Interpretation of Correlation Coefficient of Two 0-1 Sequences

Happy Rabbit

December 20, 2010

1 Interpretation

Some notes:

0 negative

1 positive

x_i estimated value (0 or 1)

y_i true value (0 or 1)

n total number of the observations

It turned out that the correlation coefficient can be represented as:

$$\rho = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

When the estimated values are the same with real values, both FP and FN are 0. In this case, ρ equals 1.

When the all the estimated values are inconsistent with real values, both TP and TN are 0. In this case, ρ equals -1.

So the correlation coefficient can be used as a Measure of Accuracy.

Table 1: Number of Different Combination

	Estimated Positive	Estimated Negative
Real Positive	True Positive (TP)	False Negative (FN)
Real Negative	False Positive (FP)	True Negative (TN)

2 Glorious Derivative Details

The correlation coefficient:

$$\rho = \frac{Exy - ExEy}{\sqrt{Var(x)}\sqrt{Var(y)}} \tag{1}$$

$$n = TP + FN + TN + FP \tag{2}$$

$$Exy = \frac{\sum_{i=1}^{n} x_i y_i}{n} \tag{3}$$

$$Ex = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4}$$

$$Ey = \frac{\sum_{i=1}^{n} y_i}{n} \tag{5}$$

$$Var(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \tag{6}$$

$$Var(y) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} \tag{7}$$

$$\sum_{i=1}^{n} x_i y_i = \#\{x_i = 1, y_i = 1 \mid 1 \leq i \leq n\} = TP \tag{8}$$

$$\sum_{i=1}^{n} x_i = \#\{x_i = 1, y_i = 0\} + \#\{x_i = 1, y_i = 0\} = FP + TP \tag{9}$$

$$\sum_{i=1}^{n} y_i = \#\{x_i = 1, y_i = 1\} + \#\{x_i = 0, y_i = 1\} = TP + FN \tag{10}$$

$$\begin{aligned}
Exy - ExEy &= \frac{\sum_{i=1}^{n} x_i y_i}{n} - \frac{\sum_{i=1}^{n} x_i}{n}\frac{\sum_{i=1}^{n} y_i}{n} \\
&= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n^2} \\
&= \frac{(TP + FN + TN + FP) \times TP - (FP + TP)(TP + FN)}{n^2} \\
&= \frac{TN \times TP - FN \times FP}{n^2}
\end{aligned}$$

$$\begin{aligned}
Var(x) &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \\
&= \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}}{n} \\
&= \frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n^2} \\
&= \frac{n(FP + TP) - (FP + TP)^2}{n^2} \\
&= \frac{(FP + TP)(FN + TN)}{n^2}
\end{aligned}$$

Similarly:

$$Var(y) = \frac{(TP + FN)(TN + FP)}{n^2}$$

So

$$\rho = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \tag{11}$$