Image from NASA: https://commons.wikimedia.org/wiki/File:Earth_Western_Hemisphere_transparent_background.png#filelinks

Remote Sensing

Accuracy Assessment

WV VIEW

AmericaView
Empowering Earth Observation Education
AmericaView.org - Est. 2003

In the prior module, we discussed methods for producing classifications from remotely sensed data using unsupervised and supervised classification methods. In this module, we will focus on methods to assess the accuracy of these products.

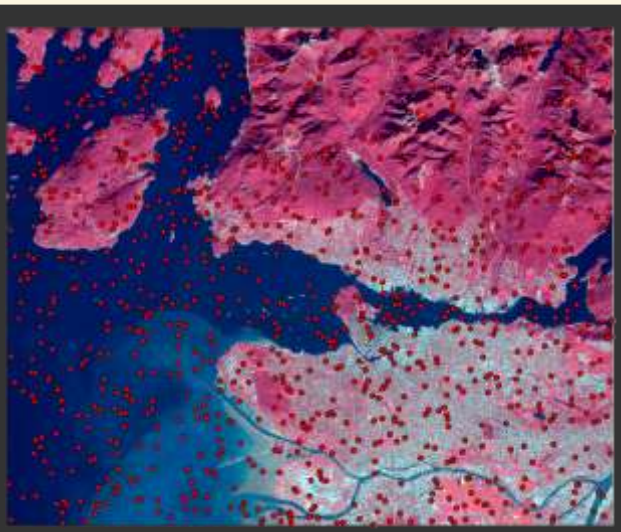
Module Topics

1. Validation data collection
2. Confusion matrix
3. Overall accuracy, precision, and recall
4. Kappa statistic and alternatives
5. Binary classification metrics: precision, recall, F1-score, specificity, and negative predictive value
6. Multi-threshold metrics: ROC curves, PR curves, and AUC
7. Regression metrics: RMSE, MAE, R-squared, AIC, and BIC
8. Multiclass aggregation
9. Confidence intervals

These are the topics covered in this module.

Good validation data should be:

1. Unbiased
2. Randomized
3. Non-overlapping with training data
4. Correctly proportioned
5. Accurate/precise



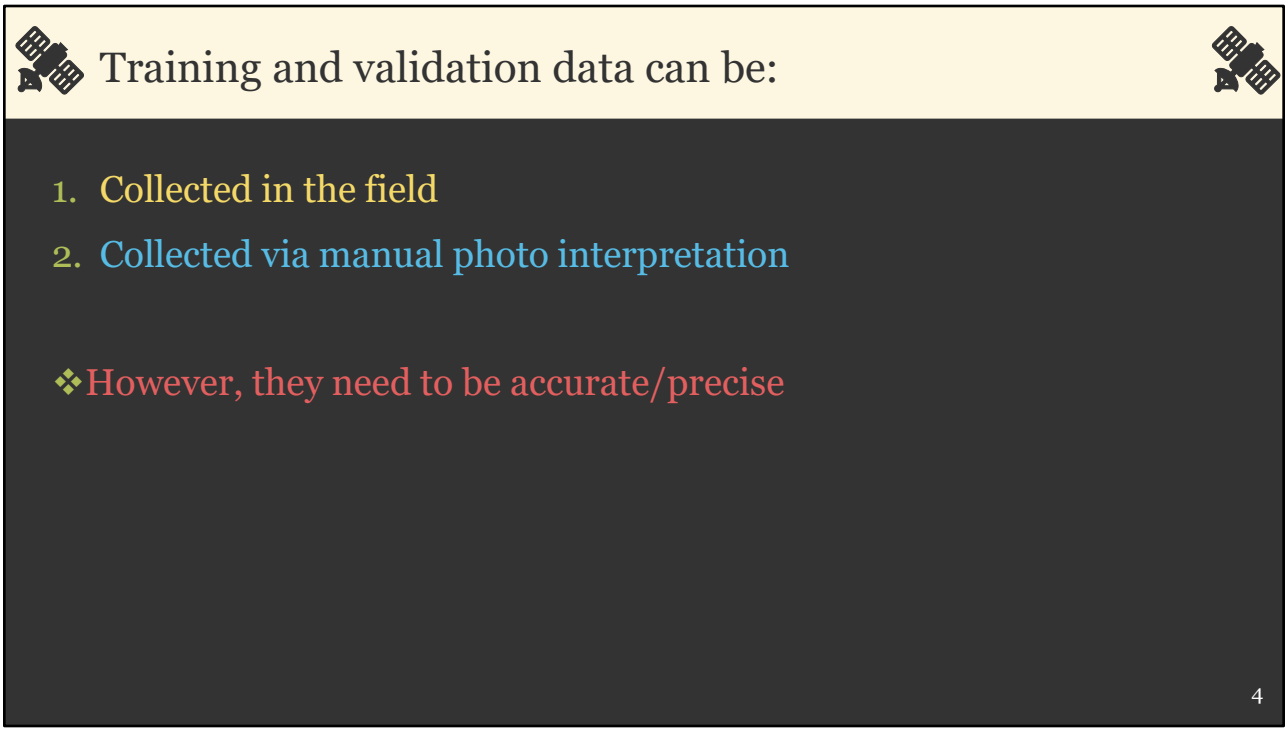
3

Assessment relies on validation data.

In order to produce an unbiased estimate of the performance, the validation data must be randomized in some way. Also, the training and validation data should not overlap, or the same samples should not be included in both the training and validation sets. This is because algorithms tend to do a better job at predicting the training samples as opposed to new locations, a phenomenon known as overfitting. So, including training samples as validation samples could inflate the reported performance.

It is also recommended that training samples be correctly proportioned relative to the map. For example, if you are mapping land cover and 70% of the mapped area is forest, then 70% of the validation samples should also be forest.

Lastly, validation data should be accurate. The goal here is to compare the map product to reference data of higher quality. It is generally assumed that no data are perfect. So, even the validation data will have some error. We try to avoid using the terms “ground truth” or “ground truthing” for this reason.



