R^2 is rescaled mean squared error

Brendan O'Connor

September 3, 2009

R^2 – "the coefficient of determination" – is a rescaling of MSE (relative to the dataset in question). Alternative definitions are (1) the regression's proportion of total sum of squares, or (2) the squared correlation between predictions and responses.

Setup: items x_i and we're targeting real-valued responses y_i by fitting a function $f(x)$. Let's be vague on training vs. test sets; all that matters is we want to evaluate the prediction function's accuracy on these items. Then

$$MSE = \sum_i (f(x_i) - y_i)^2 / N$$

Let's use definition #1 of $(1-R^2)$, that it's the "sum of squared error divided by total sum of squares". These terms are

- $SS_{tot} = (y_i - E[y])^2$: total sum of squares, which is a rescaling of response variance
- $SS_{err} = (y_i - f(x_i))^2$: sum of squared errors, a.k.a. "residual sum of squares", a rescaling of the model's predictions' MSE

So we have:

$$\begin{aligned} 1 - R^2 &= \frac{SS_{err}}{SS_{tot}} \\ &= \frac{\sum (y_i - f(x_i))^2 \quad / N}{\sum (y_i - E[y])^2 \quad / N} \\ &= \frac{MSE}{Var(y)} \\ &= \frac{(Mean) SqErr\ of\ predictions}{(Mean) SqErr\ of\ guessing\ the\ mean} \end{aligned}$$

R^2 can be thought of as a rescaling of MSE, comparing it to the variance of the outcome response.

It's nice to interpret because it's bounded between 0 and 1. Higher is better.

If MSE=0, then $R^2 = 100\%$: you have perfect predictions.

If MSE is as bad as just guessing the mean for everything, then $R^2 = 0\%$: about as bad as possible.

In fact, on your training data, if you fit a linear regression with a bias term, it's impossible to go below $R^2 = 0$.[1]

But on held-out data, it's possible to go lower than 0 if you're overfitting. I have an amusing text regression example I can show anyone who's interested.

I think it's useful to talk about held-out R^2 simply because of its intuitive scale.

You can think about it analgously to accuracy in the discrete case. Accuracy is the proportion of response labels the model gets right. R^2 is the proportion of response variance the model captures. (Or "explains", as some say.)[2]

Now, there's a second view that apparently is more popular at least among people in our class, that R^2 is the squared correlation between predictions and response. On the 4th Wikipedia page below, there's a derivation for how this is equivalent to the MSE view above, but it only applies to R^2 on the training data.

Relevant Wikipedia pages this all came from:

- http://en.wikipedia.org/wiki/Fraction_of_variance_unexplained
- http://en.wikipedia.org/wiki/Coefficient_of_determination
- http://en.wikipedia.org/wiki/Correlation
- http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

[1] I think this is true for fitting any function that has a bias term when you're trying to minimize squared error. Guessing the mean for everything is the least-squares solution for a "linear" model that only has a bias term; i.e., the maximum likelihood solution for a constant plus i.i.d. gaussian noise. If you care about measuring squared error, it's hard to imagine a fair but crappier baseline than guessing the mean, since you can always get infinitely bad MSE by guessing infinitely far away. You might as well take the best of the crappy "guess a constant everywhere" family of baselines.

[2] Though to complete the analogy of competition against a constant baseline, in the discrete case it should be the model's accuracy versus a guess-the-most-common baseline. I'm not sure how to make it such a ratio scale nicely from 0-ish to 1; maybe logs have to be involved. Like some sort of KL divergence or information gain or something measure.