Datascience fundamentals

Fernando Lovera

Mario Verstraeten

Lecture 6:

Performance Evaluation (ROC)

Agenda

-Performance eval: Regression

-Performance eval: Classification

- Confusion matrix

- Accuracy, Recall, Precision, F1

- Definition ROC, theory ROC, history ROC, meaning Roc, making ROC curve

- Interaction of ROC curve

- Multiple models with ROC

- Area under ROC curve

- Use of eval measures

Performance regression

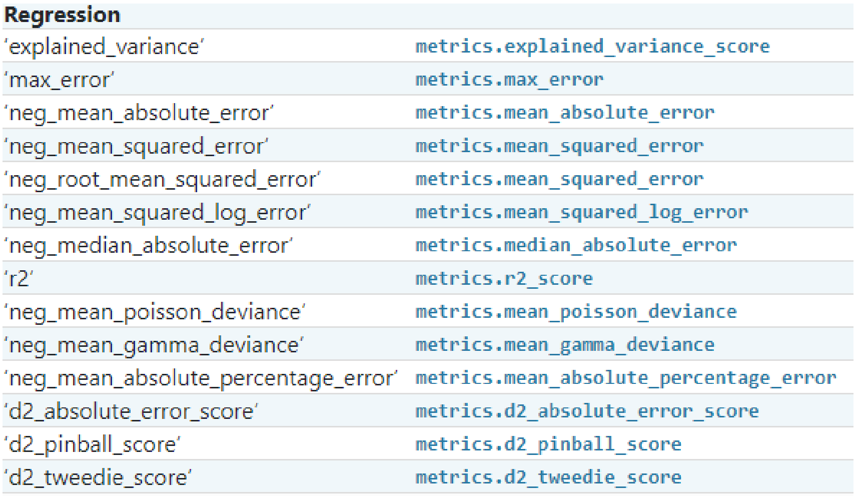
Remember:

- Mean Absolute Error
- Mean Squared Error
- Root Mean Squared Error

But there is also...

- R^2 score
- D^2 score

In scikit-learn:



Regression	
‘explained_variance’	`metrics.explained_variance_score`
‘max_error’	`metrics.max_error`
‘neg_mean_absolute_error’	`metrics.mean_absolute_error`
‘neg_mean_squared_error’	`metrics.mean_squared_error`
‘neg_root_mean_squared_error’	`metrics.mean_squared_error`
‘neg_mean_squared_log_error’	`metrics.mean_squared_log_error`
‘neg_median_absolute_error’	`metrics.median_absolute_error`
‘r2’	`metrics.r2_score`
‘neg_mean_poisson_deviance’	`metrics.mean_poisson_deviance`
‘neg_mean_gamma_deviance’	`metrics.mean_gamma_deviance`
‘neg_mean_absolute_percentage_error’	`metrics.mean_absolute_percentage_error`
‘d2_absolute_error_score’	`metrics.d2_absolute_error_score`
‘d2_pinball_score’	`metrics.d2_pinball_score`
‘d2_tweedie_score’	`metrics.d2_tweedie_score`

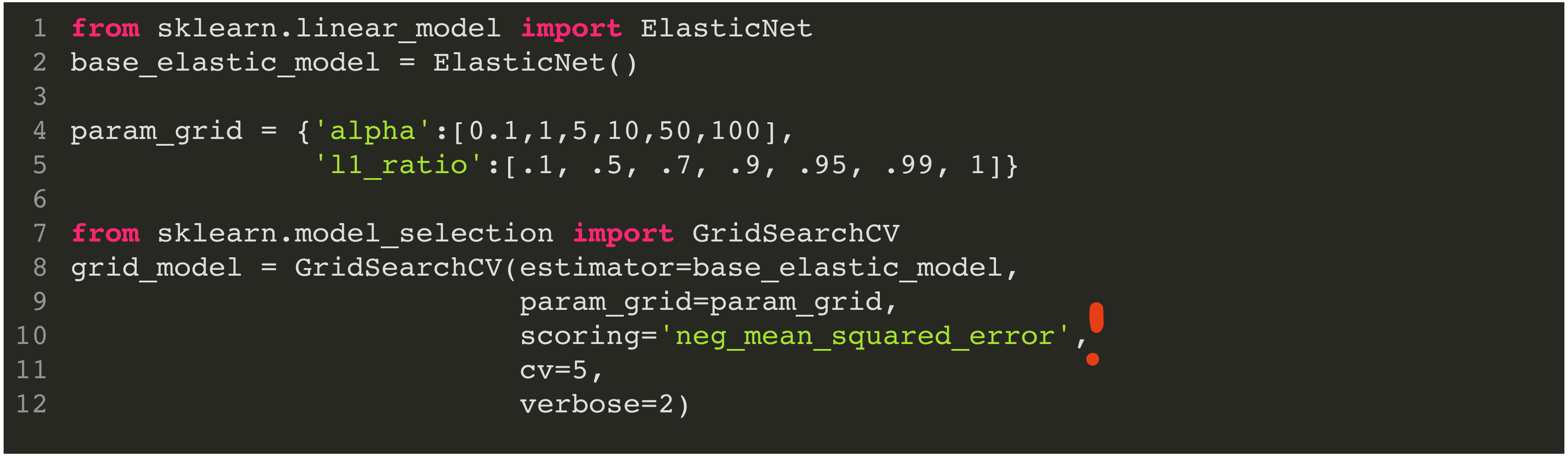
R^2? : It tells you how well the input variables explain the variation in the output variable. It ranges from 0 to 1.

A value of 1 indicates that the model perfectly predicts the output variable, while a value of 0 indicates that the model does not explain any of the variability.

D^2?: Addresses the issue of R^2 potentially increasing when more variables are added to the model, even if they don't significantly contribute to explaining the output.

