Linear Data Analysis
Cross Validating Linear Regression

Cain Susko

Queen's University
School of Computing

February 7, 2022

a Concepts in Statistical Regression

In this section we will view the problem of linear regression in the context of vector spaces.
Statistical Regression is, for a set of observations (a_1, a_2, a_n, c). Where

a_j	independent, experimentally controllable
c	dependent, experimentally observed

.

Regression is the process of relating the dependent and independent variables using an underlying function. This underlying function has an argument of weights $\vec{w}$ such that:

$$c \approx F(a_j; \vec{w}).$$

where:

c	actual value
$F\left(a_j; \vec{w}\right)$	model value
$\vec{w}$	model parameters

.

A critical part of regression is residual error which is the difference between the *actual* value the model value. The residual error depends on $\vec{w}$ as data a_j are fixed. Thus the residual error is defined as:

$$e_i(\vec{w}) =^{def} c_i - F(a_j; \vec{w}).$$

we can gather these individual errors for each observation into a error vecotr:

$$\vec{e}(\vec{w}) =^{def} \begin{bmatrix} e_1(\vec{w}) \\ e_2(\vec{w}) \\ \vdots \\ e_m\left(\vec{w}\right) \end{bmatrix}.$$

Thus each entry in the error vector depends on the same vector $\vec{w}$. Therefore we have all the errors for all the observations from 1 to m

b Concepts in Linear Regression

We shall now consider regression in the context of Linear Analysis.
Consider the Ordinary Least Squares Problem (OLS). The idea of this problem is to minimize an objective function of the error. Om objective function matches a vector input to a scalar output.
The squared error E_2 is defined as

$$E_2(\vec{w}) =^{def} \sum_{i=1}^{m} (e_i(\vec{e}))^2 = ||\vec{e}(\vec{w})||^2.$$

The geometry of this is: We want a vector $\vec{e}$ with the smallest normal. To find this we need a linear model of our data:

$$F(a_j; \vec{w}) =^{def} a_1 w_1 + a_2 + w_2 + \cdots + a_n w_n = \vec{a}^\top \vec{w}.$$

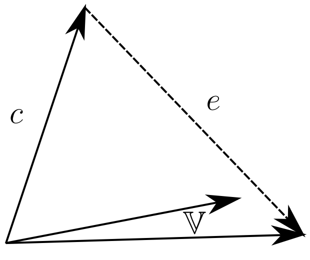
Where the Observation Residual Error is:

$$e_i(\vec{w}) =^{def} c_i - \vec{a}_i \vec{w}.$$

$$\vec{e}(\vec{w}) = \vec{c} - A\vec{w}.$$

We can solve this by projecting $\vec{c}$ to $\mathbb{V}$ spanned by the columns of A.
We were doing this earlier in lesson 11, thus the solution is the normal equation

$$[A^\top A]\vec{w} = A^\top \vec{c}.$$



c Examples of Linear Regression

Examples of some data that can be used for linear regression:

- data that appear to be in a 1D vector space (as discovered by Robert Hooke). This is to say that data passing through the origin satisfies Hooke's Law.

  A mathematical for this type of data is a constant slope:

  $$F(a, w) = Wa.$$

  The residual error is thus:

  $$e_i(w) = c_i - wa_i.$$

  Therefore the normal equation of the projection is:

  $$w = \frac{\vec{a}^\top \vec{c}}{\vec{a}^\top \vec{a}}.$$

- data that appear to be 1D but have an intercept (rather than strictly passing through the origin) The model for this type of data is a first order polynomial:

  $$F(a_j, w) = w_1 a + w_2.$$

  where w_2 is the intercept term. Our approximation of this is:

  $$c_i \approx w_1 a + w_2.$$

  Each residual error is:

  $$e_i(\vec{w}) =^{def} c_i = w_1 a_i - w_2.$$

  We can now gather the vectors into:

  $$\vec{e}(\vec{w}) = \vec{c} - A\vec{w}.$$

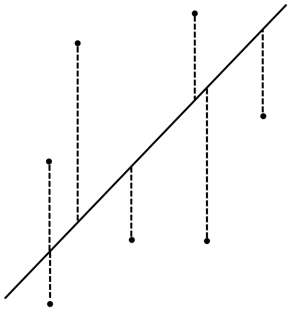
  $$A = \begin{bmatrix} a_1 & 1 \\ a_2 & 1 \\ \vdots & \vdots \\ a_n & 1 \end{bmatrix}.$$

