Methods in Open Science

Model Assessment

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

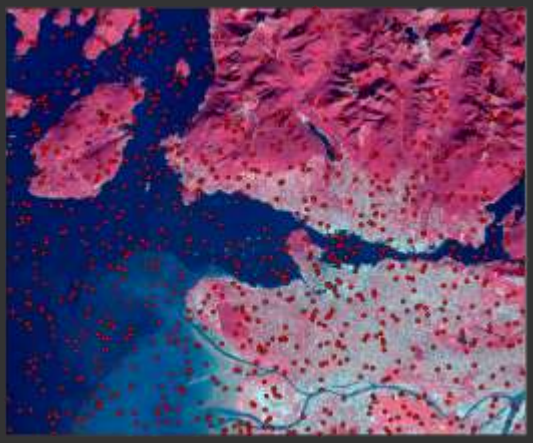
WV VIEW

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

Now that we have conceptualized the process of predictive modeling and some common methods used to create models, we will turn our attention to how models are assessed or validated.

Validation Data

- Unbiased
- Randomized
- Non-overlapping with training data
- Accurate/precise



2

Validation data must be provided to assess the model performance.

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 data points (i.e., overfitting). So, including training samples as validation samples could inflate the reported performance.

Lastly, validation data should be accurate. The goal here is to compare the prediction 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.

Regression

3

Let's begin with a discussion of metrics used to validate numeric predictions or regression models.

RMSE

- RMSE = Root Mean Square Error

$$\sqrt{\Sigma(y - \hat{y})^2/n}$$

  - In the units of y

- MSE = Mean Square Error

$$\Sigma(y - \hat{y})^2/n$$

4

Continuous or numeric predictions can be assessed using root mean square error (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 predicted 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 suggests better predictive performance.

RMSE Example

Predicted	Actual	Residual	Squared Residual
15.1	15.7	-0.6	0.36
21.5	20.8	0.7	0.49
17.5	17.2	0.3	0.09
14.8	15.1	-0.3	0.09
11.7	12.4	-0.7	0.49
		Sum of Squared Residuals	1.52
		MSE	0.304
		RMSE	0.551

5

This slide provides an example calculation of MSE and RMSE. Step through this example and make sure you understand how the result was obtained.

R-Squared (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

6

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, suggesting 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.

To calculate the total sum of squares, you subtract the mean from each data point value from the validation set then square the result and sum all squared differences. This serves as a measure of total variance in the dataset.

The residual sum of squares is calculated by subtracting the predicted value for each data point from the value provided in the validation data, squaring the difference, then summing all the differences. This represents the unexplained variance.

The residual sum of squares is then subtracted from the total sum of squares. The difference is next divided by the total sum of squares. Again, this represents the proportion of variance explained by the model and will scale for 0 to 1.

Adjusted R^2

- R^2 always increases as you add predictor variables
- Adjusted R^2

$$1 - \frac{(1-R^2)(N-1)}{(N-p-1)}$$

N = sample size, p = number of variables

- Penalization for increased number of predictor variables and model complexity

7

When more than one predictor variable is used, R^2 must be modified to obtain Adjusted R^2. R-squared will always increase as predictor variables are added, so without this adjustment, the measure can be misleading when multiple predictors are included.

The equation requires altering R^2 based on the sample size and number of predictor variables.

Note that this is only required when using the training data to calculated R-squared. If you are using a withheld test or validation set, this correction is not required.

Other 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
- Larger 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

8

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.

Summary of Key Points: Regression

- Prediction of numeric variables is often assessed using withheld testing or validation data using the RMSE and/or R-squared measures.
- R-squared is a measure of correlation as opposed to accuracy.
- Adjusted R^2 penalizes for increased number of predictor variables
- AIC and BIC consider both model performance and model complexity (e.g., the number of included predictor variables)

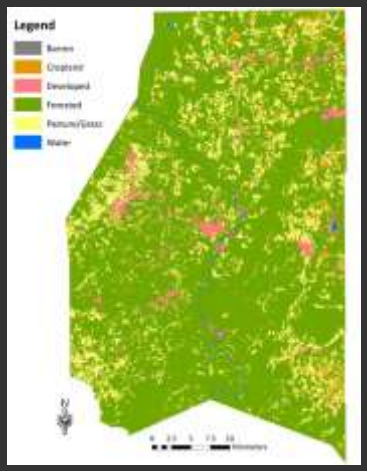
Classification

10

We will now move on to discuss metrics used to evaluate multiclass and binary classification models.