Performance eval: regression

Remember, when using GridSearchCV:



```
from sklearn.linear_model import ElasticNet
base_elastic_model = ElasticNet()

param_grid = {'alpha':[0.1,1,5,10,50,100],
              'l1_ratio':[.1, .5, .7, .9, .95, .99, 1]}

from sklearn.model_selection import GridSearchCV
grid_model = GridSearchCV(estimator=base_elastic_model,
                          param_grid=param_grid,
                          scoring='neg_mean_squared_error',
                          cv=5,
                          verbose=2)
```

The scoring parameter can be any of the previously mentioned scores

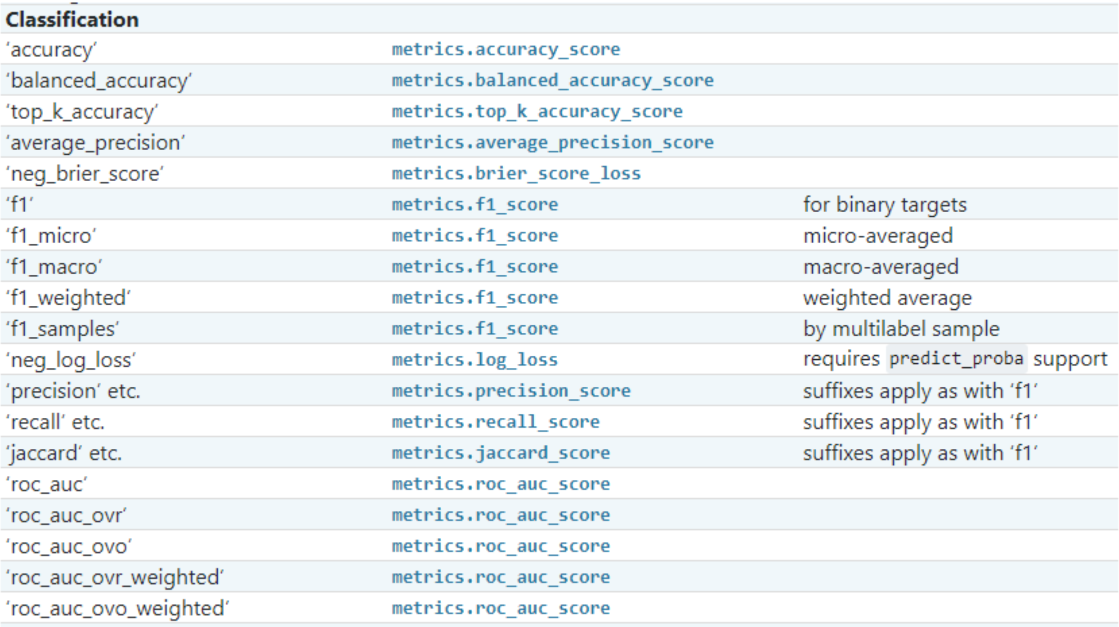
-> different results

Performance Eval: Classification

Most Common:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC

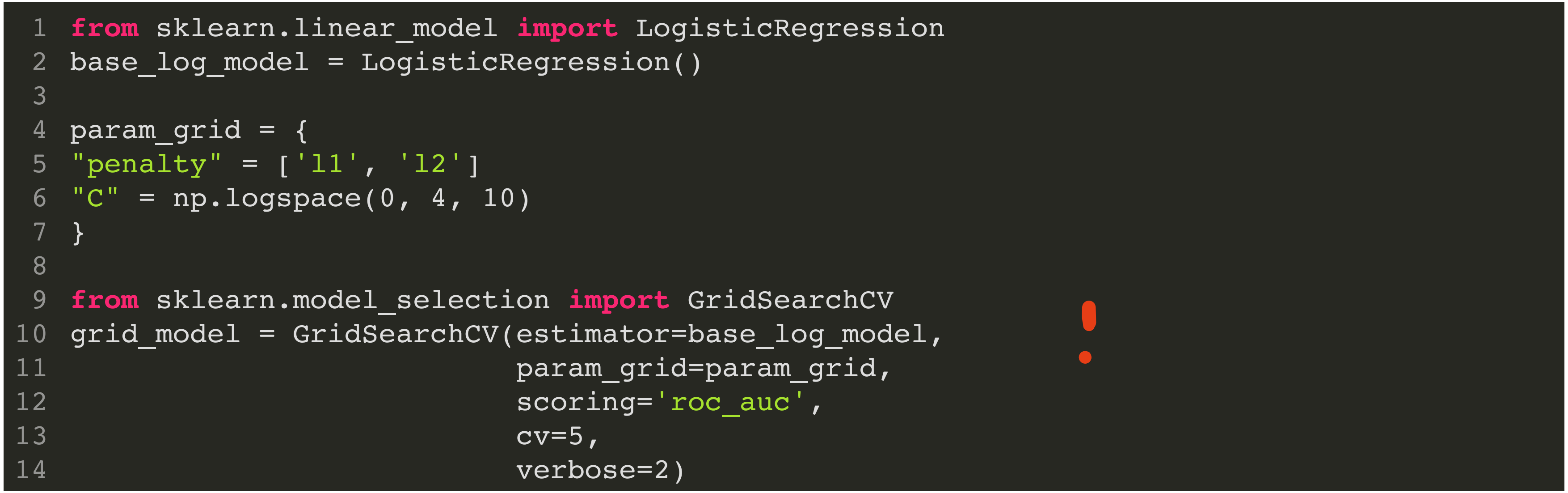
In Scikit-learn:



Classification		
'accuracy'	`metrics.accuracy_score`	
'balanced_accuracy'	`metrics.balanced_accuracy_score`	
'top_k_accuracy'	`metrics.top_k_accuracy_score`	
'average_precision'	`metrics.average_precision_score`	
'neg_brier_score'	`metrics.brier_score_loss`	
'f1'	`metrics.f1_score`	for binary targets
'f1_micro'	`metrics.f1_score`	micro-averaged
'f1_macro'	`metrics.f1_score`	macro-averaged
'f1_weighted'	`metrics.f1_score`	weighted average
'f1_samples'	`metrics.f1_score`	by multilabel sample
'neg_log_loss'	`metrics.log_loss`	requires `predict_proba` support
'precision' etc.	`metrics.precision_score`	suffixes apply as with 'f1'
'recall' etc.	`metrics.recall_score`	suffixes apply as with 'f1'
'jaccard' etc.	`metrics.jaccard_score`	suffixes apply as with 'f1'
'roc_auc'	`metrics.roc_auc_score`	
'roc_auc_ovr'	`metrics.roc_auc_score`	
'roc_auc_ovo'	`metrics.roc_auc_score`	
'roc_auc_ovr_weighted'	`metrics.roc_auc_score`	
'roc_auc_ovo_weighted'	`metrics.roc_auc_score`	

Performance Eval: Classification

Remember, when using GridSearchCV:



```python
from sklearn.linear_model import LogisticRegression
base_log_model = LogisticRegression()

param_grid = {
"penalty" = ['l1', 'l2']
"C" = np.logspace(0, 4, 10)
}

from sklearn.model_selection import GridSearchCV
grid_model = GridSearchCV(estimator=base_log_model,
                          param_grid=param_grid,
                          scoring='roc_auc',
                          cv=5,
                          verbose=2)
```

!

