Economic Efficiency and the Weighted Kappa Loss Function

Jennifer Lynn Steele

2016/06/28

1 Introduction

In the Prudential Life Insurance Kaggle competition models were fitted and scored using a weighted kappa loss function, a common loss function when using a categorical prediction. This function is easy to implement but it is unlikely to minimize the costs for prudential from an economic perspective.

These loss functions, like the least squares function for continous variables, evolved from a need to create models that were interpretable. In economics the focus is not on prediction, but rather on interpreting the model. In this case fitting a model needs to be fairly objective and free of bias, so that the interpretations are as close to 'true' as possible. With prediction models we no longer care about interpreting the model. In many cases interpretation is not possible, or in very difficult (for example XGBoost and neural network implementations). In these cases we create very complex models that treat every error in an equal prescribed manner. This may not be optimal, especially when the expected costs associated with errors differ across categories, or areas of the distribution of true values.

In this note I first review the weighted kappa loss function, including some issues with it, and then I look at a basic economic model, thinking about what the cost-minimizing mechanism might look like, and what sort of loss function would implement this optimal mechanism when there is noise. Finally I run a simulation looking at the difference between using the weighted kappa loss function and the optimal mechanism.

2 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. The Kappa coefficient was introduced by Cohen in 1960 with a weighted version introduced in 1968. The weighted version is useful for ordinal categorical data, as it allows us to penalize a prediction based on how far it is from the real value, rather than just whether or not it is correct.
The kappa function is as follows:

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

Kappa is calculated based on the number of observations with true categorical rating, i, and predicted categorical rating j.

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

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

The further apart i and j are, the more weight on that discrepancy. Weights along the diagonal are zero, and the weights are symmetrical (i.e. $w_{i,j} = w_{j,i}$)

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

$$O_{i,j} = count(trueratingi, predictedratingj)$$

If the prediction is 100% accurate the matrix will have positive values along the diagonal, and zero everywhere else. The weight on diagonal entries is zero, so the kappa is $1 - 0$ or 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 is the outer product of the true histogram vector, and the predicted histogram vector. A histogram vector is $\{count(i = 1), ..., count(i = N)\}$. This matrix is normalized such that matrices E and O have the same sum:

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

This if the distribution is heavily skewed towards true 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.

3 Drawbacks of the Weighted Kappa Loss Function

3.1 Identical error rates, but different expected prediction matrices

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

With the Weighted Kappa Loss function the squared loss term shows up in the weight matrix, increasing the penalties exponentially as the distance between the predicted and true values increases.

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 (here the columns are predicted values, the rows are true values):

Prediction 1: Histogram Matrix O_1, $\kappa_1 = 0.9464$

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

Prediction 2: Histogram Matrix O_2, $\kappa_2 = 0.9400$

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

These two matrices each have 100 true observations in each category, but they differ in the predictions. They each have 340 correct predictions, and 60 incorrect predictions, each of which differ from the true

rating by one category. However, using the weighted kappa loss function we find that the first prediction is 'better' as the kappa value is closer to 1. The difference in the kappas stems from the expected prediction matrix, with a larger summation across the expected matrix in the first case.

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. Prediction 2 has the same distribution as the true values, and is symmetric, so could even be considered a better predictor from that standpoint, at least it's not biased.

3.2 Squared loss penalty

The other thing to consider about the kappa loss function is the exponentially increasing penalty of mis-labelling. Continuing with the four category example, the weight matrix looks as follows:

	1	2	3	4
1	0	.11	.44	1
2	.11	0	.11	.44
3	.44	.11	0	.11
4	1	.44	.11	0

As the difference between the predicted and true ratings increases from 1 category to 2, the penalty increases by .33, quadrupling the penalty. Increasing further to a 3-category mistake, the weight increases by .56 to 1. This causes the model to get values 'close to' the true level rather than increasing the accuracy rates. See the following two examples to see how close misses are preferred to a higher accuracy rate with large misses:

Prediction 1: Histogram Matrix O_1, $\kappa_1 = 0.960$

	1	2	3	4
1	95	5	0	0
2	5	85	10	0
3	0	10	85	5
4	0	0	5	95

Prediction 2: Histogram Matrix O_2, $\kappa_2 = 0.948$

	1	2	3	4
1	95	3	2	0
2	3	92	3	2
3	2	3	90	3
4	0	2	5	95

For the first prediction, the accuracy rate is 90%, but the 'misses' are all by one category. The average error (including non-misses as 0) is 0.1. For the second prediction, the accuracy rate is 93% and the average error is .09. However, due to the squared loss function, the first prediction has a higher kappa, which would lead it to be chosen over the second prediction. This implies a very strict cost function, where the costs of mis-labelling are increasing in the same way as the error term. This is unlikely to be the case, as shown in the next section.

