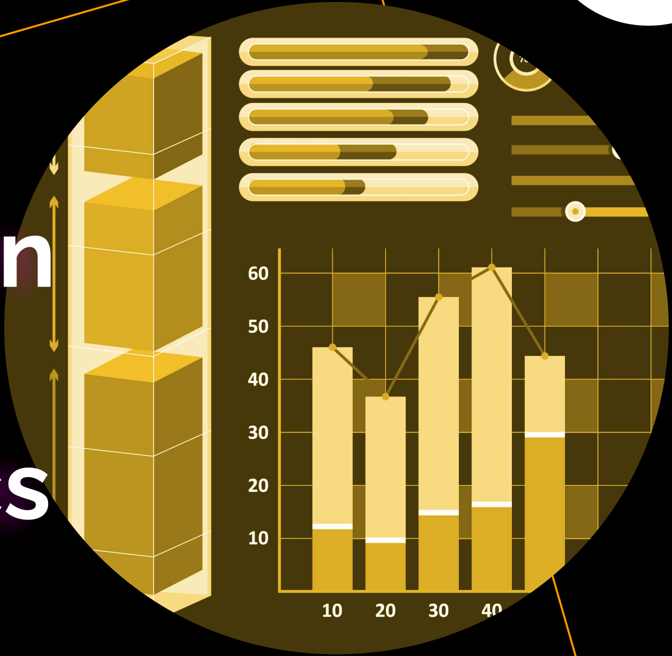
Train, Test, Validation datasets & Performance metrics



Date / Người giảng dạy

Train, Test, Validation datasets

Definition

- **Training Dataset**: The sample of data used to fit the model.
- **Validation Dataset**: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.
- **Test Dataset**: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

Train/Val/Test Splits

Common ratios of three subsets: the training set, the validation set, and the test set used are:

- 70% train, 15% val, 15% test
- 80% train, 10% val, 10% test
- 60% train, 20% val, 20% test

Train/Val/Test Splits

```
from sklearn.model_selection import train_test_split

X, y = np.arange(10).reshape((5, 2)), range(5)

X_train, X_test, y_train, y_test
    = train_test_split(X, y, test_size = 0.2, random_state = 1)

X_train, X_val, y_train, y_val
    = train_test_split(X_train, y_train, test_size = 0.25, random_state = 1) # 0.25 x 0.8 =
0.2
```

Performance Metrics

Performance Metrics

After doing the usual Feature Engineering, Selection, and of course, implementing a model and getting some output in forms of a probability or a class, the next step is to find out how effective is the model based on some metric using test datasets. Different performance metrics are used to evaluate different Machine Learning Algorithms.

Several metrics are used to evaluate classification and regression algorithms. Some metrics for classification: precision, recall, sensitivity, specificity, F-measure, Matthews correlation, etc. They are all based on the confusion matrix. Others exist for regression (continuous output variable): square error, absolute error, etc. The technique is mostly to run an algorithm on some data to get a model, and then apply that model on new, previously unseen data, and evaluate the metric on that data set, and repeat. Some techniques (actually resampling techniques from statistics): Jacknife, Cross validation, K-fold validation, bootstrap.

Accuracy and Loss

Accuracy

Accuracy = n/N

i. n: number of ‘true’ prediction

ii. N: Total number of prediction

Loss

Commonly a nonnegative number, different between value of real data and our prediction.

Confusion Matrix

Model dự đoán

Thực tế

	Bị Covid	Không bị Covid
Bị Covid	980	0
Không bị Covid	10	10

Confusion Matrix

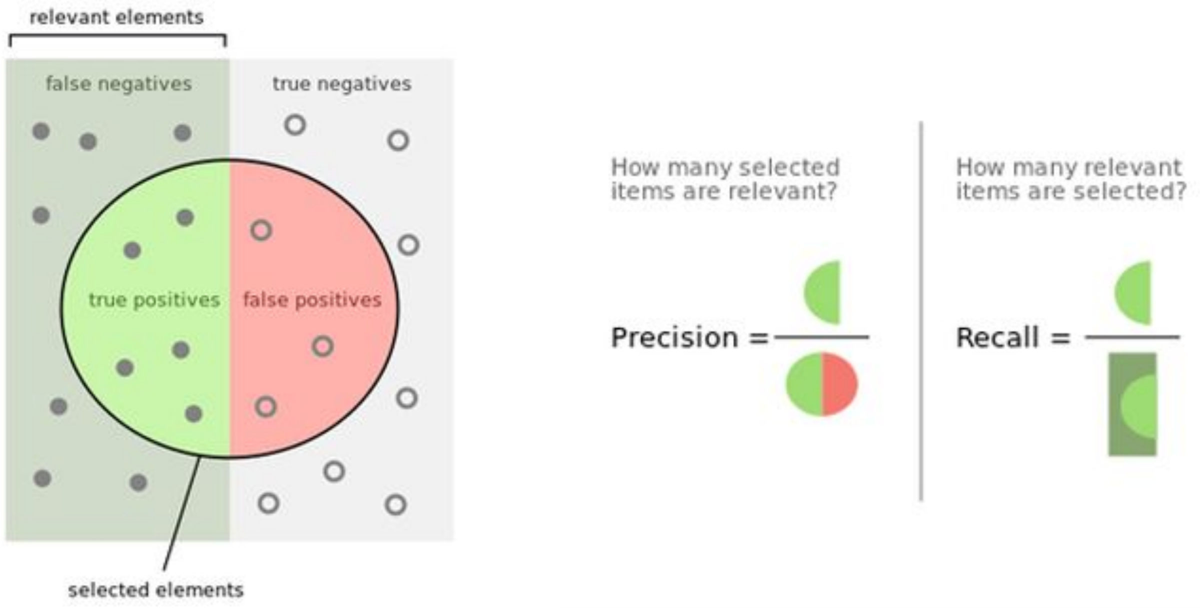
Model dự đoán (Predicted)

