^2 h}{\partial \Theta_1 \partial \Theta_2} & \cdots & \frac{\partial^2 h}{\partial \Theta_1 \partial \Theta_d} \\ \frac{\partial^2 h}{\partial \Theta_2 \partial \Theta_1} & \frac{\partial^2 h}{\partial \Theta_2 \partial \Theta_2} & \cdots & \frac{\partial^2 h}{\partial \Theta_2 \partial \Theta_d} \\ \ddots & & & \\ \frac{\partial^2 h}{\partial \Theta_d \partial \Theta_1} & \frac{\partial^2 h}{\partial \Theta_d \partial \Theta_2} & \cdots & \frac{\partial^2 h}{\partial \Theta_d \partial \Theta_d} \end{pmatrix}$
- We chose Θ so $h'(\Theta) = 0$.
- Claim: $\Delta^T h''(\Theta) \Delta \geq 0$, which suffices for global minimality

Global Minimality and Uniqueness

- $h''(\Theta)$, also called the Hessian, is exactly $2X^TX$!
- $\Delta^T X^T X \Delta = |X\Delta|^2 \geq 0$
- If X (which is $n \times d$, $n \geq d$) has full rank d then $X\Delta$ is always non-zero, so $|X\Delta|^2 > 0$, i.e., there is a unique global minimum
- Otherwise, any Δ in the nullspace of X satisfies $X\Delta = 0$ and $h(\Theta + \Delta) = h(\Theta)$, so moving around in the nullspace yields all global minima

Invertibility of $X^T X$

- $rank(X) = rank(X^T X)$. Why? Because the nullspaces are the same.
- If X has rank d then $X^T X$ is invertible and there is an unique global minimum
- If X has smaller rank, then $X^T X$ is not invertible, then how do we find solutions to $X^T X\Theta = X^T y$?
- Note $X^T X$ takes an arbitrary vector Θ and transforms it into the column space of X^T.

The Pseudo-Inverse of $X^T X$

- Since $X^T X$ is a real, symmetric, positive semi-definite $d \times d$ matrix, all its eigenvalues are real and non-negative
- The corresponding eigenvectors are either in the column space of X^T or orthogonal to it, and form an orthonormal basis
- Eigenvector v has eigenvalue 0 if it is orthogonal to all the columns of X^T and eigenvalue non-0 if it is in the column space of X^T
- $X^T X = \sum_i e_i v_i {v_i}^T$ where e_i are the non-0 eigenvalues and v_i denotes the corresponding eigenvectors

The Pseudo-Inverse of $X^T X$ Contd.

- Consider $(\sum_i \frac{1}{e_i} v_i {v_i}^T) X^T y$ as a solution for Θ
- Claim: $(X^T X)(\sum_i \frac{1}{e_i} v_i {v_i}^T) X^T y = X^T y$
- Proof:
  - $(\sum_i e_i v_i {v_i}^T)(\sum_i \frac{1}{e_i} v_i {v_i}^T) X^T y = (\sum_i v_i {v_i}^T) X^T y$
  - $(\sum_i v_i {v_i}^T) x = x$ for any vector x in the column space of X^T
  - Since $X^T y$ is in the column space of X^T, $(\sum_i v_i {v_i}^T) X^T y = X^T y$, as needed

Summarizing the Non-Invertible Case

- Since the column space of X^T has dimension at least 1, at least one eigenvalue of $X^T X$ is non-zero, hence there is at least one solution satisfying $h'(\Theta) = 0$, namely $\Theta = (\sum_i \frac{1}{e_i} v_i {v_i}^T) X^T y$

- This is a global minimum, though not unique, any $h(\Theta + \Delta)$ where Δ is in the nullspace of X (equvalently, orthogonal to column space of X^T) is also a global minimum

- There is a unique global minimum in the column space of X^T

Running Time

- Compute $X^T X$
- Compute eigenvalues and eigenvectors
- Compute $(\sum_i \frac{1}{e_i} v_i {v_i}^T) X^T y$
- Naively, $O(d^2 n + d^3)$
- There are ways to compute $X^T X$ faster, approximately, via random sampling

The Non-Linear Case

- Say $f_\Theta(x) = (x^T\Theta)^2$ but not $f_\Theta(x) = (x^2)^T\Theta$, where x^2 denotes the vector x with entries squared; assume well-behaved (continuously differentiable)
- $h(\Theta + \Delta) = h(\Theta) + \Delta^T h'(\Theta) + \Delta^T h''(\Theta)\Delta + O(\Delta^3)$ for small enough Δ
- $h'(\Theta) = 0$ can no longer be solved using linear equations. Needs a general root-finding procedure
- $h''(\Theta)$ need not be of the form $A^T A$, which means $\Delta^T h''(\Theta)\Delta$ could be negative, i.e., a root need not be local minimum, let alone a global one