The scoring parameter can be any of the previously mentioned scores

-> different results

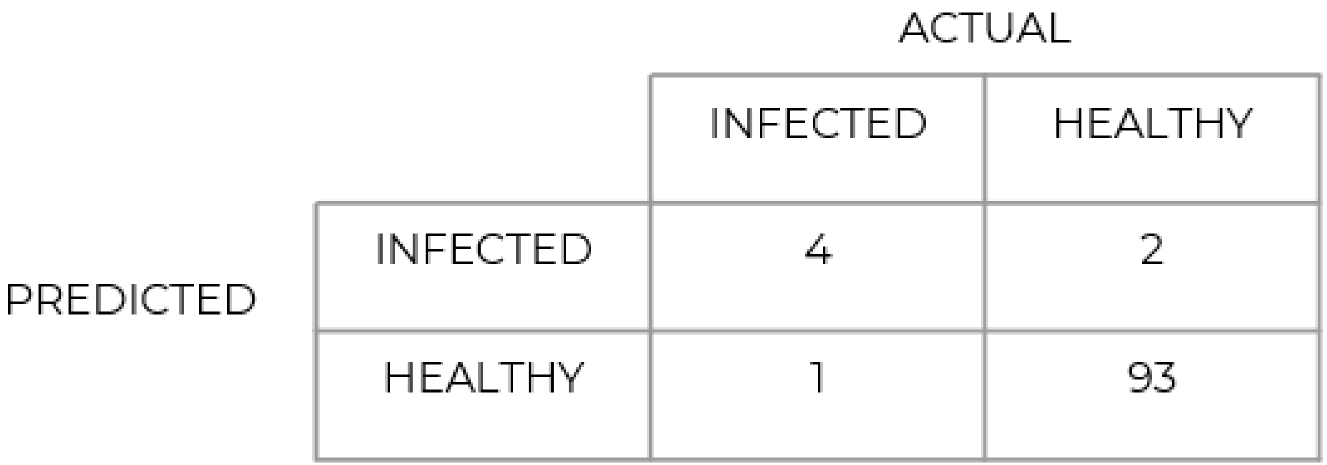
Confusion Matrices

- Imagine we've developed a test or model to detect presence of a virus infection in a person based on some biological feature.
- We could treat this as a Binary Classification, predicting:
  - 0 - Not Infected (Tests Negative)
  - 1 - Infected (Tests Positive)
- It is unlikely our model will perform perfectly. There are 4 possible outcomes:
  - Infected person tests positive.
  - Healthy person tests negative.
  - Infected person tests negative.
  - Healthy person tests positive.

Confusion Matrices

Imagine a test group of 100 people:

- 5 are infected. 95 are healthy.

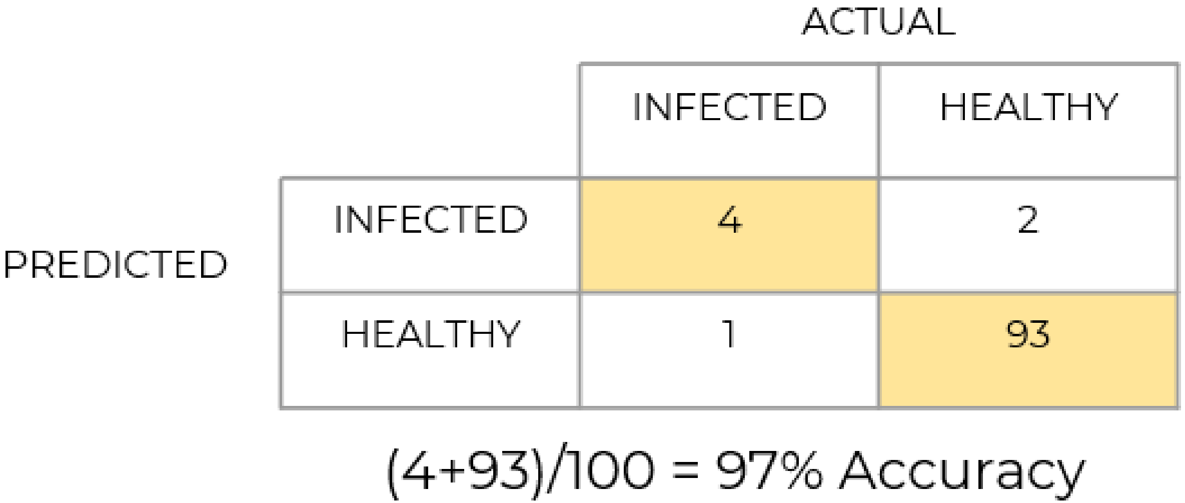


		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Accuracy

Accuracy:

- How often is the model correct?



		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

(4+93)/100 = 97% Accuracy

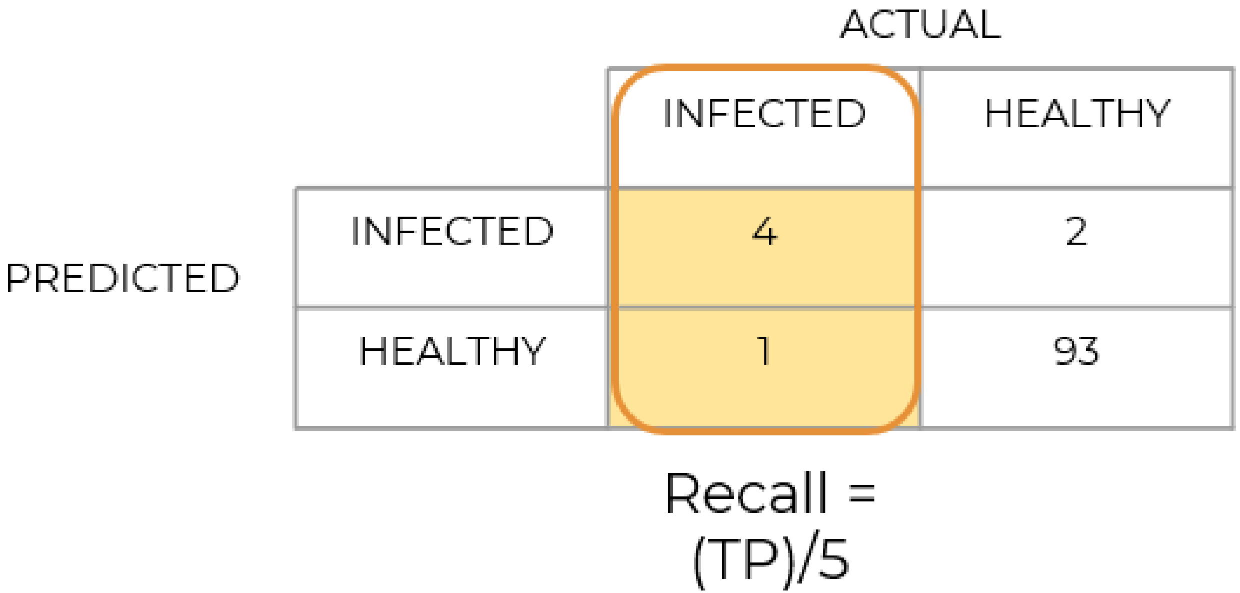
Accuracy

- This is the accuracy paradox!
  - Any classifier dealing with **imbalanced** classes has to confront the issue of the accuracy paradox.
  - **Imbalanced** classes will always result in a distorted accuracy reflecting better performance than what is truly warranted.