Optimally, validation data will be collected in the field by visiting and documenting field sites. However, this is not always possible due to access, time, or financial constraints. Field sampling may not be possible or practical. As a result, it is common for training and validation data to be generated in the office using manual photo interpretation methods. In fact, it is more likely for training and validation data to be collected from manual photograph interpretation than field surveys due to practicality and constraints. Sometimes manual photo interpretation is augmented with limited field validation.

I would argue that manual photo interpretation is adequate for validation data collection, as long as a trained analyst can accurately identify or differentiate mapped classes from these sources. Ancillary data used to make interpretations should generally be temporally aligned with the data that were classified and of a higher spatial resolution. If high spatial resolution data are used as input to the classification, it is okay to use the same data for training and validation data collection.

The most important factor here is that validation data are accurate and precise, regardless of the methods used to generate them. For validation data, it is important that the samples are randomized and unbiased so that the assessment is representative of the confusion or error in the map product. For

example, it is common to select validation samples as random image objects or point locations.

The process of validating classification products is generally difficult and time consuming. However, it is a very important component of the thematic map generation process that should always be included. It is difficult to assess the value and appropriate use of data without a rigorous assessment of the error present.

Multiclass Classification Metrics

5

We will begin by discussing assessment metrics commonly used to assess multiclass products, or when there are more than two classes differentiated.

Confusion Matrix or Error Matrix

- Explores confusion between classes
- How were the classes confused?
- Should be created from randomized validation data

		Reference Data							
		Forested	*Pasture / Grass*	*Barren*	*Cropland*	*Developed*	*Water*	Total	1 – Commission Error
Classified Data	*Forested*	91	8	1	11	0	2	113	81%
	Pasture/ Grass	4	81	0	15	0	0	100	81%
	Barren	0	0	87	2	10	2	101	86%
	Cropland	3	11	0	69	0	0	83	83%
	Developed	0	0	10	0	73	0	83	88%
	Water	2	0	2	3	17	96	120	80%
	Total	100	100	100	100	100	100		
	1 – Omission Error	91%	81%	87%	69%	73%	96%		

6

Many assessment metrics for classification problems are based on the confusion matrix, which is also commonly referred to as an error matrix.

Using validation samples, a confusion matrix compares the correct classification at a location to the predicted class. Diagonal cells, shaded gray in the example, represent correct classifications. For example, 91 samples in the example were of the forested class and were correctly labelled as forested. All off-diagonals cells represent errors or misclassifications. For example, 4 samples were forested but were incorrectly labelled as pasture/grass.

The confusion matrix is very informative since it doesn't just summarize the total amount of error, but also differentiates sources of error. An analysis of the confusion matrix allows for an understanding of what classes are well mapped or poorly mapped and what classes are most confused with one another. For example, 15 cropland locations were misclassified as pasture/grass while 11 pasture/grass samples were misclassified as cropland. This indicates that these two classes are commonly confused. In contrast, only 4 of the water samples were misclassified to another class, suggesting that water is well mapped or differentiated.

It is important to note that the quality of the assessment will depend greatly on the quality of the validation data. In remote sensing, we tend to avoid using the

term “ground truth” for validation samples since it is assumed that even ground samples will have some errors or miss-classifications. Instead, we tend to use the term validation data or testing data. Even if there are errors, it is assumed that the validation data are more accurate than the classification that is being assessed. So, it is important to compare your classification to a dataset this is more accurate. Also, it is generally best that validation samples are collected using randomized sampling methods so as not to bias the assessment. If an analyst hand-picks validation samples, these samples are likely to not represent the true landscape conditions.

Confusion Matrix or Error Matrix

- Explores confusion between classes
- How were the classes confused?
- Should be created from randomized validation data

		Reference Data			
		Forested	*Not Forest*	Total	1 – Commission Error
Classified Data	*Forested*	94	15	109	86%
	Not Forest	6	85	91	93%
	Total	100	100		
	1 – Omission Error	94%	85%		

7

This slide provides another example of an error matrix with only two classes.

Metrics

- What can be calculated from a confusion matrix?
- Overall Accuracy = $\frac{\text{Number of Features Correctly Classified}}{\text{Total Number of Features}}$ X 100
- Class User's Accuracy = $\frac{\text{Number of Features of Specific Class Correctly Classified}}{\text{Row Total}}$ X 100
- Class Producer's Accuracy = $\frac{\text{Number of Features of Specific Class Correctly Classified}}{\text{Column Total}}$ X 100

8

From the confusion matrix, it is possible to derive measures of overall classification accuracy and class-level accuracies.

Overall accuracy is simply the number of correctly classified samples divided by the total number of samples. In other words, it is the sum of the diagonal divided by the sum of the entire table. This metric is generally reported either as a proportion (0 to 1) or as a percentage (0% to 100%).

We will discuss the class-level accuracies on the next slide.

Metrics

User's Accuracy

- 1 - Commission error
- Commission Error = error associated with samples being included in the wrong class
- Probability that a feature classified on the map actually represents that category

$$\frac{\text{Number of Features of Specific Class Correctly Classified}}{\text{Row Total}} \times 100$$

Producer's Accuracy

- 1 - Omission error
- Omission Error = error associated with samples being excluded from the right class
- Probability of a reference feature being correctly classified

$$\frac{\text{Number of Features of Specific Class Correctly Classified}}{\text{Column Total}} \times 100$$

9

For each mapped class, two different measures of accuracy can be calculated. User's accuracy is 1 minus the commission error (error associated with samples being included in the wrong class). In contrast, producer's accuracy is 1 minus omission error (error associated with samples not be included in the correct class). User's accuracy represents the probability that a feature classified on the map actually represents that category on the ground whereas producer's accuracy relates to the probability of a reference feature being correctly classified. Similar to overall accuracy, these metrics are reported as proportions (0 to 1) or percentages (0% to 100%). Proportions are most common.

The standard way to generate a confusion matrix is to define the columns using the reference data classifications and the rows using the predicted classes. When the rows and columns are defined using this standard, user's accuracy for each class is calculated as the number correct for the class divided by the row total, which represents the number of samples mapped to that class. In contrast, producer's accuracy for a class is the number correct for the class divided by the column total, which represents the number of samples from that class in the validation data.