Confusion Matrix

		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%		



11

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 for data points 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 labeled as forested. All off-diagonal cells represent errors or misclassifications. For example, 4 samples were forested but were incorrectly labeled 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 predicted or poorly predicted 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 predicted.

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

“ground truth” for validation samples since it is assumed that even manually labeled samples will have some errors or misclassifications. 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 Summary

$$\text{Overall Accuracy} = \frac{\text{Number of Features Correctly Classified}}{\text{Total Number of Features}} \times 100$$

$$1 - \text{Commission Error} = \frac{\text{Number of Features of Specific Class Correctly Classified}}{\text{Row Total}} \times 100$$

$$1 - \text{Omission Error} = \frac{\text{Number of Features of Specific Class Correctly Classified}}{\text{Column Total}} \times 100$$

12

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%).

For each differentiated class, two different measures of accuracy can be calculated. 1 minus the commission error relates to samples being included in the wrong class. In contrast, 1 minus omission error is associated with samples not be included in the correct class. 1 minus commission error represents the probability that a feature classified to a specific class actually represents that category in the real world whereas 1 minus omission error 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%).

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, 1 minus commission error for each class is calculated as the number correct for the

class divided by the row total, which represents the number of samples predicted to that class. In contrast, 1 minus omission error 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.

Kappa

- Kappa, Kappa Statistic, Cohen's Kappa, K-Hat ($\widehat{K}$)
- Sometimes you are right by random chance
- An adjustment of overall accuracy that considers random or chance agreement

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

13

Sometimes features are correctly classified simply by random chance. The Kappa statistic offers a correction of overall accuracy that takes into account chance agreement.

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 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.

Example Calculations (1)

		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

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

$= 0.828$ or 82.8%

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

$= 0.794$

14

Here is an example of calculating overall accuracy, Kappa, and class-level 1-omission and 1-commission errors. Make sure you understand how these metrics were calculated.

Example Calculations (2)