The Gauss-Newton Approach

- Start with a particular Θ_0
- Linearize $f_\Theta(x)$ in the neighborhood of this Θ_0: $f_{\Theta_0+\Delta}(x) \sim f_{\Theta_0}(x) + [f'_{\Theta_0}(x)]^T \Delta$
- Minimize $hh(\Delta) = \sum_i |f_{\Theta_0}(x_i) + [f'_{\Theta_0}(x_i)]^T \Delta - y_i|^2$
- Use linear regression to find $\Delta = (Z^T Z)^{-1} Z^T (y - f_{\Theta_0}(x))$ that minimizes $hh(\Delta)$; iterate with $\Theta_0 = \Theta_0 + \Delta$
- Where the ith row of Z is the vector $[f'_{\Theta_0}(x_i)]^T$

Convergence of the Gauss-Newton Approach

- If $f_{\Theta_0+\Delta}(x) \sim f_{\Theta_0}(x) + [f'_{\Theta_0}(x)]^T \Delta$ is *close* to an equality then $\Theta_0 + \Delta$ provides a strictly better fit than Θ_0 unless Θ_0 is already a local minimum

- So if we start very close to a local minimum (so the linear approximation holds in a very close neighborhood), this will converge

- But no guarantee if you start further away.

Modified Gauss-Newton with Smaller Step Sizes

- $\Delta = (Z^T Z)^{-1} Z^T (y - f_{\Theta_0}(x))$, which equals $-\frac{1}{2}(Z^T Z)^{-1} h'(\Theta_0)$ (why?)
- $h(\Theta) = (f_\Theta(x) - y)^T (f_\Theta(x) - y)$ and $h'(\Theta) = 2Z^T (f_\Theta(x) - y)$
- $\Theta_0 + \alpha\Delta = \Theta_0 - \frac{\alpha}{2}(Z^T Z)^{-1} h'(\Theta_0)$
- For small enough $\alpha > 0$, $h(\Theta_0 + \alpha\Delta) \sim h(\Theta_0) - \frac{\alpha}{2} h'(\Theta_0)^T (Z^T Z)^{-1} h'(\Theta_0) < h(\Theta_0)$, at least if Z is full rank and Θ_0 is not already a local minimum

The Modified Gauss-Newton Algorithm for the Non-Linear Case

- Start with some Θ_0 and then move to
  $\Theta_0 - \alpha(Z^T Z)^{-1} h'(\Theta_0) = \Theta_0 - \alpha(Z^T Z)^{-1} Z^T (f_{\Theta_0}(x) - y)$
- Here the ith row of Z is the vector $[f_{\Theta_0}{}'(x_i)]^T$
- α can be determined by single-parameter minimization
  - e.g., bisection: start with an interval, evaluate derivative in the middle, if negative, go right, otherwise go left
- Stop when there isn't much progress

Some More Modifications: Levenberg

- Instead of $\Theta_0 - \alpha(Z^T Z)^{-1} h'(\Theta_0)$ one could simply follow steepest gradient to $\Theta_0 - \alpha I h'(\Theta_0)$
- The latter always leads to improvement but could be slow, the former is quicker if $f_{\Theta_0}(x)$ is linear in the neighbourhood of Θ_0
- Levenberg: Combine the two approaches, i.e., use $\Theta_0 - \alpha(Z^T Z + \lambda I)^{-1} h'(\Theta_0)$
- Adjust λ in each iteration, if there is quick descent then keep it low, otherwise increase it so you are closer to steepest descent

Some More Modifications: Marquardt

- Marquardt: Instead of $\Theta_0 - \alpha(Z^T Z + \lambda I)^{-1} h'(\Theta_0)$, use $\Theta_0 - \alpha(Z^T Z + \lambda \mathbf{diag}(Z^T Z))^{-1} h'(\Theta_0)$
- $$\mathbf{diag}(Z^T Z) = \begin{pmatrix} \sum_i (\frac{\partial f_{\Theta_0}(x_i)}{\partial \Theta_1})^2 & & & \\ & \sum_i (\frac{\partial f_{\Theta_0}(x_i)}{\partial \Theta_2})^2 & & \\ & & \ddots & \\ & & & \sum_i (\frac{\partial f_{\Theta_0}(x_i)}{\partial \Theta_d})^2 \end{pmatrix}$$
- When λ is large so one is in the steepest descent regime, move faster in directions where the partial derivative is smaller, in a bid to accelerate

Finishing Up

- Can we use these regression methods, particularly Levenberg-Marquardt, to identify the orbit of Mars, based on data in the previous lecture?