I find these terms to be very confusing. I prefer to simply report commission and omission error, as I feel these terms are more easily understood. However,

reporting user's and producer's accuracy is the standard in remote sensing.

Kappa

- Kappa, Kappa Statistic, Cohen's Kappa, K-Hat ($\widehat{K}$)
- Sometimes you are right by random chance
- Kappa offers an adjustment of overall accuracy that takes into account random or change agreement
- It is considered to be more robust than overall agreement

$$\frac{[\text{Total Number of Features} * \text{Number of Correct Features}] - [\text{Sum of All Row Totals} * \text{Column Totals}]}{[\text{Total Number of Features squared}] - [\text{Sum of All Row Totals} * \text{Column Totals}]}$$

10

Sometimes features are correctly classified simply by random chance. The Kappa statistic offers a correction of overall accuracy that takes into account chance agreement. It is a commonly reported metric in remote sensing.

The numerator of Kappa is calculated as the total number of features multiplied by the number of correct features. From this, you subtract the sum of all the row and associate column totals. The denominator is calculated as the square of the number of samples with the sum of all the row and associate column totals subtracted.

Other terms used for Kappa include the Kappa statistic, Cohen's Kappa, and K-Hat ($\widehat{K}$). $\widehat{K}$ indicates that a population statistic is being estimated from a sample.

Kappa is generally reported as a proportion (0 to 1) as opposed to a percentage. It is also possible to calculate confidence intervals and statistically compare Kappa values for different classifications. We will not discuss these additional statistical considerations in this course.

Kappa

		Reference Data							
		Forested	*Pasture/ Grass*	*Barren*	*Cropland*	*Developed*	*Water*	Total	1 – Commission Error
Classified Data	*Forested*	91	8	1	11	0	2	113	81%
	Pasture/ Grass	4	81	0	15	0	0	100	81%
	Barren	0	0	87	2	10	2	101	86%
	Cropland	3	11	0	69	0	0	83	83%
	Developed	0	0	10	0	73	0	83	88%
	Water	2	0	2	3	17	96	120	80%
	Total	100	100	100	100	100	100		
	1 – Omission Error	91%	81%	87%	69%	73%	96%		

Overall Accuracy

$$= \frac{91+81+87+69+73+96}{600}$$

$$= 0.828 \text{ or } 82.8\%$$

Kappa

$$= \frac{(((91+81+87+69+73+96)*600) - (100*113+100*100+100*101+100*83+100*83+100*120))}{(600^2 - (100*113 +100*100+100*101+100*83+100*83+100*120))}$$

$$= 0.794$$

11

Class user's and producer's accuracies have already been calculated on the margins of this example confusion matrix. Make sure you understand how these measures were derived.

I have now provided the calculations for overall accuracy and Kappa. Make sure you understand how these calculations were derived. For this example, overall accuracy was 0.828 while Kappa was 0.794.

Kappa

		Reference Data			
		Forested	*Not Forest*	Total	1 – Commission Error
Classified Data	*Forested*	94	15	109	86%
	Not Forest	6	85	91	93%
	Total	100	100		
	1 – Omission Error	94%	85%		

Overall Accuracy

$$= \frac{94+85}{200}$$

$$= 0.895 \text{ or } 89.5\%$$

Kappa

$$= \frac{(((94+95)*200) - (100*109+100*100*91))}{(200^2 - (100*109+100*100*91))} = 0.890$$

12

This slide provides the metric calculations for the second example.

Should we use Kappa?

International Journal of Remote Sensing
Vol. 32, No. 15, 10 August 2011, 4407–4429

Taylor & Francis

Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment

ROBERT GILMORE PONTIUS JR* and MARCO MILLONES
School of Geography, Clark University, Worcester, MA 01610, USA

(Received 27 August 2010; in final form 20 December 2010)

The family of Kappa indices of agreement claim to compare a map's observed classification accuracy relative to the expected accuracy of baseline maps that can have two types of randomness: (1) random distribution of the quantity of each category and (2) random spatial allocation of the categories. Use of the Kappa indices has become part of the culture in remote sensing and other fields. This article examines five different Kappa indices, some of which were devised by the first author in 2000. We expose the indices' properties mathematically and illustrate their limitations graphically, with emphasis on Kappa's use of randomness as a baseline, and the often-ignored conversion from an observed sample matrix to the estimated population matrix. This article concludes that these Kappa indices are useless, misleading and/or flawed for the practical applications in remote sensing that we have seen. After more than a decade of working with these indices, we recommend that the profession abandon the use of Kappa indices for purposes of accuracy assessment and map comparison, and instead summarize the cross-tabulation matrix with two much simpler summary parameters: quantity disagreement and allocation disagreement. This article shows how to compute these two parameters using examples taken from peer-reviewed literature.

Remote Sensing of Environment

Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification

Giles M. Foody

Foody, G.M., 2020. Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. *Remote Sensing of Environment*, *239*, p.111630.

Pontius Jr, R.G. and Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, *32*(15), pp.4407-4429.

13

Some researchers have argued that the Kappa statistic should no longer be used. Specifically, they argue that the adjustment for chance agreement is not necessary and misleading. Despite these suggestions, Kappa is still a very commonly reported metric. So, I have included an explanation of it in this course. However, you should be aware that its use in the field is somewhat controversial.

If you are interested in this debate, I would recommend the papers referenced on this slide.

Alternatives to Kappa

- Quantity disagreement = Difference in proportion of classes
- Allocation disagreement = Difference in spatial allocation of classes (further divided into exchange and shift)
- Quantity Disagreement + Allocation Disagreement = Overall Disagreement

14

Pontius and Millones (2011) suggested some alternative metrics to replace Kappa. Specifically, they have argued for the use of quantity disagreement and allocation disagreement. Quantity disagreement relates to the differences in proportions of classes between the reference data and classification result whereas allocation disagreement, which is further divided into exchange and shift disagreement, relates to differences in the location or spatial allocation of the refence data and classification result.