		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

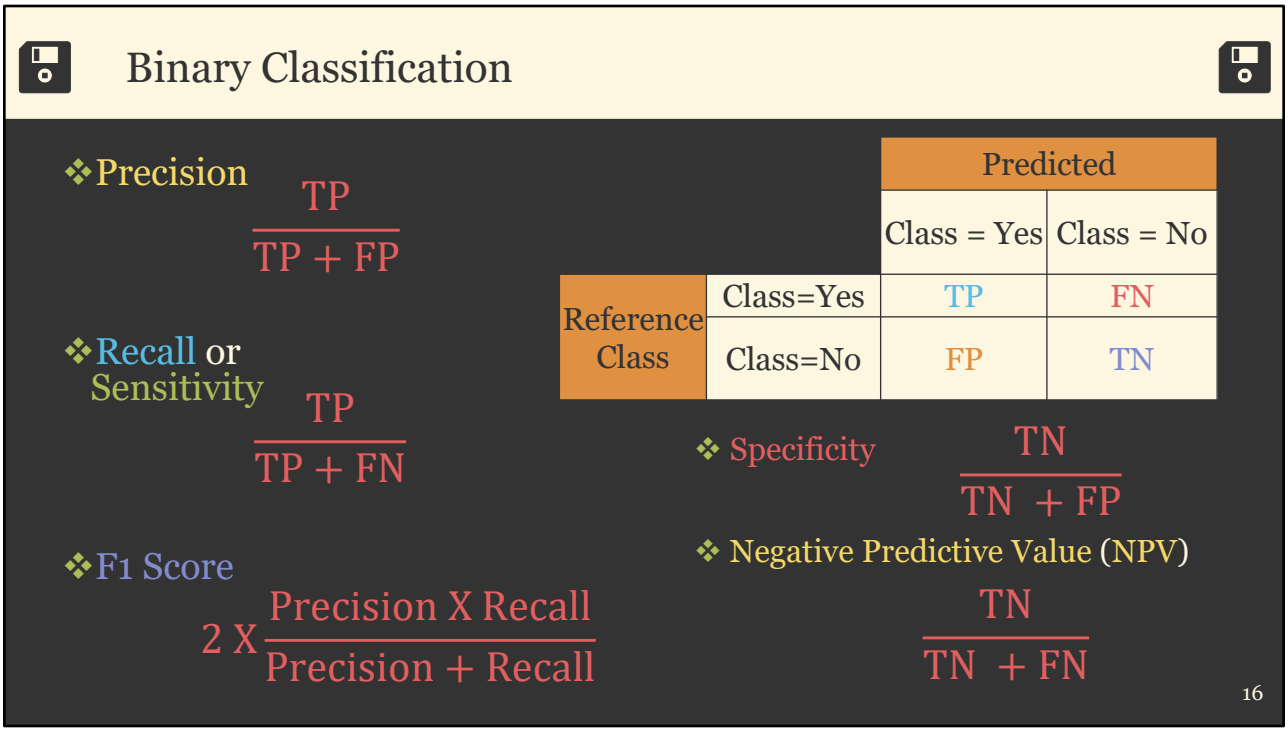
$$= \frac{94+85}{200}$$

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

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

15

This slide provides the metric calculations for a second example.



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.

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

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

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

https://stephenallwright.com/micro-vs-macro-f1-score/

17

It is possible to aggregate class-level metrics to a single value. Examples include class aggregated recall, precision, 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.

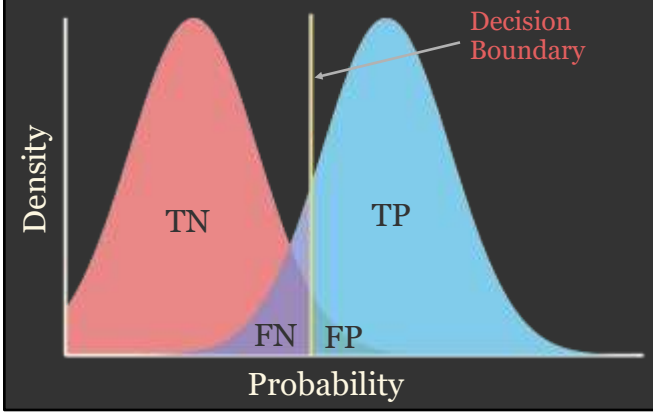
Micro-averaging consists of aggregating all samples before calculating the metric. Effectively, the confusion matrix is aggregated to TP, TN, FP, and FN counts prior to calculating metrics. When this averaging method is used, classes with more samples have a larger weight in the final calculation.

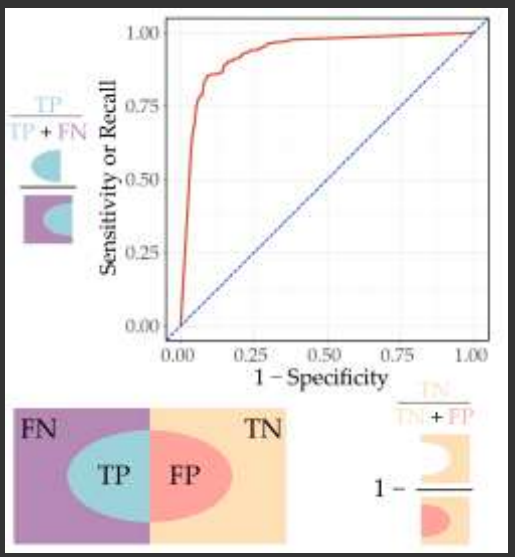
Since commission errors are omission errors when referenced relative to another class, it turns out that micro-averaged precision, recall, and F1-score

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.

ROC Curves

- True Positive Rate = fraction of cases predicted as true that were true
- False Positive Rate = fraction of cases predicted as true that were false