  Which we can then solve using the normal equation:

  $$[A^\top A]\vec{w} = A^\top \vec{c}.$$

  we can then solve this equation for the weights of our model $\vec{w}$

Visually, Residual Error in Linear Regression is the 'vertical' error



d Data Standardization for Linear Regression

Consider a simple set of data where m data is in one columns.
Use the intercept term.
Thus, the data matrix is $A = \begin{bmatrix}\vec{a} & \vec{1}\end{bmatrix}$
Find the linear regression

$$A\vec{w} \approx \vec{c}.$$

We must first get the zero mean of the data:

$$M = \begin{bmatrix}\vec{m} & \vec{0}\end{bmatrix}.$$

the standardized form of our $\vec{c}$ is thus:

$$\vec{b} = \vec{c} - \overline{c}\vec{1}.$$

given that there is a $\vec{0}$ in M, do we have $M\vec{v} \approx b$?
This would be:

$$\begin{bmatrix}\vec{m} & \vec{0}\end{bmatrix}\begin{bmatrix}v_1 \\ v_2\end{bmatrix} = v_1\vec{m} + v_2\vec{0}.$$

Which thus means that we **cannot** determine v_2!

To solve this issue we need a simpler model; as when we include an intercept term and simplify to zero mean, the intercept term is ‘too complicated’ and we should instead do e the following model:

$$\vec{m}u \approx \vec{b}.$$

Which needs to be the same as

$$\begin{aligned} \vec{a}w_1 + w_2 &\approx \vec{c} \\ \vec{m}u &\approx \vec{c} \\ [\vec{a} - \overline{a}\vec{1}]u &\approx \vec{c} - \overline{c} \\ \vec{a}u + [-a\vec{1}u] &\approx \vec{c} - \overline{c}\vec{1} \\ \vec{a}u + [\overline{c}\vec{1} - \vec{a}\vec{1}u] &\approx \vec{c}. \end{aligned}$$

Thus, $w_1 = u$ and the intercept is equal to:

$$w_2 = \overline{c}\vec{1} - \overline{a}\vec{1}u.$$

This thus means that

$$\vec{x} = \frac{\vec{m}}{\sigma_m}.$$

$$\vec{y} = \frac{\vec{b}}{\sigma_b}.$$

Where $\vec{x}$ is the zero mean unit variance form of our original data and $\vec{y}$ is the standardized *dependent* observations. Given these, we want to equate:

$$\begin{aligned} \vec{m}u &\approx \vec{b} \\ \vec{x}z \approx \vec{y}. \end{aligned}$$

where $\vec{b}$ is the zero mean dependent data Thus we can say:

$$\begin{aligned} \vec{x}z &\approx \vec{y} \\ \frac{\vec{m}}{\sigma_m}z &\approx \frac{\vec{b}}{\sigma_b} \\ \frac{\sigma_b\vec{m}}{\sigma_m}z7 \approx \vec{b} \\ . \end{aligned}$$

And we must find

$$u = \frac{\sigma_b}{\sigma_m}z.$$

e Residual Error in Linear Regression

IN order to asses regression one must use the same measure of residual error that was minimized. This is typically called the 'root mean square' (RMS) error.

$$RMS(A, \vec{c}; \vec{w}) =^{def} \sqrt{\frac{e_1^2 + e_2^2 + \cdots + e_m^2}{m}} = \frac{||\vec{e}(\vec{w})||}{\sqrt{m}}.$$

Usually: use all the data when evaluating a method. Although it may be useful to keep some data to test, after regression.
If all data is independent within regression, then the linear regression becomes explanation (Principle Component Analysis).
If data is linearly independent but not orthogonal. Linear regression is projection thus:

- Uses column space $\mathbb{V}$ of matrix A
- projects dependents $\vec{c}$ to $\vec{p} \in \mathbb{V}$
- in normal equation, regression inverts the matrix $[A^\top A]$

A practical problem in this situation is however, that $[A^\top A$ is often ill conditioned typically because the SVD is of A has a large $\frac{\sigma_1}{\sigma_r}$. A possible solution for this situation is to use an orthogonal or orthonormal basis $\vec{q_i}$ of $\mathbb{V}$

Learning Outcomes

students should now be able to:

- solve linear regression with the normal equation
- use zero mean data in linear regression
- use standardized data in linear regression
- find the error of a linear regression