I feel that these metrics are informative and worth including in accuracy assessment workflows and reports. However, they have not gained wide use and acceptance in the field.

A tool for calculating and exploring a confusion matrix

- Excel Spreadsheet by Dr. Gil Pontius of Clark University

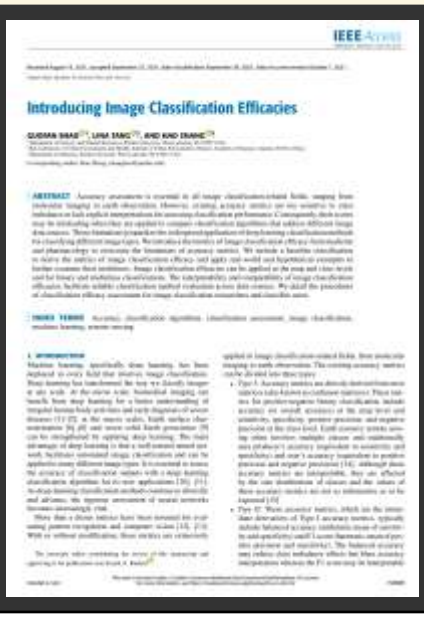
Excel Spreadsheet: http://www2.clarku.edu/~rpontius/

R Package: https://cran.r-project.org/web/packages/diffeR/index.html

15

If you would like to explore these alternative metrics, I would recommend the Excel Spreadsheet linked here and provided by Dr. Gil Pontius of Clark University. These metrics can also be calculated in the R data science environment using the diffeR package.

MICE



- Map-Level Image Classification Efficacy
- Alternative to Kappa

Shao, G., Tang, L. and Zhang, H., 2021. Introducing image classification efficacies. *IEEE Access*, *9*, pp.134809-134816.

Tang, L., Shao, J., Pang, S., Wang, Y., Maxwell, A., Hu, X., Gao, Z., Lan, T. and Shao, G., 2024. Bolstering performance evaluation of image segmentation models with efficacy metrics in the absence of a gold standard. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, pp.1-12.

16

Another alternative to Kappa that has been recently suggested is the map-level image classification efficacy (MICE) metric. It can be calculated using the micer R package.

MICE

- Adjusts the accuracy rate relative to a random classification baseline.
- Only the proportions from the reference labels are considered, as opposed to the proportions from the reference and predictions, as is the case for the Kappa statistic.

$$\text{Accuracy of a random classification } (OA_0) = \sum_{j=1}^{J} \left(\frac{n_j}{n}\right)^2$$

$$\text{Map image classification efficacy (MICE)} = \frac{OA - OA_0}{1 - OA_0}$$

17

MICE offers and adjustment of overall accuracy (OA) relative to a random classification baseline (OA_0). For class j, the probability that a random sample is correctly labeled to that class is equivalent to the square of the proportion of the total count of samples, n, that belong to class j (n_j). The estimated accuracy of a random classification is defined as the sum of all these class-level probabilities, and OA is adjusted to obtain MICE by subtracting OA_0 and dividing by 1 – OA_0.

MICE

Reference-total-based image classification efficacy (RE_j) =

$$\frac{PA_j - \frac{n_j}{n}}{1 - \frac{n_j}{n}}$$

Classification-total-based image classification efficacy (CE_j) =

$$\frac{UA_j - \frac{n_j}{n}}{1 - \frac{n_j}{n}}$$

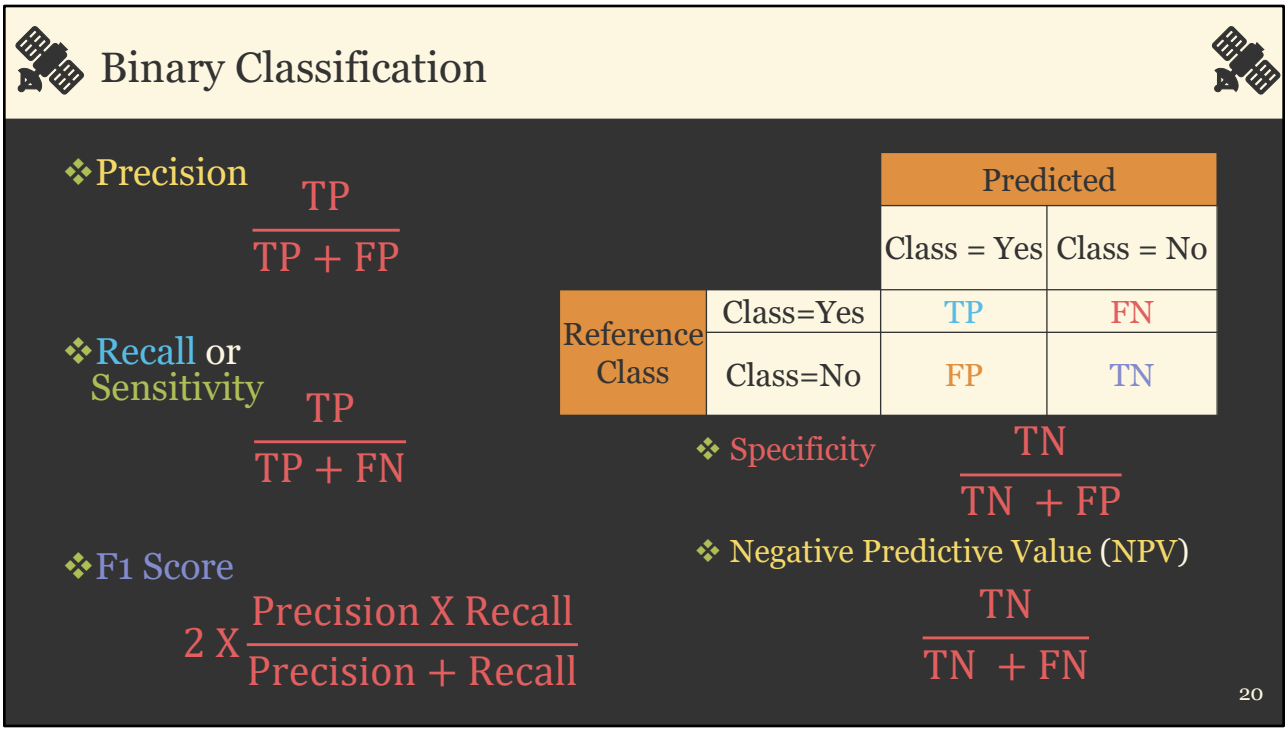
18

It is also possible to apply MICE-style adjustments to class-level metrics, such as producer's accuracy (recall) and user's accuracy (precision).