=> If a class is only a small percentage (**n%**), then a classifier that always predicts the majority class will always have an accuracy of (1-n).

Recall

- Recall:
  - When it actually is a positive case, how often is it correct?
  - (TP)/Total Actual Positives

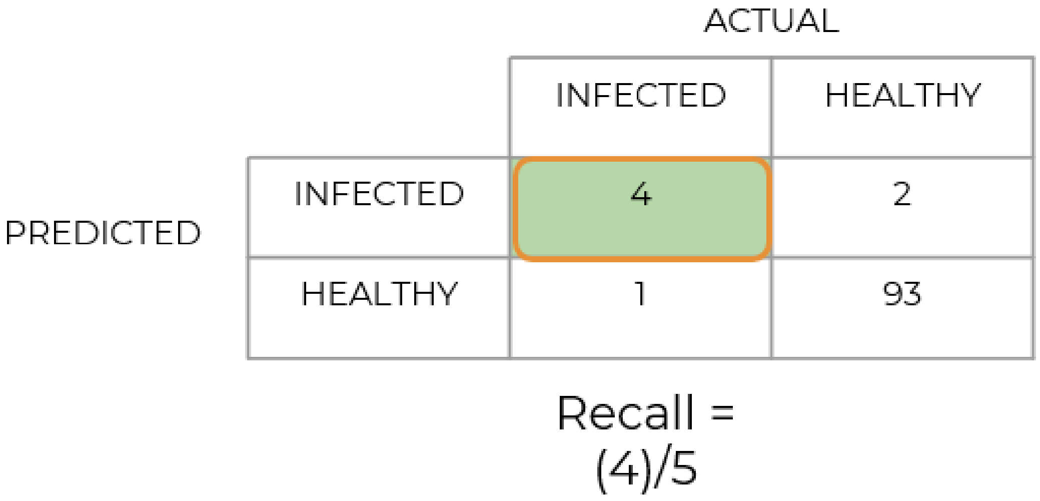


		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Recall =
(TP)/5

Recall

- Recall:
  - How many relevant cases are found?
  - (TP)/Total Actual Positives

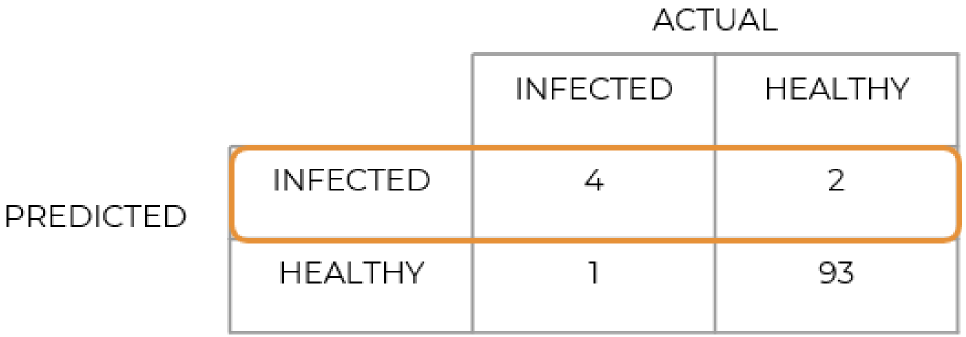


		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Recall =
(4)/5

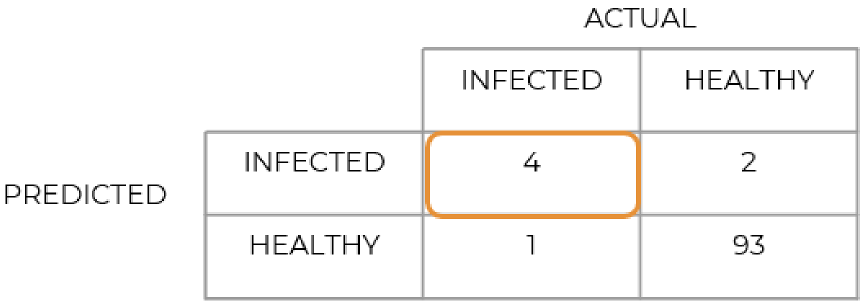
Precision

- Precision:
  - When prediction is positive, how often is it correct
  - (TP)/Total Predicted Positives



		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Precision =
(TP)/6



		ACTUAL	
		INFECTED	HEALTHY
PREDICTED	INFECTED	4	2
	HEALTHY	1	93

Precision =
(4)/6

F1-score

- Recall and Precision can help illuminate our performance specifically in regards to the relevant or positive case.
- Since precision and recall are related to each other through the numerator (TP), we often also report the F1-Score, which is the harmonic mean of precision and recall.The harmonic mean (instead of the normal mean) allows the entire harmonic mean to go to zero if **either** precision or recall ends up being zero.