4 Economic Losses and Insurance Markets

In the Kaggle challenge the weighted kappa loss function was used for the evaluation of the results and selection of the best model. While this loss function is easy to implement and straight forward in its interpretation, it's unlikely to be the welfare-maximizing loss function for Prudential.

Setting aside for the moment the question of how the ratings are determined, and selection bias due to the price menu offered to consumers (and possible risk tolerances), we'll look first at a simple example.

4.1 The Assumption of Non-Linear Cost

As discussed earlier the exponentially increasing loss function biases the model towards near-misses rather than accuracy rates. This is important because at an equilibrium (the fitted model) the computer has considered all of the tradeoffs between the different variables, and the marginal effect they have on the loss function. With a non-linear loss function, a change that moves one observation from a 2-category miss to a 1-category miss has the same effect on the Kappa as a change that moves three observations from a 1-category miss to an accurate prediction. It is not clear that the expected cost of the mis-labelling is as punishing.

For example, consider a case where true risk is distributed uniformly along a line from 0 to 1, and the risk categories are equal segments of the line. In order for the weighted kappa to be an appropriate loss function, the expected cost of mis-labelling has to be increasing exponentially, If the true value is 4, and the loss associated with a predicted rating of 3 is \$100, the loss associated with a predicted rating of 5 would have to be approximately $\$100^2$. This would require the expected costs associated with any risk profile to follow the same general shape as the squared loss function.

4.2 Asymmetric Loss Profile

The second concern with the weighted kappa loss function is the symmetric treatment of inaccurate predictions. Assuming that expected costs are increasing with risk, there is no model where the expected losses from mis-labelling are symmetric. If all individuals buy insurance, then labelling an individual as more risky than their true value would increase Prudential's revenues (assuming the price of a policy is increasing with risk category). However labelling an individual as less risky than their true value would lower Prudential's revenues, with no effect on costs (their true value, and thus expected cost, remains unchanged).

What is more likely is that if an individual is mis-labelled as more risky than their true value, that individual will choose not to buy insurance, and a greater error will not increase the cost to Prudential, so mis-labelling above the true value may be beneficial for the first one or two categories, and then will fall to a constant loss as the individual chooses not to buy insurance. For the same individual, the expected cost of predicting a category below the true value will have increasing costs as the price of insurance falls while the expected cost remains the same.

For example, consider an individual again with a true risk category of 4, and for simplicity assume that Prudential's optimal pricing mechanism is price = risk category x 10, and the expected cost is risk category x 9. Also assume at this point that the individual will buy insurance at any price below \$45.

In this case if the individual is accurately predicted as a risk category of 4, Prudential's gain is 40 - 36 = 4. If the individual is predicted as any risk category above 4, they will not buy insurance, and Prudential loses \$4. For risk categories 3, 2, and 1, Prudential loses \$10, \$20, and \$30 respectively. It is clear that the assumption of symmetric loss penalties is not appropriate in this case.

4.3 Losses across the distribution

The final main concern with the use of the Weighted Kappa Loss Function is the identical treatment of mis-labelling regardless of the true value of the risk. A true value of 7 with a predicted value of 8 has the same effect as a true value of 1 with a predicted value of 2. It is unlikely that mis-labelling has the same effect across true values. The easiest way to see this is with an example.

Consider a population where risk is distributed normally, and the 8 risk categories are set up to be all identical in mass. This means that risk categories 1 and 8 encompass the tails of the distribution, and categories 4 and 5 have the smallest risk intervals. If the indivdual's true value is way out in the tail of 8, labelling them as a 7 is unlikely to have the same loss as labelling an individual with a true value of 5 as a 4.

5 Does the loss function matter?

If the model does a very good job predicting the true values, meaning both a high accuracy and small errors, the loss function is not a huge factor. However in datasets where there is lots of noise it is likely that the loss function will play a fairly large role in what the final set of predictions looks like. In the Kaggle competition the winning submission had a Kappa of 0.67909, suggesting that the loss function may play a large role in the final set of predictions.

In this section I will look at a simple linear regression model with regularization and see how the use of different loss functions affects the predictions.

still to come

References

[1] J. Cohen, Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.*, Psychological Bulletin, Vol 70(4), Oct 1968, 213-220.*

[2] J. Cohen, A Coefficient of agreement for nominal scales.*, Educational and Psychological Measurement, Vol 20(1): 37-46*

[3] Kaggle Prudential Life Insurance Assessment Competition (completed) `https://www.kaggle.com/c/prudential-life-insurance-assessment`