Accuracy Plots for Categorical Variables

Jared L. Deutsch and Clayton V. Deutsch

The cross validation of numerical models is an important step in any geostatistical study. For continuous variables, there are a number of choices for the display of cross validation results including accuracy plots and scatterplots. This note introduces a program, `accplt_cat`*, for the display of cross validation results from categorical variables. Accuracy plots are constructed on a by-category basis and overall. Relevant statistics, including the B value and Shannon H entropy are calculated to gauge the information contained in categorical variable estimates.*

Introduction

Cross validation is one of the most critical steps in a geostatistical modeling program. Cross validation can help tune modeling parameters and compare different algorithms or modeling strategies. This can be in the form of estimating statistics or data values using either 1) a subset of the data or 2) data not used in the generation of the geostatistical model. Ideally, cross validation would consider new data not used in making the numerical model, however this is not always practical.

The cross validation and display of cross validation results for continuous variables has been previously discussed (Deutsch, 2010). A program, `accplt_ns`, was introduced to construct an accuracy plot for the display of normal scored cross validation results. An accuracy plot bins the predicted probabilities into intervals such as 0.0 – 0.1, 0.1 – 0.2 to 0.9 – 1.0. The average of the predicted probabilities in these bins is then compared to the actual fraction of each category in the bin. For good predictions, the average predicted probability in each bin should be close to the occurrence of the category.

Indicator Formalism and Some Statistical Measures

Consider a categorical variable k with k=1,…,K categories. These categories are taken to be mutually exclusive and exhaustive. At every point, the data event i_k is 1 if category k is present and 0 if it is not. Although a categorical variable can only take a value of 0 or 1, estimates are continuous probabilities. The probability of category k being present is p_k where:

$$\sum_{k=1}^{K} p_k = 1 \text{ and } p_k \in [0,1] \text{ where } k = 1,...,K \tag{1}$$

When probabilistic estimates of a categorical variable are made for i=1,…,N data values, there are a number of useful statistics to determine the quality of estimates including the B value and Shannon H entropy. The B value is defined as the difference between the average predicted probability when the true value is 1 and the predicted probability when the true value is 0 (Deutsch, 2010). A high value indicates that the presence and absence of categories are being predicted correctly.

$$B = \frac{1}{\sum_{i=1}^{N} i_{i_k=1}} \sum_{i=1}^{N} \sum_{k=1}^{K} p_{k,i_k=1} \sum p_{k,i_k=1} - \frac{1}{\sum_{i=1}^{N} i_{i_k=0}} \sum_{i=1}^{N} \sum_{k=1}^{K} p_{k,i_k=0} \tag{2}$$

The Shannon H entropy, given here in nats (natural logarithm units) is a measure of the variability of all values (Eq. 3). The lowest possible entropy occurs when we know exactly what the value of a categorical variable will be at a certain location, ie: p_k = 1 for i_k = 1 and p_k = 0 for i_k = 0. This instance has an entropy value of zero. The maximum possible entropy occurs when we have no idea what the true value could be so p_k = 1/K for all k.

$$H = -\sum_{k=1}^{K} p_k \ln\left(p_k\right) \tag{3}$$

For $p_k = 0$, the term $p_k \ln(p_k) = 0$ which can be derived by taking the limit as p_k decreases to 0. Two quantities are reported in accplt_cat: the average H entropy (Eq. 4) and maximum possible H entropy which is calculated given the number of categories with $p_k = 1/K$ (Eq. 5).

$$H_{avg} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} p_k \ln\left(p_k\right) \tag{4}$$

$$H_{\max} = \sum_{k=1}^{K}\frac{1}{K}\ln\left(K\right) \tag{5}$$

A Bootstrap-Like Approach to Probability Intervals

The expected fraction in each interval is equal to the mean of the probability interval; however with either a low number of data in the interval or wide range of values taken in the interval, deviations are expected. A proposed bootstrap-like approach to quantify this expected deviation by the construction of a 90% probability interval. The algorithm calculates a 90% probability interval for each bin separately:

1. For each bin (probability interval), find the probability values in the bin
2. Generate a random number [0,1]. If the random number is lower than the probability value then $i^*_k = 1$, else $i^*_k = 0$
3. Calculate the fraction of these random indicators in the bin
4. After this has been done 1000 times, get the 5% and 95% quantiles to get the 90% probability interval

This calculation assumes that the probability values are unbiased estimates of the category and that probability values are independent. For the cross validation of a kriged categorical variable, the probability variables will be conditionally unbiased given a sufficiently large search (see (Deutsch and Deutsch, 2012), this report), however the probability estimates will not be independent. The 90% probability interval constructed with this algorithm is therefore a conservative estimate.