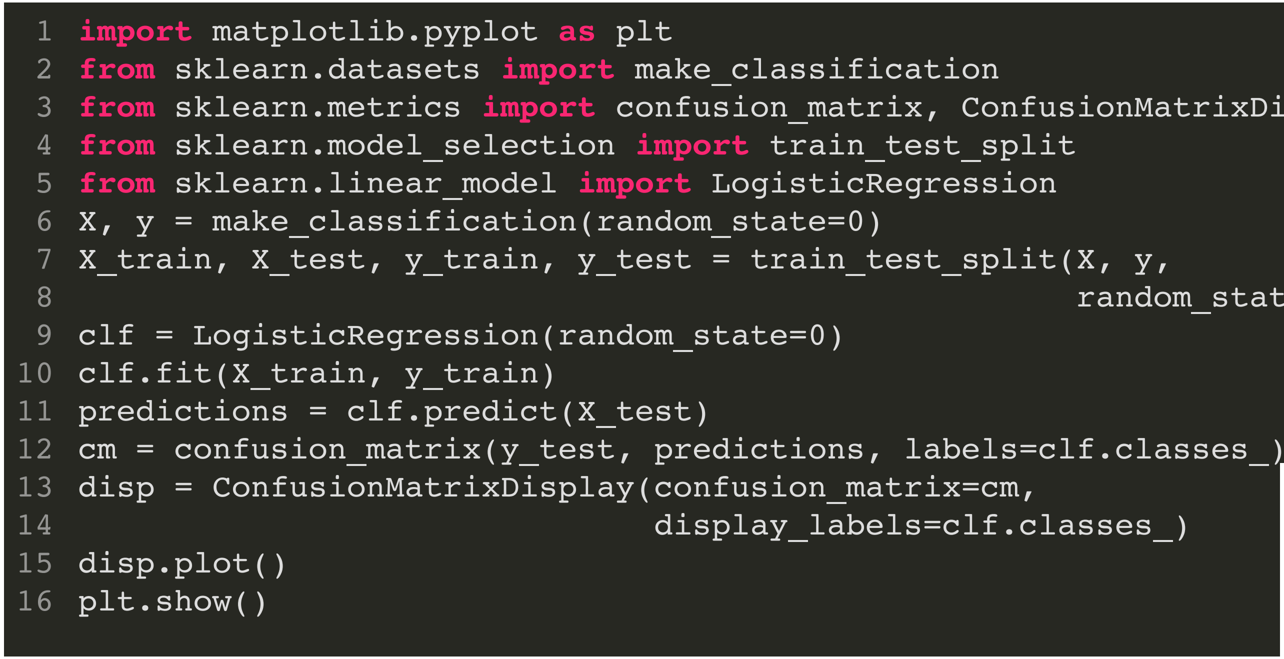
$$F = \frac{2 \times precision \times recall}{precision + recall}$$

F1-score

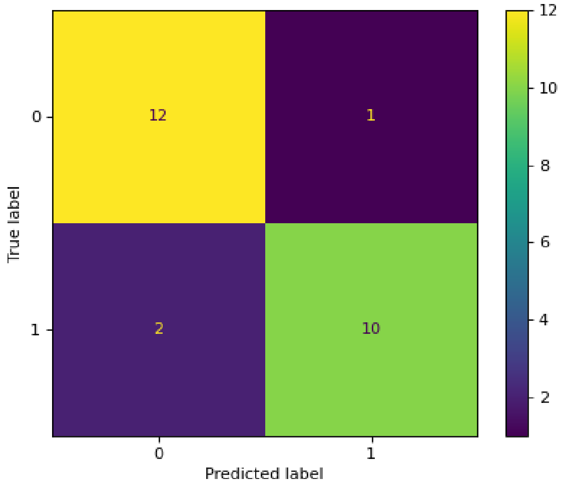
- Recall and Precision can help illuminate our performance specifically in regards to the relevant or positive case.
- Since precision and recall are related to each other through the numerator (TP), we often also report the F1-Score, which is the harmonic mean of precision and recall.The harmonic mean (instead of the normal mean) allows the entire harmonic mean to go to zero if **either** precision or recall ends up being zero.

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

Confusion Matrices in Sklearn



```python
import matplotlib.pyplot as plt
from sklearn.datasets import make_classification
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
X, y = make_classification(random_state=0)
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    random_state=0)
clf = LogisticRegression(random_state=0)
clf.fit(X_train, y_train)
predictions = clf.predict(X_test)
cm = confusion_matrix(y_test, predictions, labels=clf.classes_)
disp = ConfusionMatrixDisplay(confusion_matrix=cm,
                              display_labels=clf.classes_)
disp.plot()
plt.show()
```



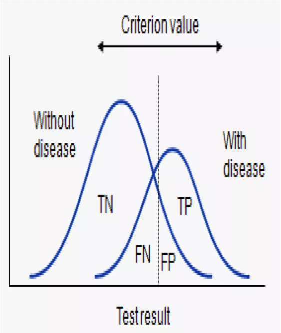
ROC Curves - definition

A receiver operating characteristic curve (ROC) curve is graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination is varied.

A ROC cirve is a way to compare diagnostic tests. It is a plot of the true positive rate against the false

ROC Curves - theory

When considered two populations' results of a disease, you will rarely observe a perfect separation between the two groups. The distribution of the tests results overlap, as shown in the following figure.



For every possible cut-off point on criterion value you select to discriminate between the two populations, there will some cases with disease correctly classified as positive (TP - True Positive), but some cases with the disease will be incorrectly classified as negative (FN - False Negative). On the other hand, some cases without the disease will be correcly classified as negative (TN - True Negative), finally, some cases without the disease will be classified as positive (FP - False positive)

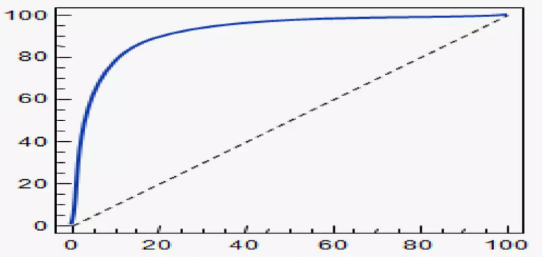
ROC Curves - history

- The name "Receiver Operating Charactersitic" came signal detection theory developed during World War II for the analysis of rada images.
- Radar operators had to decide whether a blip on the screen represented an enemy target, a friendly ship or just noise.
- Signal detection measures the ability of radar receiver operators to make these important directions.
- Their ability to do so was called: Receiver operating characteristic.

ROC Curves - meaning

- ROC plot shows... the relationship between sensitivity and specicity. For example, a decrease in sensitivity results in an increase in specificity. It also shows the test accuracy; the closer the graph is to the top and left-hand borders, the more accurate the test. Likewise, the closer the graph to the diagonal, the less accurate the test.
- A perfect test would go straight from zero up the top-left corner and then straight across the horizontal. The likelihood ration; given by the derivative at any particular cut point.

True positive sensitivity)



False positive (specificity)

ROC Curves - meaning

- ROC plot shows... trade-offs between sensitivity anf specificity. The ROC plot is a model-wide evaluation that is based on two basic evaluation measures - specificity and sensitivity.
- Specificity is a performance measure of the whole negative part of a dataset, whereas sensitivity is a performance measure of the whole positive part.