Thực tế (Actual)

	Positive	Negative
Positive	True Positive Dự đoán Đúng là Positive	False Negative Dự đoán Sai là Negative
Negative	False Positive Dự đoán Sai là Positive	True Negative Dự đoán Đúng là Negative

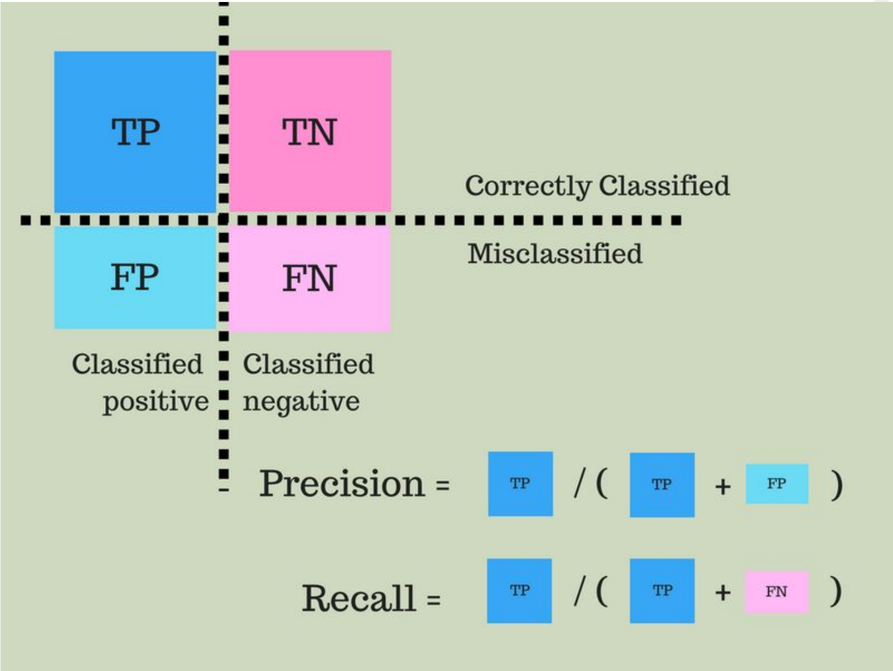
Precision & Recall

Source:
https://www.digital-mr.com/media/cache/5e/b4/5eb4dbc50024c306e5f707736fd79c1e.png



Precision & Recall

Source: https://nlpforhackers.io/wp-content/uploads/2017/01/Precision-Recall.png



F1-Score, ROC Curve

F1=2 (Precision∗Recall)/(Presicion+Recall)

ROC (Receiver Operating Characteristic) curve

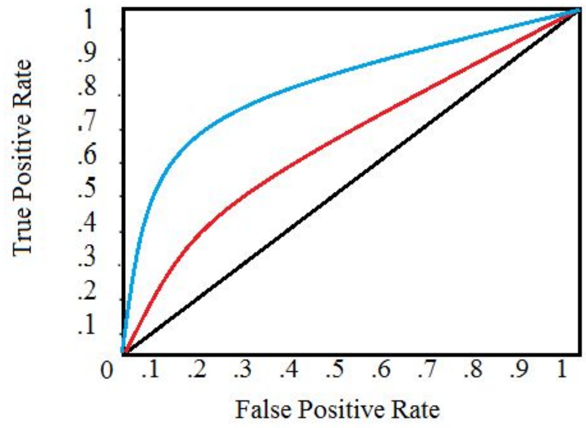
-True Positive Rate (TPR) = Recall

-False Positive Rate (FPR) = FP/(FP+TN)