Implementation

The generation of accuracy plots considered all categories together (combined) and individually by category has been implemented in `accplt_cat`. The statistical measures and 90% probability interval calculation introduced are implemented in the program. A text file is also generated containing all of the data from the postscript plots. The true values and probabilities must be contained in the same data file; these may come from the GSLIB programs `kt3d` or `ik3d` in cross validation mode. The parameter file is given in Table 1.

Table 1: Parameter file for `accplt_cat`

Line	Output
1	Parameters for ACCPLT_CAT
2	************************
3	
4	START OF PARAMETERS:
5	ik3d_3lots.out -file with true values and cond. probs.
6	3 - number of categories
7	0 1 2 - category IDs
8	1 2 3 4 - columns for category probabilities, true
9	0.1 -probability increment (ie: 0.05 or 0.1)
10	1 69396 -bootstrap check (yes=1)?, random number seed
11	1 -show number of data in each bin
12	accplt_cat.ps -file for accuracy plot for combined
13	accplt_by_cat.ps -file for accuracy plot by category
14	accplt_cat.dat -file for text output

For categories with low frequencies, it is unlikely that there will be a large number of locations where the probability of the category is high. For this reason, the number of data in each bin is output at the top of the bin. This, in addition to the bootstrap-like 90% probability intervals provide a measure of how precise the accuracy measurement for a given probability interval is.

A sample accuracy plot considering all categories together is shown in Figure 1. This accuracy plot, constructed by indicator kriging of a large 2D domain in cross validation mode demonstrates that the extreme high and low probabilities are slightly under and over-estimated, respectively. Points which fall outside of the 90% probability interval are flagged. In addition, the number of bins which fall inside the 90% probability intervals are reported. The number of points in each probability interval are reported on top of the accuracy plot (2876, 178,…). The user has the option of turning this off in the parameter file. Accuracy plots are also generated on a by-category basis. A sample plot is shown in Figure 2.

There are a number of observations that can be gathered from these plots. When the occurrence of a category is very low, there are not enough points to construct intervals for higher probability values (Figure 2). For this reason, the higher probability intervals are omitted by `accplt_cat`. As the number of data points in the bin decreases, the 90% probability interval size increases as larger deviations are expected. In addition, there is no requirement on a categorical variable accuracy plot to monotonically increase or decrease. Depending on data configuration and category occurrence, points may fall above or below the 45° line. This is expected; only in the presence of a bias then points would consistently fall above or below the line.

A portion of the text output from `accplt_cat` is shown below (Table 2). The data from all plots generated by-category and overall, are included in this text output file.

Table 2: Portion of text output from `accplt_cat`

Line	Output
1	`------------------------------`
2	`Accuracy Plot for Category    0`
3	`------------------------------`
4	`Avg Probability =  0.015`
5	`True Fraction =  0.015`
6	`Number of P Intervals =     6`
7	`Avg when true=0 =  0.011`
8	`Avg when true=1 =  0.284`
9	`B value =  0.273`
10	`% inside 90% PI =  100.0`
11	
12	`Accuracy Plot Values`
13	`probmean,fracinint,ninbin, 0.05b , 0.95b`
14	`   0.001   0.002    1764   0.000   0.003`

Conclusion

The cross validation of numerical models is an essential modeling step. This note has introduced a program, `accplt_cat`, for the plotting and checking of cross validation results of categorical variables. A 90% probability interval is calculated using a bootstrap-like approach to quantify what level of deviation is reasonable. The calculated probability interval will be conservative as the calculation assumes independence of predicted probabilities, but is useful for approximating what level of deviation is expected.