ROC Curves - making the curve

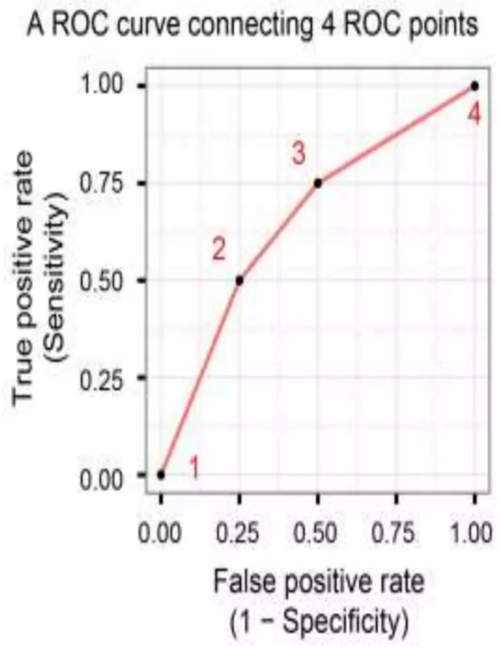
- A ROC point is a point with a pair of x and y values in the ROC space where x is 1 a - specificity anf y is sensitivity.
- A ROC curve is created by connecting al ROC points of a classifier in the ROC space. Two adjacent ROC points can be connected by a straight line, and the curve starts at (0.0, 0.0) and ends at (1.0, 1.0).

ROC Curves - making the curve

- Consider the following example, to make a ROC curve by connecting several ROC points. Let us assume that we have calculated sensitivity and specificity values from multiple confusion matrices for four different threshold values.

Threshold	**Sensitivity**	**Specificity**	**1 specificity**
1	0.0	1.0	0.0
2	0.5	0.75	0.25
3	0.75	0.5	0.5
4	1.0	0.0	1.0

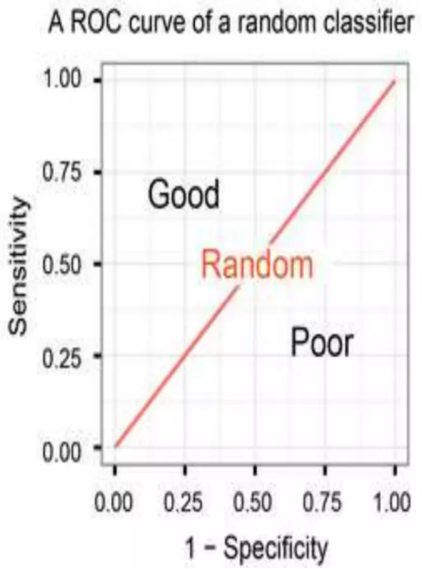
ROC Curves - making the curve



- We first added four points that matches with the pairs of sensitivity and specificity values and then connected the points to create a ROC curve.
- The plot shows a ROC curve connecting four ROC points.

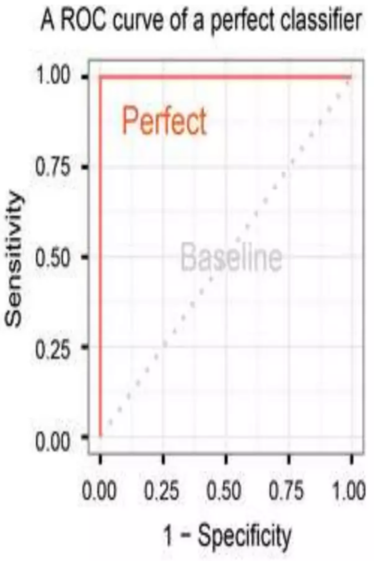
ROC Curves - interaction of ROC curves with a classifier

- A classifier with the random performance level always shows a straight line from the origin (0.0, 0.0) to the top right corner (1.0, 1.0).
- Two areas spearated by this ROC curve indicates a simple estimation of the performance level. ROC curves in the area with the top left corner (0.0, 1.0) indicate good performance levels. Whereas ROC curves in the other area with the bottom right corner (1.0, 0.0) indicate poor performance



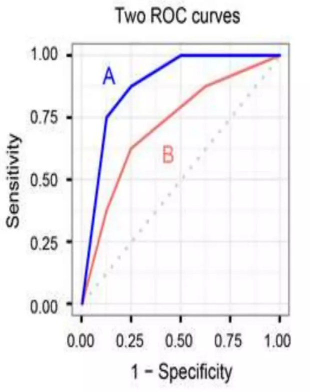
ROC Curves - interaction of ROC curves with a classifier

- A ROC curve represents a classifier with the random performance level. The curve separates the soace into two areas for good and poor performance levels
- A classifier with the perfect performance level shows a combination of two straight lines - from the origin (0.0, 0.0) to the top left (0.0, 1.0) and further to the top right corner (1.0, 1.0).
- It is important to notice that classifiers with meaningful performace levels usually lie in the area between the random



ROC Curves - multiple models

- Comparison of multiple classifiers is usually straight forward especially when no curves cross each other. Curves close to the perfect ROC curve have a better performance level than the ones clases to the baseline
- Two ROC curves represent the performance levels of two classifiers A and B. Classifier A clearly outperforms classifier B.

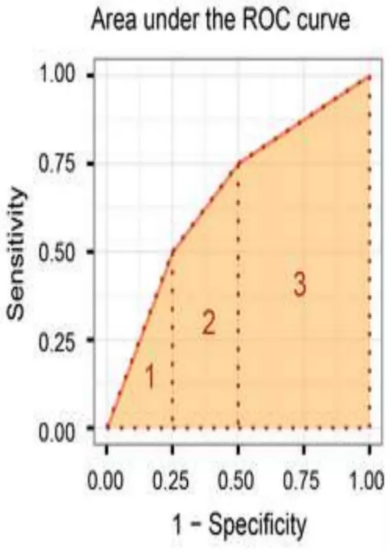


ROC Curves - area under ROC curve score

- Another advantage of using ROC plot is a single measure called the auc (area under the ROC curve) score. As the name indicates, it is an area under the curve calculated in the ROC space.
- One of the easy ways to calculate the AUC score is using the trapezoidal rule, which is adding up all trapezoids under the curve
- The AUC score can be calculated by the trapezoidal rule, which is adding up all trapezoids under the curve. The areas of the three trapezoids 1,2,3 are 0.0625, 0.15625 and 0.4375. The AUC score is then 0.65625

ROC Curves - area under ROC curve score

- The AUC score can be calculated by the trapezoidal rule, which is adding up all trapezoids under the curve. The areas of the three trapezoids 1,2,3 are 0.0625, 0.15625 and 0.4375. The AUC score is then 0.65625.