-

ROC curve shows the relationship between TPR and FRR

AUC (Area Under ROC Curve)



Source: https://www.statisticshowto.com/wp-content/uploads/2016/08/ROC-curve.png

Source: http://www.damienfrancois.be/blog/files/modelperfcheatsheet.pdf

Confusion matrix for two possible outcomes *p* (positive) and *n* (negative)

		Actual p	Actual n	Total
Predicted	p'	**true positive**	**false postive**	P
	n'	**false negative**	**true negative**	N
	total	P'	N'	

Classification accuracy

(TP + TN) / (TP + TN + FP + FN)

Error rate

(FP + FN) / (TP + TN + FP + FN)

Paired criteria

Precision: (or Positive predictive value) proportion of predicted positives which are actual positive

TP / (TP + FP)

Recall: proportion of actual positives which are predicted positive

TP / (TP + FN)

Sensitivity: proportion of actual positives which are predicted positive

TP / (TP + FN)

Specificity: proportion of actual negative which are predicted negative

TN / (TN + FP)

True positive rate: proportion of actual positives which are predicted positive

TP / (TP + FN)

True negative rate: proportion of actual negative which are predicted negative

TN / (TN + FP)

Positive likelihood: likelihood that a predicted positive is an actual positive

sensitivity / (1 - specificity)

Negative likelihood: likelihood that a predicted negative is an actual negative

(1 - sensitivity) / specificity

Combined criteria

BCR: Balanced Classification Rate

½ (TP / (TP + FN) + TN / (TN + FP))

BER: Balanced Error Rate, or **HTER**: Half Total Error Rate: 1 - BCR

F-measure harmonic mean between precision and recall

2 (precision . recall) / (precision + recall)

F_β-measure weighted harmonic mean between precision and recall

$(1+\beta)^2$ TP / ($(1+\beta)^2$ TP + β^2 FN + FP)

The harmonic mean between specificity and sensitivity is also often used and sometimes referred to as F-measure.

Youden's index: arithmetic mean between sensitivity and specificity

sensitivity - (1 - specificity)

Matthews correlation correlation between the actual and predicted

(TP . TN – FP . FN) / ((TP+FP) (TP+FN) (TP + FP) (TN+FN))$^{1/2}$

comprised between -1 and 1

Discriminant power normalised likelihood index

sqrt(3) / π .

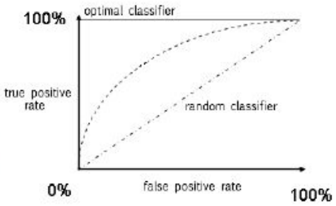
(log (sensitivity / (1 – specificity)) + log (specificity / (1 - sensitivity)))

<1 = poor, >3 = good, fair otherwise

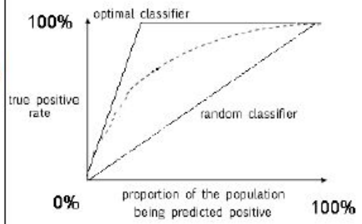
Graphical tools

ROC curve receiver operating characteristic curve : 2-D curve parametrized by one parameter of the classification algorithm, e.g. some threshold in the « true postivie rate / false positive rate » space

AUC The area under the ROC is between 0 and 1



(Cumlative) Lift chart plot of the true positive rate as a function of the proportion of the population being predicted positive, controlled by some classifier parameter (e.g. a threshold)



Relationships

sensitivity = recall = true positive rate

specificity = true negative rate

BCR = ½ . (sensitivity + specificity)

BCR = 2 . Youden's index - 1

F-measure = F_1measure

Accuracy = 1 – error rate

References

Sokolova, M. and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45, 4 (Jul. 2009), 427-437.

Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7 (2006) 1–30

Source: http://www.damienfrancois.be/blog/files/modelperfcheatsheet.pdf

Let $D = \{(x_i, y_i)\}$ be a set of input/output pairs and f a function such that for $i = 1..n$,

$$y_i = f(x_i) + \epsilon_i$$

Squared error

SSE Sum of Squared Errors, or
RSS Residual Sum of Squares

$$\sum_i \epsilon_i^2$$

MSE Mean Squared Error

$$\frac{1}{n}\sum_i \epsilon_i^2$$

RMSE Root Mean Squared Error

$$\sqrt{\frac{1}{n}\sum_i \epsilon_i^2}$$

NMSE Normalised Mean Squared Error

$$\frac{SSE}{var(\{y_i\})}$$

where var is the empirical variance in the sample.

