Economic Efficiency and the Weighted Kappa loss function

Description of weighted kappa loss function

The weighted kappa loss function is used when the dependent variable is categorical and ordinal. It measures the squared distance from the true value, and is discounted by the expected squared loss if categories were randomly assigned.

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}$$

The weight matrix, w, is calculated on the squared difference between the categories, discounted by number of categories:

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^s}$$

The further apart I and j are, the more weight on that discrepancy. Weights along the diagonal (where i=j) are zero.

The histogram matrix O is a count of each true vs. predicted rating, with I being the 'true' rating and B being the predicted rating, It's an NxN matrix with the following entry at each location I,j:

$$O_{i,j} = count\ (true\ rating\ i, predicted\ rating\ j)$$

If the prediction is 100% accurate, the matrix will have positive values along the diagonal, and zero everywhere else. In this case the weight is zero, so the sum of weight times loss = 0, and $\kappa = 1$

Finally, the weighted kappa loss function is discounted by the expected prediction. This takes into account the effect of varying distributions across categories. If most of the observations are at the extremes, the histogram matrix O multiplied by weights is likely greater than if the observations are uniformly distributed across categories.

The expected prediction matrix is the outer product of the actual histogram vector, and the predicted histogram vector. Allowing the histogram vector to be {count(i=1) , . . . , count(i=N)) the raw entry $E_{i,j}$ would be count(actual rating i) * count(predicted rating j). This matrix is normalized such that matrices E and O have the same sum

$$E_{i,j} = count(actual\ rating\ i) * count(predicted\ rating\ j) * \frac{\sum_i \sum_j O_{i,j}}{\sum_i \sum_j E_{i,j}}$$

Thus if the distribution is heavily skewed towards actual ratings of 1 and N, the matrix E will take this into account, discounting the loss function from the histogram matrix O. This expected prediction matrix changes as the count of predicted ratings j changes.

Application of Weighted Kappa Loss function:

With a continuous dependent variable the standard loss function is least squares, the objective being to minimize the error term, $(y - \Theta X)^2$, this puts a greater penalty on observations whose predicted value is further from the actual value. The goal becomes getting predicted values sufficiently close to the actual value.

With the Weighted Kappa Loss function the squared loss term shows up in the weight matrix, penalizing observations further from the actual category. Weights on the diagonal are zero, and weights increase exponentially as we move away from the diagonal. Multiplying this by our histogram matrix O, we find that inaccurate predictions are more costly the further they are from the diagonal.

However, the addition of the expected prediction matrix means that predicted distributions with the same error rate, and the same summation from their weighted histogram matrix O have different kappas: (the columns are predicted values, the rows are actual values)

Prediction 1:
Histogram Matrix O

	1	2	3	4
1	100	0	0	0
2	30	70	0	0
3	0	0	70	30
4	0	0	0	100

Kappa = 0.9464

Prediction 2:
Histogram Matrix O

	1	2	3	4
1	90	10	0	0
2	10	80	10	0
3	0	10	80	10
4	0	0	10	90

Kappa 0.9400

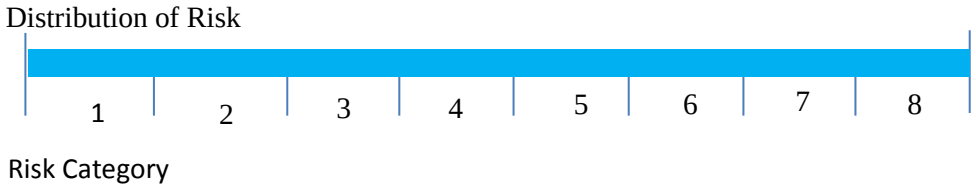
The difference between these two lies in the expected prediction matrix, with more weight on the extreme values (1 and 4) in the first example, the sum of the weighted expected prediction matrix increases, increasing the kappa value. Thus by using the kappa method a fit that biases the results towards the extremes rather than symmetrically will have a higher kappa score, and therefore be deemed 'better'.

From an economic perspective there is no reason to assume that the first prediction is better than the second. Each prediction is accurate 85% of the time, and off by one rank 15% of the time. At least prediction 2 has the same distribution as prediction 1, so the errors are symmetric.

Application to Insurance Markets:

In the Kaggle Prudential Insurance markets the weighted kappa loss function was used for evaluation of results. While this loss function is easy to implement, and straight forward in its interpretation, it's not clear that it's the welfare-maximizing loss function for Prudential.

As an example, suppose that actual risk is distributed uniformly across the population, and that the risk categories break the distribution up into equal size chunks. Therefore the risk interval for category 1 is the same size as the risk interval for category 2, etc.



Therefore, allowing individual's actual risk to vary from 0 to 8, risk category one contains individuals with actual risk between 0 and 1, risk category two contains individuals with actual risk between 1 and 2, and so on.

Now, suppose that the expected cost of insuring a person is linear in their true risk factor, and due to risk aversion, Prudential would optimally offer a category-specific insurance price such that everyone will buy insurance if they were able to accurately predict risk.

Now, let's look at the expected cost of prediction error. If the difference between price and expected cost is the same for each risk category, an inaccurate prediction has one of two possible results. If their predicted risk category is below their actual risk, they will buy insurance, and Prudential will lose the difference between the prices of the two categories. If their predicted risk category is above their actual risk, the individual will choose not to buy insurance, and Prudential will lose price of category i – expected cost of category i.

It is not clear that these losses are symmetric, with a sufficiently competitive insurance sector the loss in the difference between the prices will likely be greater than the difference between the price and expected cost within a category. Therefore more weight should be put on a prediction that underestimates risk rather than a prediction that overestimates risk.

Furthermore, if the distribution of risk types is not uniform, but follows a normal distribution, the effects of a bad prediction are even larger. In this case if the categories

each have an equal proportion of individuals, the tails will have a much larger standard deviation in expected costs than the categories in the middle.

This will lead to prices increasing quickly at either end of the distribution, and minimal price differences in the middle. Now the cost of a prediction error is greater at the extremes than in the middle of the distribution. Using a weighted kappa variable, as shown in the examples above, the 'better' model is the one that mis-labels people to the extremes, and with the large differences in expected costs at the tails, this model may be more costly in terms of revenue – costs than prediction 2.