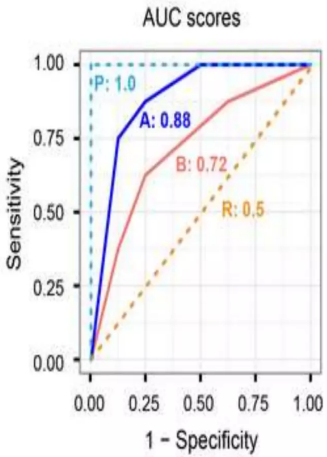


ROC Curves - area under ROC curve score

- Althoughthe theoretical range of AUC score is between 0 and 1, the actual scores of meaningful classifiers are greater than 0.5, which is the AUC score of a random classifier.
- It shows four AUC scores. The score is 1.0 for the classifier with the perfect performance level(P) and 0.5 for the classifier with the random performance level (R) ROC curves clearly shows classifier A outperforms classifier B, which is also supported by their AUC scores (0.88 and 0.72)

ROC Curves - area under ROC curve score

- It shows four AUC scores. The score is 1.0 for the classifier with the perfect performance level(P) and 0.5 for the classifier with the random performance level (R) ROC curves clearly shows classifier A outperforms classifier B, which is also supported by their AUC scores (0.88 and 0.72).



Performance Evaluation (ROC)

A classifier assigns an object to one of a predefined set of categories or classes.

Examples:

A metal detector either sounds an alarm or stays quiet when someone walks through.

A credit card application is either approved or denied.

A medical test's outcome is either positive or negative.

This talk: only two classes, “positive” and “negative”.

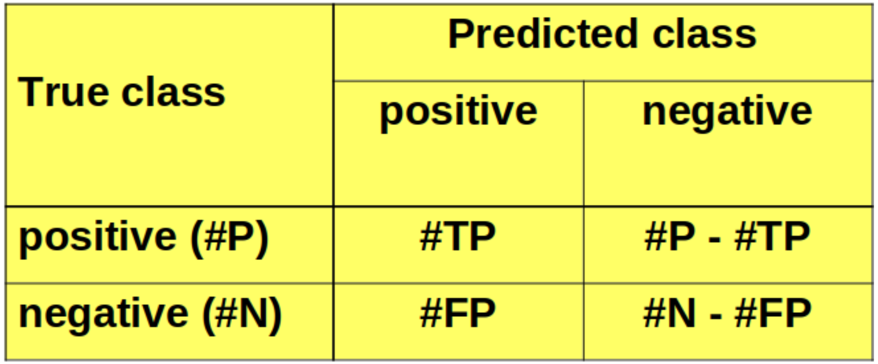
Types of error

False positive (“false alarm”), FP
alarm sounds but person is not carrying metal

False negative (“miss”), FN
alarm doesn’t sound but person is carrying metal

Reduce the 4 numbers to two rates
true positive rate = TP = (#TP)/(#P)
false positive rate = FP = (#FP)/(#N)

Rates are independent of class ratio*



True class	Predicted class	
	positive	negative
positive (#P)	#TP	#P - #TP
negative (#N)	#FP	#N - #FP

Types of error

True	Predicted	
	pos	neg
pos	40	60
neg	30	70

Classifier 1
TP = 0.4
FP = 0.3

True	Predicted	
	pos	neg
pos	70	30
neg	50	50

Classifier 2
TP = 0.7
FP = 0.5

True	Predicted	
	pos	neg
pos	60	40
neg	20	80

Classifier 3
TP = 0.6
FP = 0.2

Assumptions

Standard Cost Model

correct classification costs 0

cost of misclassification depends only on the class, not on the individual example

over a set of examples costs are additive

Costs or Class Distributions:

are not known precisely at evaluation time

may vary with time

may depend on where the classifier is deployed

Assumptions

True FP and TP do not vary with time or location, and are accurately estimated.correct classification costs 0

cost of misclassification depends only on the class, not on the individual example

over a set of examples costs are additive

Costs or Class Distributions:

are not known precisely at evaluation time

may vary with time

may depend on where the classifier is deployed

True FP and TP do not vary with time or location, and are accurately estimated.

How to evaluate performance?

Scalar Measures

- Accuracy
- Expected cost
- Area under the ROC curve

Visualization Techniques

- ROC curves
- Cost Curves

How to evaluate performance?

NEVERTHELESS…

How to evaluate performance?

A scalar does not tell the whole story.

There are fundamentally two numbers of interest (FP and TP), a single number invariably loses some information.

How are errors distributed across the classes ?

How will each classifier perform in different testing conditions (costs or class ratios other than those measured in the experiment) ?

A scalar imposes a linear ordering on classifiers.

what we want is to identify the conditions under which each is better.

How to evaluate performance?

A table of scalars is just a mass of numbers.

- No immediate impact
- Poor way to present results in a paper
- Equally poor way for an experimenter to analyze results

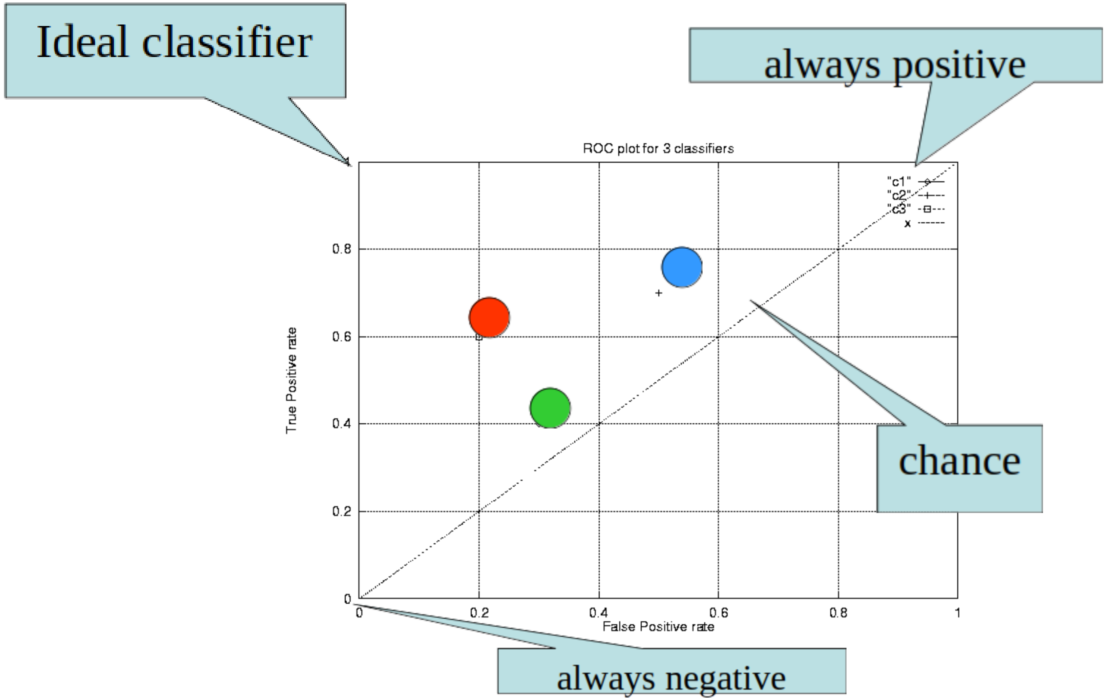
Some scalars (accuracy, expected cost) require precise knowledge of costs and class distributions.