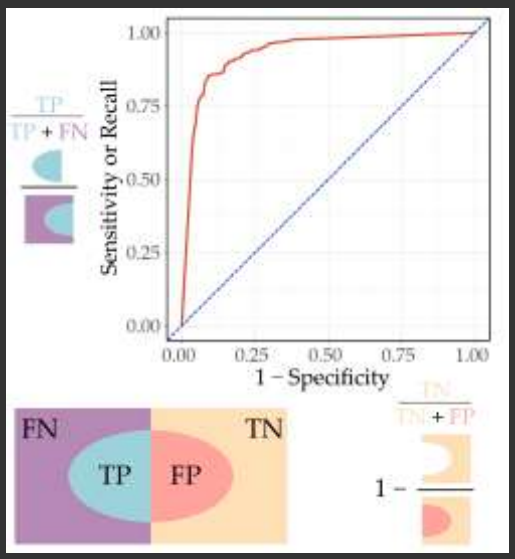
18

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



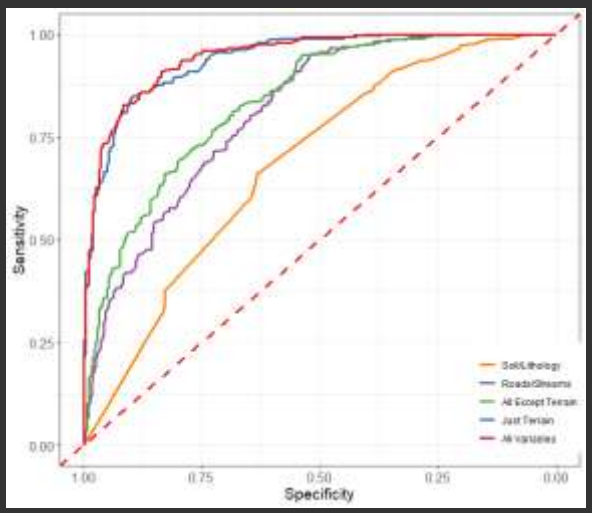
19

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)



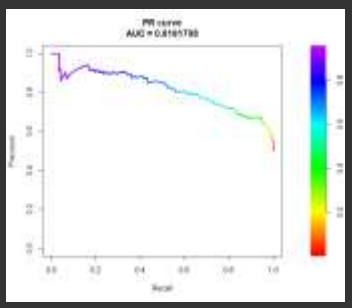
20

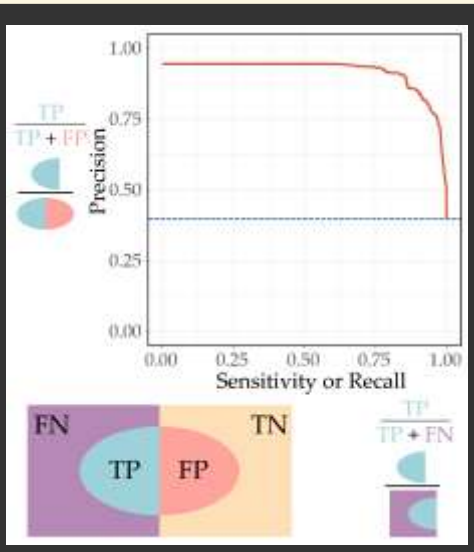
Using the ROC curve, the AUC, or area under the ROC curve, can be calculated. As the name implies, this is simply the area under the ROC curve. This measure is scaled from 0-1, where 1 is the entire area of the ROC graph. This is generally interpreted like a grade scale, as described on the slide. So, 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

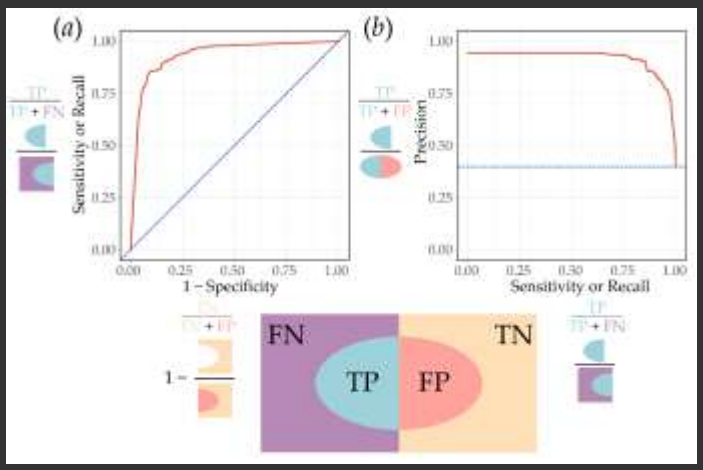




Assessments based on ROC curves do not take into account data imbalance, or the impact of a different number of samples in the two classes. The precision-recall (P-R) curve offers an alternative that takes this imbalance into account. Instead of using specificity, precision is graphed against recall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



22

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.

Summary of Key Points: Classification

- Predictions of a nominal variable are assessed using withheld testing or validation data.
- The reference and predicted classification are compared and summarized using a confusion matrix. Several statistics can be derived from the confusion matrix.
- Different terminology is commonly used when assessing multiclass and binary classification
- The receiver operating characteristic curve, precision-recall curve, and associated area under the curve measures offer a means to assess probabilistic outputs not reliant on a single decision threshold.
- The impact of class imbalance is important to consider when assessing classification output.

This is the end of this lecture module.

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

Thanks! Hope you found this useful.