Binary Classification Metrics

19

We will now explore metrics used for a binary classification and when there is a defined positive or presence class and a negative for background class.



When only a single class is differentiated from the background, as is common in feature extraction tasks, or when only two classes are differentiated to generate a binary output, an alternative terminology is commonly used.

Specifically, if one category represents a positive case while the other represents a background or negative case, the terms true positive (TP), false positive (FP), true negative (TN), and false negative (FN) are used. TP samples are those that are in the positive class and are correctly predicted as positive while FPs are not in the positive class but are incorrectly predicted as a positive case. TNs are in the negative class and are correctly predicted as negative while FNs are predicted as negative when they are actually positive. From this cross tabulation, it is possible to derive a variety of metrics.

Precision represents the proportion of the samples that is correctly classified within the samples predicted to be positive, and it is equivalent to 1 – commission error for the positive class. Recall or sensitivity represents the proportion of the reference data for the positive class that is correctly classified, and it is equivalent to 1 – omission for that class.

The F1 score is the harmonic mean of precision and recall while specificity represents the proportion of negative reference samples that is correctly predicted and is thus equivalent to 1 – omission error for the negative class.

The negative predictive value (NPV) is 1 – commission error for the negative class.

Binary and Multiclass Assessment Crosswalk

Metric	Equation	Relation to Multi-Class Metrics
Overall Accuracy (OA)	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall Accuracy
Recall (Sensitivity) True Positive Rate (TPR)	$\frac{TP}{TP + FN}$	Producer's Accuracy for positive class
Precision Positive Predictive Value (PPV)	$\frac{TP}{TP + FP}$	User's Accuracy for positive case
F1-Score (Dice Coefficient)	$2 \text{ X} \frac{\text{Precision X Recall}}{\text{Precision} + \text{Recall}}$	Harmonic mean of Producer's Accuracy and User's Accuracy for the positive case
Specificity True Negative Rate (TNR)	$\frac{TN}{TN + FP}$	Producer's Accuracy for negative case
Negative Predictive Value (NPV)	$\frac{TN}{TN + FN}$	User's Accuracy for negative case

21

The goal of this slide is to provide a crosswalk between binary and multiclass assessment metric terminology.

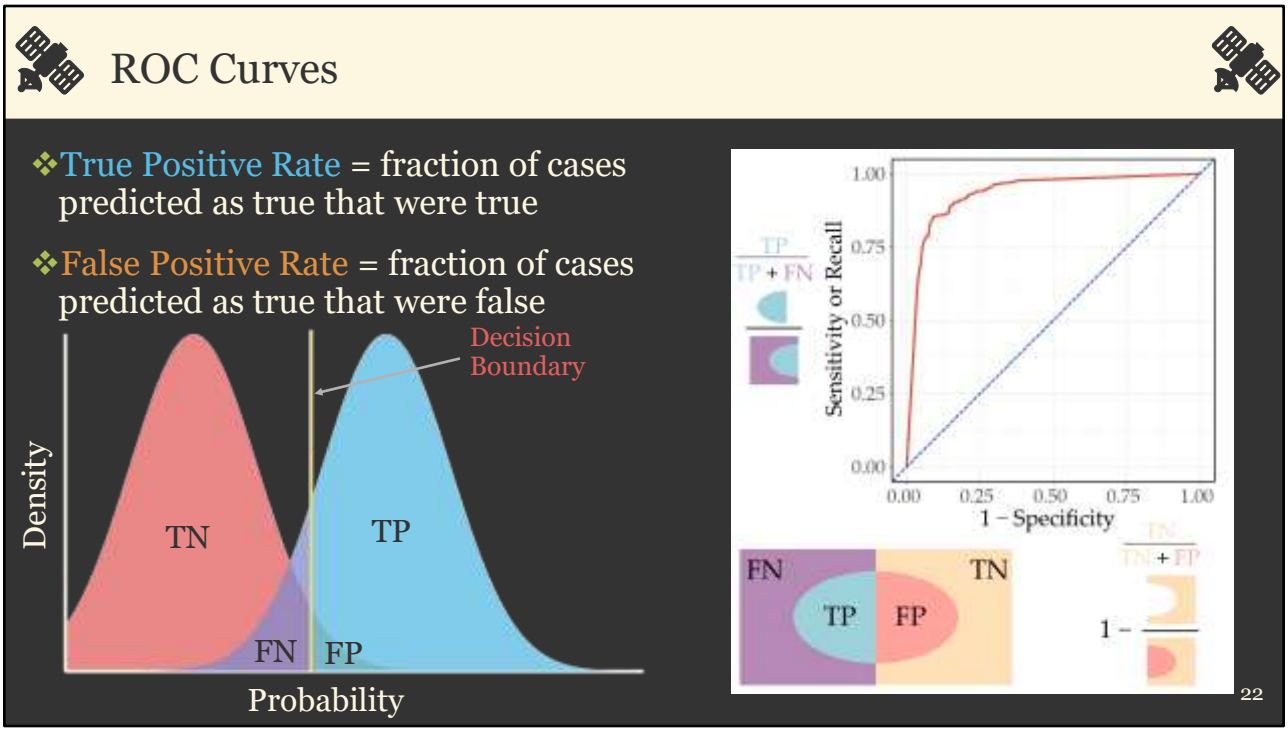
Overall accuracy is calculated the same for binary and multiclass classification problems as the total correct divided by the total number of samples.

Recall (also sometimes referred to as sensitivity or true positive rate (TPR)) is equivalent to the producer's accuracy for the positive case. All these metrics quantify 1 – omission error.