References

Deutsch, C.V., 2010. Display of Cross Validation/Jackknife Results. Centre for Computational Geostatistics, 12:406.

Deutsch, J.L. and Deutsch, C.V., 2012. Kriging, Stationarity and Optimal Estimation: Measures and Suggestions. Centre for Computational Geostatistics, 14:306.

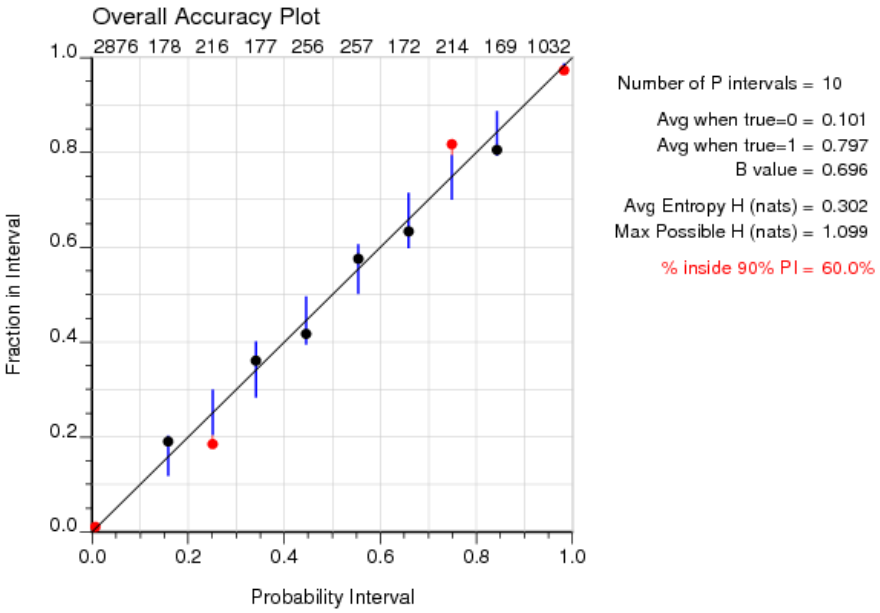

Figure 1: Sample overall accuracy plot when all categories are considered together.

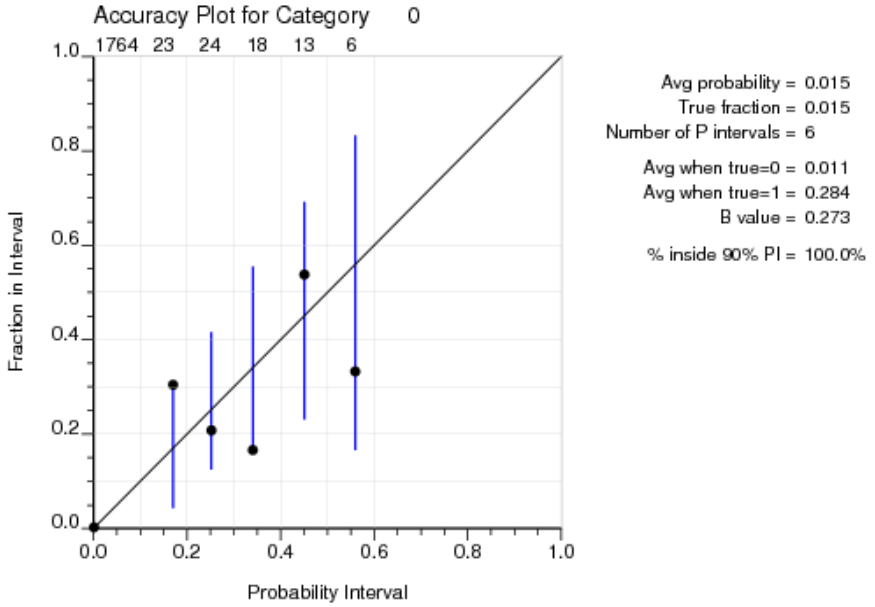

Figure 2: Sample accuracy plot for a single category. Accuracy plots on a by-category basis are included in the same postscript file, overflowing to a new page if more than 6 categories are present.