tool monkey   About Erik Erlandson

Deriving an Incremental Form of the Polynomial Regression Equations

Jul 5, 2012

Incremental, or on-line, algorithms are increasingly popular as data set sizes explode and web enabled applications create environments where new data arrive continuously (that is, incrementally) from clients out on the internet.

Recently I have been doing some experiments with applying one of the *oldest* incremental algorithms to the task of rating predictions: computing a linear regression with a coefficient of correlation. The incremental formulae look like this:

To find coefficients a_0, a_1 of the linear predictor $y = a_0 + a_1 x$:

$$a_1 = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} \qquad a_0 = \frac{\Sigma y - a_1 \Sigma x}{n}$$

The correlation coefficient of this predictor is given by:

$$\rho(x, y) = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

As you can see from the formulae above, it is sufficient to maintain running sums

$$n, \Sigma x, \Sigma y, \Sigma x^2, \Sigma y^2, \Sigma xy$$

and so any new data can be included incrementally - that is, the model can be updated without revisiting any previous data.

Working with these models caused me to wonder if there was a way to generalize them to obtain incremental formulae for a quadratic predictor, or generalized polynomials. As it happens, there is. To show how, I'll derive an incremental formula for the coefficients of the quadratic predictor:

$$y = a_0 + a_1 x + a_2 x^2$$

Recall the matrix formula for polynomial regression:

$$\vec{a} = \left(X^T X\right)^{-1} X^T \vec{y}$$

where, in the quadratic case:

$$\vec{a} = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ & \vdots & \\ 1 & x_n & x_n^2 \end{pmatrix} \qquad \vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Note that we can apply the definition of matrix multiplication and express the two products $X^T X$ and $X^T \vec{y}$ from the above formula like so:

$$X^T X = \begin{pmatrix} n & \Sigma x & \Sigma x^2 \\ \Sigma x & \Sigma x^2 & \Sigma x^3 \\ \Sigma x^2 & \Sigma x^3 & \Sigma x^4 \end{pmatrix} \qquad X^T \vec{y} = \begin{pmatrix} \Sigma y \\ \Sigma xy \\ \Sigma x^2 y \end{pmatrix}$$

And so now we can express the formula for our quadratic coefficients in this way:

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} n & \Sigma x & \Sigma x^2 \\ \Sigma x & \Sigma x^2 & \Sigma x^3 \\ \Sigma x^2 & \Sigma x^3 & \Sigma x^4 \end{pmatrix}^{-1} \begin{pmatrix} \Sigma y \\ \Sigma xy \\ \Sigma x^2 y \end{pmatrix}$$

Note that we now have a matrix formula that is expressed entirely in sums of various terms in x and y, which means that it can be maintained incrementally, as we desired. If you have access to a matrix math package, you might very well declare victory right here, as you can easily construct these matrices and do the matrix arithmetic at will to obtain the model coefficients. However, as an additional step I applied sage to do the symbolic matrix inversion and multiplication to give:

$$a_0 = \frac{1}{Z}\left(-\left(\Sigma x^3 \Sigma x - \left(\Sigma x^2\right)^2\right)\Sigma x^2 y + \left(\Sigma x^4 \Sigma x - \Sigma x^3 \Sigma x^2\right)\Sigma xy - \left(\Sigma x^4 \Sigma x^2 - \left(\Sigma x^3\right)^2\right)\Sigma y\right)$$

$$a_1 = \frac{1}{Z}\left(\left(n\Sigma x^3 - \Sigma x^2 \Sigma x\right)\Sigma x^2 y - \left(n\Sigma x^4 - \left(\Sigma x^2\right)^2\right)\Sigma xy + \left(\Sigma x^4 \Sigma x - \Sigma x^3 \Sigma x^2\right)\Sigma y\right)$$

$$a_2 = \frac{1}{Z}\left(-\left(n\Sigma x^2 - \left(\Sigma x\right)^2\right)\Sigma x^2 y + \left(n\Sigma x^3 - \Sigma x^2 \Sigma x\right)\Sigma xy - \left(\Sigma x^3 \Sigma x - \left(\Sigma x^2\right)^2\right)\Sigma y\right)$$

where:

$$Z = n\left(\Sigma x^3\right)^2 - 2\Sigma x^3 \Sigma x^2 \Sigma x + \left(\Sigma x^2\right)^3 - \left(n\Sigma x^2 - \left(\Sigma x\right)^2\right)\Sigma x^4$$