Precision (also sometimes referred to as positive predictive value (PPV)) is equivalent to the user's accuracy for the positive case. These metrics quantify 1 – omission error.

F1-score, which is also referred to as the Dice coefficient, is equivalent to the harmonic mean of the user's and producer's accuracies for the positive case.

For the negative case, specificity represents producer's accuracy (i.e., 1 – omission error) while negative predictive value (NPV) represents user's accuracy (i.e., 1 – commission error).

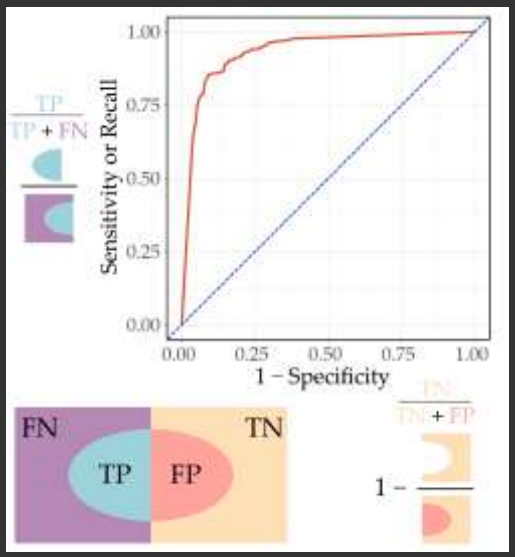


Probabilistic predictions are commonly assessed using Receiver Operating Characteristic Curves or ROC curves. These curves compare the true positive and false positive rates. True positives represent cases predicted as true that were actually true whereas false positives are cases predicted as true that were actually false. For example, if you were trying to predict which subjects had a certain disease, the true positive rate would be the fraction of the total number of subjects that were predicted as having the disease that actually have it. In contrast, the false positive rate would be the fraction of the total number of subjects that were predicted to have the disease but did not.

Using the terminology from the binary confusion matrix, false positive rate is equivalent to recall while false negative rate is equivalent to 1 – specificity.

ROC Curves

- Receiver Operating Characteristic (ROC) Curve
- Plots False Positive Rate and True Positive Rate at all decision/probability thresholds
- Sensitivity = True Positive Rate (Recall)
- Specificity = 1 – False Positive Rate



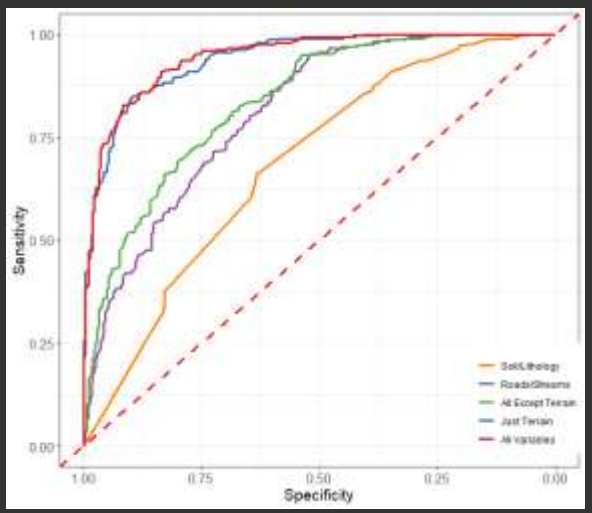
23

The ROC curve specifically plots sensitivity, which is equivalent to recall, against specificity. Sensitivity is the true positive rate (or recall) while specificity is 1 minus the false-positive rate. The rates depend on what probability threshold is used. So, instead of performing the assessment at one threshold, assessment is performed at all thresholds to plot a continuous curve.

ROC Curves

- Area Under Curve (AUC) = area under ROC Curve (0-1)
- Larger suggest better models
- Can be used to compare models

0.90-1 = excellent (A)
0.80-0.90 = good (B)
0.70-0.80 = fair (C)
0.60-0.70 = poor (D)
0.50-0.60 = fail (F)



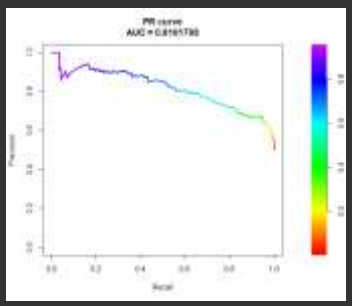
24

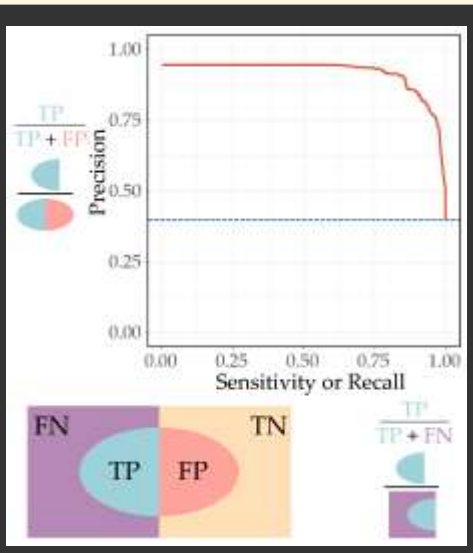
Using the ROC curve, the AUC, or area under the ROC curve, can be calculated. As the name implies, this is simply the area under the ROC curve. This measure is scaled from 0-1, where 1 is the entire area of the ROC graph. This is generally interpreted like a grade scale, as described on the slide. So, larger values are better. An AUC of 0.5 indicates that your model is not doing better than simply taking a random guess.

Since many algorithms generate class probabilities as a means to obtain the hard classification, ROC curves and the AUC measure are sometimes used to assess classification models. This is especially true for binary classifications. However, a multiclass version of the ROC curve can be generated, and a multiclass AUC can be calculated.