R-squared

$$1 - \frac{SSE}{var(y_i)}$$

where var is the empirical variance in the sample

Absolute error

MAD Mean Absolute Deviation

$$\frac{1}{n}\sum|\epsilon_i|$$

MAPE Mean Absolute Percentage Error

$$\frac{1}{n}\sum_i \frac{|\epsilon_i|}{y_i}$$

Predicted error

PRESS Predicted REsidual Sums of Squares

$$\frac{1}{n}\|diag(XX^T)(XX^T - I)Y\|_2^2$$

where X is a matrix built by stacking the x_i in rows. Y is the vector of y_i

GCV Generalised Cross Validation

$$\frac{\frac{1}{n}\|(I - X(X^TX + nI)^{-1}X^T)Y\|^2}{(\frac{1}{n}Trace(I - X(X^TX + nI)^{-1}X^T)^2}$$

where X is a matrix built by stacking the x_i in rows. Y is the vector of y_i

Information criteria

AIC Akaike Information Criterion

$$n \log MSE + 2k$$

where k is the number of parameters in the model

BIC Bayesian Information Criterion

$$n \log MSE + k \cdot \log n$$

where k is the number of parameters in the model

Robust error measures

Median Squared error

$$median(\epsilon_i^2)$$

α -trimmed MSE

$$\frac{1}{\#I}\sum_{i \in I} \epsilon_i^2$$

where I is the set of residuals ϵ_i where α percents of the largest values are discarded.

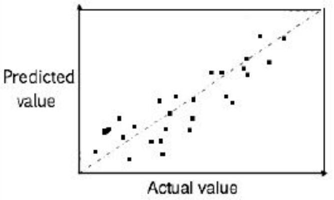
M-estimators

$$\frac{1}{n}\sum_i \rho(\epsilon_i)$$

where \rho is a non-negative function with a mininmum in 0, like the parabola, the Hubber function, or the bisquare function.

Graphical tool

Plot of predicted value against actual value. A perfect model places all dots on the diagonal.



Resampling methods

LOO – Leave-one-out: build the model on $n - 1$ data elements and test on the remaining one. Iterate n times to collect all ϵ_i and compute mean error.

X-Val – Cross validation. Randomly split the data in two parts, use the first one to build the model and the second one to test it. Iterate to get a distribution of the test error of the model.

K-Fold – Cut the data into K parts. Build the model on the K-1 first parts and test on the Kth one. Iterate from 1 to K to get a distribution of the test error of the model.

Bootstrap – Draw a random subsample of the data with replacement. Compute the error on the whole dataset minus the training error of the model and Iterate to get a distribution of such values. The mean of the distribution is the optimism. The bootstrap error estimate is the training error on the whole dataset plus the optimism.

PM: Classification

MaSSP

- *Binary classification*

```python
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
from sklearn.metrics import roc_auc_score
from sklearn.metrics import log_loss
X_actual = [1, 1, 0, 1, 0, 0, 1, 0, 0, 0]
Y_predic = [1, 0, 1, 1, 1, 0, 1, 1, 0, 0]
results = confusion_matrix(X_actual, Y_predic)
print ('Confusion Matrix :')
print(results)
print ('Accuracy Score is',accuracy_score(X_actual, Y_predic))
print ('Classification Report : ')
print (classification_report(X_actual, Y_predic))
print('AUC-ROC:',roc_auc_score(X_actual, Y_predic))
print('LOGLOSS Value is',log_loss(X_actual, Y_predic))
```

```
Confusion Matrix :
[[3 3]
 [1 3]]
Accuracy Score is 0.6
Classification Report :
              precision    recall  f1-score   support

           0       0.75      0.50      0.60         6
           1       0.50      0.75      0.60         4

    accuracy                           0.60        10
   macro avg       0.62      0.62      0.60        10
weighted avg       0.65      0.60      0.60        10

AUC-ROC: 0.625
LOGLOSS Value is 13.815750437193334
```

PM: Regression

```
from sklearn.metrics import r2_score
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
X_actual = [5, -1, 2, 10]
Y_predic = [3.5, -0.9, 2, 9.9]
print ('R Squared =',r2_score(X_actual, Y_predic))
print ('MAE =',mean_absolute_error(X_actual, Y_predic))
print ('MSE =',mean_squared_error(X_actual, Y_predic))
```

```
R Squared = 0.9656060606060606
MAE = 0.42499999999999993
MSE = 0.5674999999999999
```

MaSSP

Appendix

References

- *What is the Difference Between Test and Validation Datasets?*
- *Best Use of Train/Val/Test Splits, with Tips for Medical Data*
- https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
- https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i
- https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics
- "Giới thiệu về k-fold cross-validation," *Trí tuệ nhân tạo*, Ngày xuất bản: 30/01/2020, URL: https://trituenhantao.io/kien-thuc/gioi-thieu-ve-k-fold-cross-validation/, Ngày truy cập: 09/06/2021.
- http://www.jaist.ac.jp/~bao/VNAlectures/Evaluation-TQKhoat%20(A6-A7).pdf
- http://www.jaist.ac.jp/~bao/VNAlectures/