Inspecting the quadratic derivation above, it is now fairly easy to see that the general form of the incremental matrix formula for the coefficients of a degree-m polynomial looks like this:

$$\begin{pmatrix} a_0 \\ a_1 \\ \end{pmatrix} = \begin{pmatrix} n & \Sigma x & \cdots & \Sigma x^m \\ \Sigma x & \Sigma x^2 & \cdots & \Sigma x^{m+1} \\ \end{pmatrix}^{-1} \begin{pmatrix} \Sigma y \\ \Sigma xy \\ \end{pmatrix}$$

$$
\left(\begin{array}{c} \vdots \\ a_m \end{array} \right)^{-} \left(\begin{array}{cccc} \vdots & & \ddots & \vdots \\ \Sigma x^m & \Sigma x^{m+1} & \cdots & \Sigma x^{2m} \end{array} \right) \left(\begin{array}{c} \vdots \\ \Sigma x^m y \end{array} \right)
$$

Having an incremental formula for generalized polynomial regression leaves open the question of how one might generalize the correlation coefficient. There is such a generalization, called the coefficient of multiple determination, which is defined:

$$
r = \sqrt{\vec{c}^T R^{-1} \vec{c}}
$$

Where

$$
\vec{c} = \left(\begin{array}{c} \rho(x, y) \\ \rho(x^2, y) \\ \vdots \\ \rho(x^m, y) \end{array} \right) \qquad R = \left(\begin{array}{cccc} 1 & \rho(x, x^2) & \cdots & \rho(x, x^m) \\ \rho(x^2, x) & 1 & \cdots & \rho(x^2, x^m) \\ \vdots & & \ddots & \vdots \\ \rho(x^m, x) & \rho(x^m, x^2) & \cdots & 1 \end{array} \right)
$$

and $\rho(x, y)$ is the traditional pairwise correlation coefficient.

But we already have an incremental formula for any pairwise correlation coefficient, which is defined above. And so we can maintain the running sums needed to fill the matrix entries, and compute the coefficient of multiple determination for our polynomial model at any time.

So we now have incremental formulae to maintain any polynomial model in an on-line environment where we either can't or prefer not to store the data history, and also incrementally evaluate the 'generalized correlation coefficient' for that model.

Readers familiar with linear regression may notice that there is also nothing special about polynomial regression, in the sense that powers of x may also be replaced with arbitrary functions of x, and the same regression equations hold. And so we might generalize the incremental matrix formulae further to replace products of powers of x with products of functions of x:

for a linear regression model $y = a_1 f_1(x) + a_2 f_2(x) + \cdots + a_m f_m(x)$:

$$
\left(\begin{array}{c} a_1 \\ a_2 \\ \vdots \\ a_m \end{array} \right) = \left(\begin{array}{cccc} \Sigma f_1(x) f_1(x) & \Sigma f_1(x) f_2(x) & \cdots & \Sigma f_1(x) f_m(x) \\ \Sigma f_2(x) f_1(x) & \Sigma f_2(x) f_2(x) & \cdots & \Sigma f_2(x) f_m(x) \\ \vdots & & \ddots & \vdots \\ \Sigma f_m(x) f_1(x) & \Sigma f_m(x) f_2(x) & \cdots & \Sigma f_m(x) f_m(x) \end{array} \right)^{-1} \left(\begin{array}{c} \Sigma y f_1(x) \\ \Sigma y f_2(x) \\ \vdots \\ \Sigma y f_m(x) \end{array} \right)
$$

The coefficient of multiple determination generalizes in the analogous way.

tool monkey

tool monkey
erikerlandson@yahoo.com

erikerlandson

manyangled

adventures of an unfrozen caveman programmer