P-R Curves

- Precision-Recall Curve
- Graph precision vs. recall at different decision/probability thresholds
- Can calculate area under the curve, similar to ROC



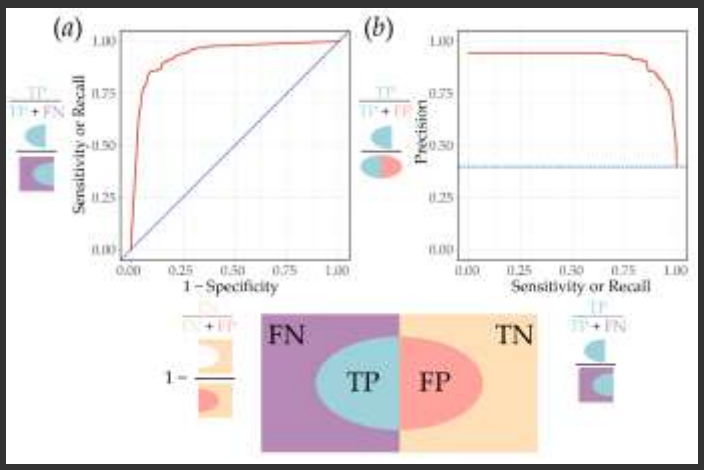


25

Assessments based on ROC curves do not take into account data imbalance, or the impact of a different number of samples in the two classes. The precision-recall (P-R) curve offers an alternative that takes this imbalance into account. Instead of using specificity, precision is graphed against recall or sensitivity. Similar to the ROC curve, the area under the P-R curve can be calculated as an assessment metric.

ROC Curve vs. P-R Curve

- Both are used for binary classification problems
- ROC Curve can be flawed when highly imbalanced data are used
- ROC Curve only considers omission error
- P-R Curve considers omission and commission error for positive case
- P-R Curve is better if data set contains moderate to large class imbalance



26

Since ROC curves and the associated AUC ROC metric rely on recall and specificity, which are both insensitive to data imbalance, they can be misleading in cases where data imbalance should be taken into account, such as when the classes make up very different proportions of the population. Specifically, reported metrics can be overly optimistic in cases of severe class imbalance and/or when the class of interest makes up a small percentage of the population.

For example, if the goal is to map the locations of wetlands in a landscape where this class makes up less than 1% of the landscape, recall provides a quantification of the percentage of wetland samples that were correctly mapped as wetlands while specificity represents the percentage of not wetland samples that were correctly mapped. These two metrics and the associated ROC curve do not incorporate precision, which quantifies the percentage of the samples that were predicted to be wetlands that were actually wetlands. If a large percentage of the landscape is not wetlands, then the ROC curve and AUC ROC measure will not capture issues of FP cases, which may be an important criteria in determining the quality of the classification and the amount of manual labor necessary to improve the results (i.e., manually re-labelling the FP cases to not wetland).

In such cases, a precision-recall (PR) curve may be more informative since it

does incorporate precision, which is sensitive to class imbalance and quantifies the percentage of samples predicted to the positive class that were TPs. This curve plots sensitivity or recall to the x-axis and precision to the y-axis. Similar to the ROC curve, it is possible to generate an area under the curve (AUC PR) metric to obtain a single summary statistic.

Regression Metrics

27

We will now discuss metrics used to assess predictions of numeric values as opposed to categories.

Assessing a Numeric/Continuous Prediction: RMSE

- RMSE = $\sqrt{\Sigma(y - \hat{y})^2/n}$
- Root Mean Square Error (RMSE) will be in the units of y
- Mean Square Error (MSE) = $\Sigma(y - \hat{y})^2/n$

28

Continuous or numeric predictions can be assessed using root mean square error or RMSE. MSE, or Mean Square Error, is simply RMSE without the square root applied. RMSE will be reported in the units of the variable being predicted while MSE will be in the square of the units. For example, if you are predicting chemical concentration in parts per million, then RMSE will be reported in parts per million and MSE will be reported in parts per million squared. RMSE is calculated by comparing the predict value to the value provided in the validation data. The values are subtracted to obtain a residual or error. The residuals are then squared, summed, then divided by the number of samples. This will provide MSE. The square root is taken to obtain RMSE. Lower RMSE and MSE are better.

Since we are focusing on the assessment of classifications here, RMSE would not be appropriate. However, it can be used to assess a prediction of a continuous measure or the performance of a regression model.

R^2

- $R^2 = \frac{Total\ Sum\ of\ Squares - Residual\ Sum\ of\ Squares}{Total\ Sum\ of\ Squares}$
- Total Sum of Squares = $\Sigma(y - \bar{y})^2$
- Residual Sum of Squares = $\Sigma(y - \hat{y})^2$
- Proportion of variance in y explained by the model

29

Another means by which to assess continuous predictions is R^2. This value represents the proportion of variance in the quantity being predicted explained by the model. It is scaled from 0 to 1. 0 indicates that no variance is explained, or this is a poor model, while 1 indicates that all variance is explained. So, higher values are better. Again, this measure would not be appropriate for assessing a classification.

R^2 is not strictly a measure of model accuracy or performance. Instead, it is a measure of variance explained. However, it is still useful for assessing models.

Adjusted R^2

- When you have more than one x (multiple regression) you have to use adjusted R^2
- Adjusted $R^2 = 1 - \frac{(1-R^2)(N-1)}{(N-p-1)}$
- N = sample size, p = number of variables

30

When using the training data and when more than one predictor variable is used, R^2 must be modified to obtain adjusted R^2 to deal with inflation. The equation requires altering R^2 based on the sample size and number of predictor variables.

Note that it is not necessary to perform this adjustment when the withheld test or validation data are used to obtain R^2 as opposed to the training data.

Other Regression Metrics

Akaike Information Criterion (AIC)
- Used for model comparisons
- Based in information theory
- Considers goodness of fit
- Penalizes for increased number of predictor variables/model complexity
- Lower value is better

Bayesian Information Criterion (BIC)
- Based on Bayes' Theorem
- Considers goodness of fit
- Penalizes for increased number of predictor variables/model complexity
- Lower value is better

31

Another option is the Akaike Information Criterion or AIC. This is a measure of the relative quality of statistical models based on information theory. It takes into account both the goodness of fit (i.e., accuracy) and the complexity of the model. Specifically, models are penalized for having a larger number of included predictor variables. AIC is generally used to compare between models and select the best performing models while also taking into account the number of predictor variables. Larger values are preferred.