- Often these are not known precisely and might vary with time or location of deployment.

Do we need to visualize performance?

- Shape of curves more informative than a single number
  - Curve informs about
  - all possible misclassification costs*
  - all possible class ratios*
  - under what conditions C1 outperforms C2
- Immediate impact (if done well)

Do we need to visualize performance?



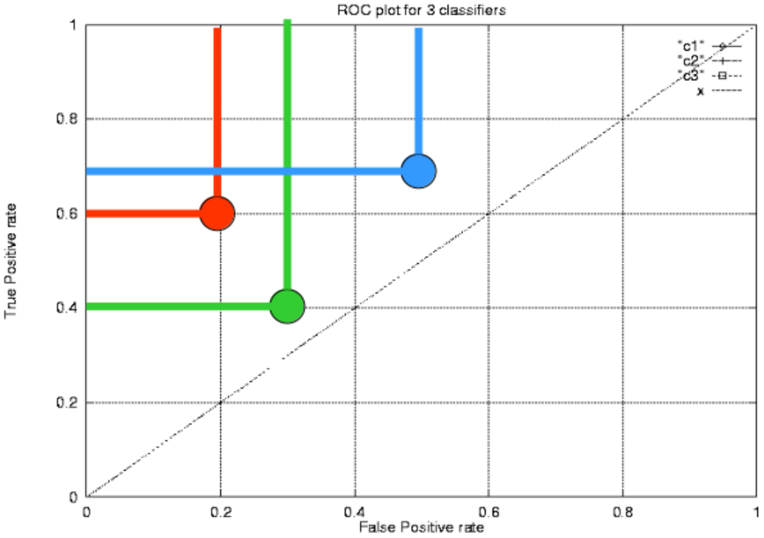
Ideal classifier

always positive

chance

always negative

Do we need to visualize performance?



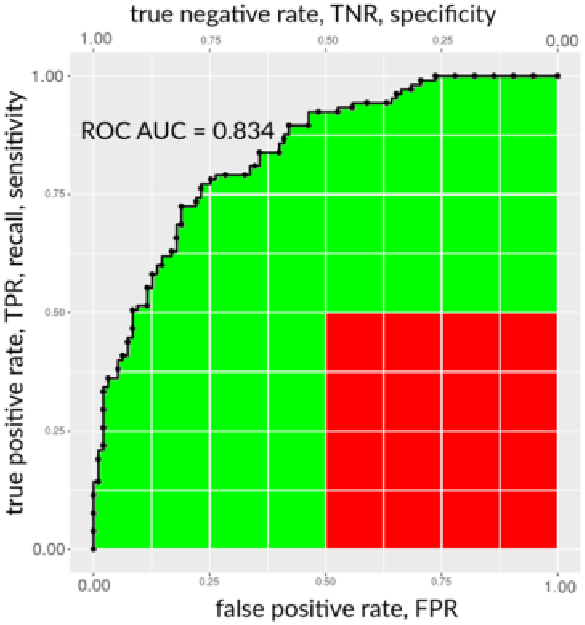
ROC Curve

A classifier produces a single ROC point.

If the classifier has a “sensitivity” parameter, varying it produces a series of ROC points (confusion matrices).

Alternatively, if the classifier is produced by a learning algorithm, a series of ROC points can be generated by varying the class ratio in the training set.

ROC Curve - What's wrong with ROC Curves



The main criticism to the ROC curve regards the incorporation of areas with low sensitivity and low specificity (both lower than 0.5) for the calculation of the total area under the curve (AUC), as described in the plot on the left.

Multiclass ROC curves

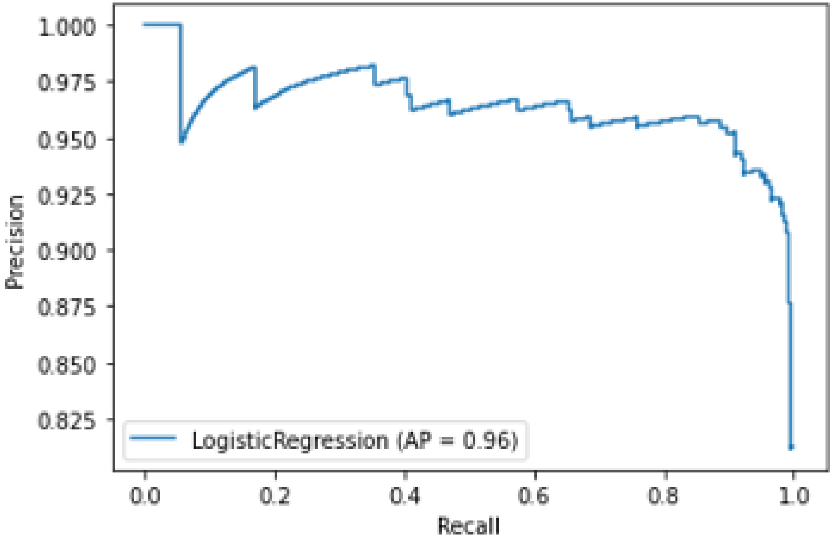
- How can we plot ROC curves, when it's not about binary classification, but multi-class classification?
  - One-vs-Rest multiclass ROC
  - One-vs-One multiclass ROC (And then average them by a certain measurement)

More info:

Multiclass Receiver Operating Characteristic (ROC) — scikit-learn 1.3.1 documentation

Precision-Recall Curve

- Can also create precision vs. recall curves:



AP = Average Precision