Weka Error Measurements

When the class value is nominal, the kappa statistic is given. When the class value is numeric, the correlation coefficient is given.

Kappa statistic – chance-corrected measure of agreement between the classifications and the true classes. Calculate by taking the agreement expected by change away from the observed agreement and dividing by the maximum possible agreement. A value greater than 0 means that your classifier is doing better than chance.

The equation for κ is:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

where po is the relative observed agreement among raters, and pe is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then κ = 1. If there is no agreement among the raters other than what would be expected by chance (as given by pe), $\kappa \leq 0$.

A kappa value of 0 means that the result is the same as would be expected by chance.

Example:
Suppose that you were analyzing data related to a group of 50 people applying for a grant. Each grant proposal was read by two readers and each reader either said "Yes" or "No" to the proposal. Suppose the dis/agreement count data were as follows, where A and B are readers, data on the diagonal slanting left shows the count of agreements and the data on the diagonal slanting right, disagreements:

		B	
		Yes	No
A	Yes	20	5
	No	10	15

Note that there were 20 proposals that were granted by both reader A and reader B, and 15 proposals that were rejected by both readers. Thus, the observed proportionate agreement is po = (20 + 15) / 50 = 0.70

To calculate pe (the probability of random agreement) we note that:

- Reader A said "Yes" to 25 applicants and "No" to 25 applicants. Thus reader A said "Yes" 50% of the time.
- Reader B said "Yes" to 30 applicants and "No" to 20 applicants. Thus reader B said "Yes" 60% of the time.

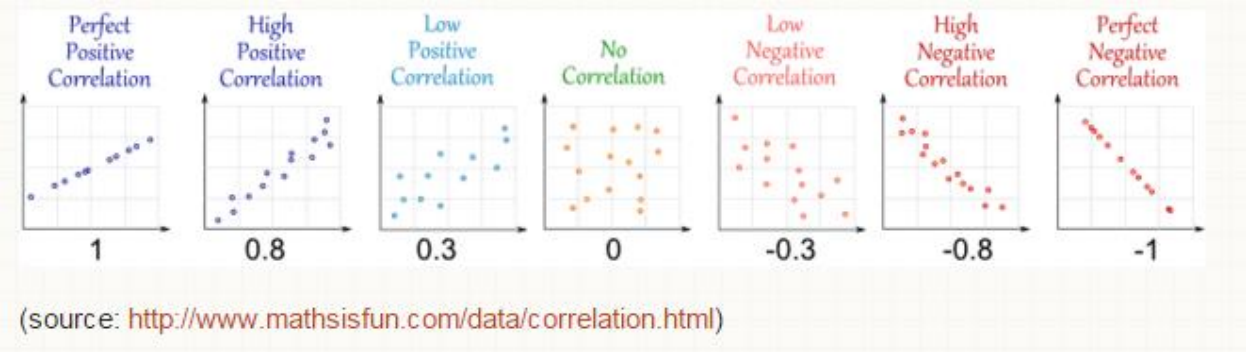
Therefore the probability that both of them would say "Yes" randomly is 0.50 · 0.60 = 0.30 and the probability that both of them would say "No" is0.50 · 0.40 = 0.20. Thus the overall probability of random agreement is Pr(e) = 0.3 + 0.2 = 0.5.

So now applying our formula for Cohen's Kappa we get:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.70 - 0.50}{1 - 0.50} = 0.40$$

From https://en.wikipedia.org/wiki/Cohen%27s_kappa

Correlation coefficient – measure of correlation and dependence between observed values. The values range between -1 and 1, where 0 is no relation, 1 is a very strong relation and -1 is an inverse relation (see the following):



From https://en.wikipedia.org/wiki/Correlation_coefficient

Mean absolute error – quantity used to measure how close forecasts or predictions are to the eventual outcomes.

The mean absolute error is given by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|\,.$$

As the name suggests, the mean absolute error is an average of the absolute errors $|e_i| = |f_i - y_i|$, where f_i is the prediction and y_i the true value. Note that alternative formulations may include relative frequencies as weight factors.

From https://en.wikipedia.org/wiki/Mean_absolute_error

The mean absolute error is like the variance, but rather than square the difference, use its absolute value. (If the scores are spread closely around the mean, the variance will be smaller than the mean absolute error. If the scores are not spread closely, squaring the distance will lead to larger variances. Taking the absolute value assigns equal weight to the spread of data whereas squaring emphasizes the extremes. Squaring, however, makes the algebra easier to work with and relates to Pythagorean Theorem.)

Root mean square error - measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. It represents the sample standard deviation of the differences between predicted values and observed values. It aggregates the magnitudes of the errors in predictions for various times into a single measure of predictive power. It is a good measure of accuracy, but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale-dependent. It is also called the root-mean square deviation, RMSD

The RMSD of predicted values $\hat{y}_t$ for times t of a regression's dependent variable y is computed for n different predictions as the square root of the mean of the squares of the deviations:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y)^2}{n}}.$$

From https://en.wikipedia.org/wiki/Root-mean-square_deviation

Difference between relative and absolute - Absolute error is how much your result deviates from the real value. Relative error is a measure in percent compared to the real value.

Relative absolute error
Root relative square error

Area under ROC curve (Receiver Operating Characteristic – since it came from signal processing, AUC) – In the result section of applying a classification algorithm, Weka gives a ROC Area for each class value (look under "Detailed Accuracy by Class" and scroll to the right). The ROC area measures discrimination, that is, the ability of the test to correctly classify those with and without the disease. Consider the situation in which patients are already correctly classified into two groups. You randomly pick one from the disease group and one from the non-disease group. Run the classifier on both. The area under the curve is the percentage of randomly drawn pairs where both are correctly classified. Computing the area is difficult to explain and beyond the scope of this introductory material. Two methods are commonly used: a non-parametric method based on constructing trapeziods under the curve as an approximation of area and a parametric method using a maximum likelihood estimator to fit a smooth curve to the data points.

A ROC area of 1 represents a perfect test; an area of .5 represents a worthless test. A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system:

.90-1 = excellent (A)
.80-.90 = good (B)
.70-.80 = fair (C)
.60-.70 = poor (D)
.50-.60 = fail (F)