An alternative to AIC is the Bayesian Information Criterion (BIC). Similar to AIC, this method takes into account model performance while also penalizing for a large number of predictor variables. In contrast to AIC, it is based on Bayes' Theorem, and lower values indicate a better model as opposed to larger values, which is the case for AIC.

Other Considerations

Aggregated Multiclass Metrics

- Macro-Averaging
- Micro-Averaging

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Dice} = 2\ X \frac{\text{Precision X Recall}}{\text{Precision+Recall}}$$

$$\text{Dice} = 2\ X \frac{2\ X\ TP}{(TP+FP)+(TP+FN)}$$

Class	TP	FP	FN	Precision	Recall	F1-Score/ Dice
A	120	17	41	0.88	0.75	0.81
B	96	12	17	0.89	0.85	0.87
C	111	32	21	0.78	0.84	0.81

$$\text{Dice (Macro)} = \frac{0.81+0.87+0.81}{3l} = 0.83$$

$$\text{Dice (Micro)} = \frac{2*(120+96+111)}{(120+96+111+17+12+32)+(120+96+111+41+17+21)} = 0.83$$

33

It is possible to aggregate class-level metrics to a single value. Examples include class aggregated recall (i.e., producer's accuracy), precision (i.e., user's accuracy), and F1-score (i.e., the Dice coefficient). This slide provides an example for the Dice coefficient or F1-score. When more than two classes are differentiated, or there is not a positive and background case, background or negative case metrics, such as specificity and negative predictive value, do not have meaning.

There are generally two means to aggregated class-level metrics: macro-averaging and micro-averaging. Macro-averaging consists of calculating the metric separately for each class then averaging the results. When macro-averaging is used, all classes are treated equally in the average, regardless of the relative number of samples assigned to each class. This can be useful when minority classes are of importance, or you are attempting to balance performance across all classes.

Micro-averaging consists of aggregating all samples before calculating the metric. Effectively, the confusion matrix is aggregated to TP, TN, FP, and FN counts prior to calculating metrics. When this averaging method is used, classes with more samples have a larger weight in the final calculation.

Since commission errors are omission errors when referenced relative to

another class, it turns out that micro-averaged precision, recall, and F1-score are all equivalent, Further, they are actually equivalent to overall accuracy; as a result, if class aggregated metrics are reported, we recommend reporting the macro-averaged versions alongside overall accuracy.

A note on the Population Matrix

- In order to estimate overall map accuracy, the validation data should occur in the same proportion as the classes on the ground
- For example, if 75% of the area is forest, 75% of the validation data should be forest
- If this is not the case, you should re-proportion the confusion matrix to maintain an accurate estimate of map accuracy

34

In order to assess the accuracy of a map product, it is generally recommended that the proportion of samples in each class in the reference set approximate the proportion of land area of the class in the mapped extent. If a purely random sampling routine is used to select validation locations, this should hold true, since the random chance of a location in a certain class being selected is proportional to the class's land area. However, when other sampling methods are used, such as stratified random sampling, then proportions of samples by class will not align with the proportion of land area.

A population matrix or estimated population confusion matrix is a confusion matrix in which the proportions of land area of each class is approximated by the proportion of refence samples. If random sampling was not used to select the validation data locations, it is possible to approximate a population matrix from a confusion matrix mathematically if the proportion of each class on the landscape can be estimated.

When assessing map output, it is generally recommended to do so using a population matrix. If a class that takes up a large proportion of the land area is commonly misclassified, this should have a higher weight in the assessment than more rare classes since this will have a larger impact on map accuracy.

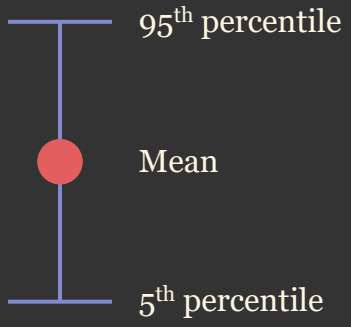
Despite arguments for the use of population matrices, many published studies

do not make the required adjustments.

Confidence Intervals with Bootstrapping

- Bootstrap = random sample with replacement

1. Generate large set of bootstrap replicates (> 1,000)
2. Calculate metric of interest for each replicate
3. Obtain percentiles (5% and 95% for 95% confidence interval)



35

It can be informative to include an estimate of uncertainty around assessment metrics. One way to do this is to calculate a confidence interval (CI).

More specifically, confidence intervals can be estimated using bootstrap replicates from the validation set. Bootstrapping is random sampling with replacement such that a sample is generated that is the same size as the original sample, but some samples are repeated while others are not sampled or included. The assessment metric of interest can then be calculated for each of the replicates. To obtain accurate CIs, a large number of replicates are used (generally > 1,000).

Once the assessment metric is calculated for each replicate, the mean can be reported along with an associated CI. For a 95% confidence interval, the 5th and 95th percentiles can be calculated from the large set of bootstrap replicates.

Statistical Test for Model Difference

- Generate a set of bootstrap samples of the test set to calculate the assessment metric of interest
- Since the samples are not independent, the difference between the metric of interest for each unique bootstrap sample pair is then calculated to serve as a dependent variable
- Use standard analysis of variance (ANOVA) (more than two models) test or a paired t-test (two models)
- The null hypothesis is that there is no difference between the model performance for each pair while the alternative hypothesis is that there is a difference. Since this is a formal statistical test, it yields a *p*-value.
- If more than two models are compared and the ANOVA analysis suggests statistical difference, a post-hoc pairwise comparison can then be used to compare each pair of models.

https://www.tmwr.org/compare

36

In order to statistically compare models or classifications, I recommend using either analysis of variance (ANOVA) when more than two results are compared or a t-test if two results are compared. Since the samples are not intendent, the difference between pairs should be used.

Chapter 11 of *Tidy Modeling with R* describes how to implement these tests in R. It is linked on this slide.

This is the end of this lecture module.

Please return to the West Virginia View Webpage for additional content.

37

Thanks! Hope you found this useful.