Model Evaluation

Thanh-Nghi Do

Can Tho University

dtnghi@cit.ctu.edu.vn

Can Tho
Dec. 2019

Content

- Introduction
- Protocol
- Performance metrics

Content

- <u>**Introduction**</u>
- Protocol
- Performance metrics

Introduction

- How well is learning model doing?
  - How the model generalizes on unseen data
  - High generalization accuracy or low generalization error
  - Error on the training data is not a good indicator of performance on future data
  - New data (unseen data) will probably not be exactly the same as the training data!
  - Overfitting: the model fits the training data too precisely - usually leads to poor results on new data
  - Underfitting: the model does not fit the training data well enough - usually leads to poor results on training and test data

Introduction

- Possible evaluation measures
  - Classification Accuracy
  - Total cost/benefit - different errors involve different costs
  - Lift and ROC curves
  - Error in numeric predictions
- How reliable are the predicted results ?

Introduction

- Natural performance measure for classification problems: error rate
  - Success: example's class is predicted correctly
  - Error: example's class is predicted incorrectly
  - Error rate: proportion of errors made over the whole set of examples
- Training set error rate: is way too optimistic!

Content

- Introduction
- **Protocol**
- Performance metrics

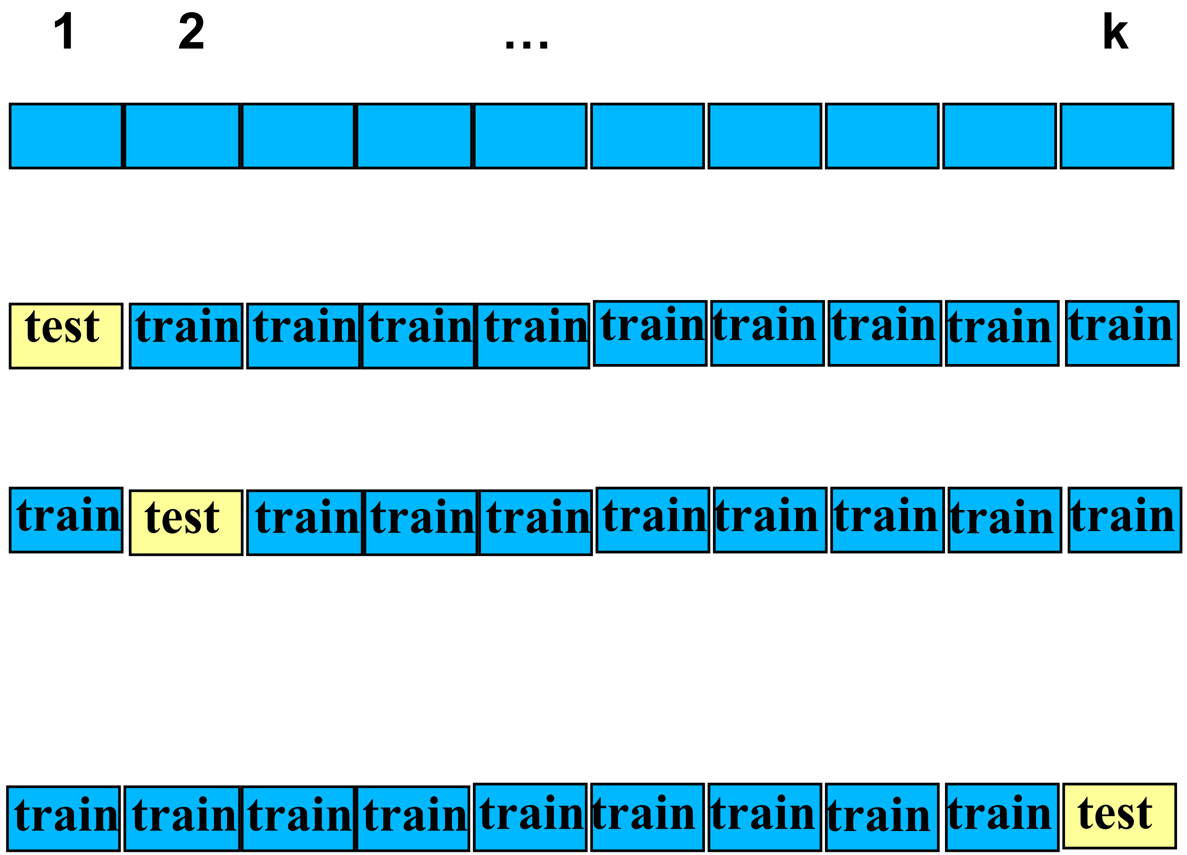
Protocol

- Given **dataset**: **trainset**, **testset**
  - Using **trainset** to build model **M**
  - Reporting the result on **testset** using resulting model **M**

Protocol: cross-validation

- *k*-fold (**dataset, #examples > 300**)
  - *k*-fold
  - it = *1*
  - it = *2*

    …
  - it = *k*



Protocol: cross-validation

- Given **dataset**
- Protocol: **hold-out**
  - Randomly split **dataset** into training set (2/3 for **trainset**) and test sets (1/3 for **testset**)
- Protocol **leave-1-out**
  - *k*-fold: **k = #examples**
- Protocol **bootstrap .632**
  - Bootstrap **B**: sampling with replacement from **dataset**
  - Out-of-bag **OOB**: examples in **dataset** are left out of the bootstrap **B**
  - **Err(M, dataset) = 0,382 Err(M, OOB) + 0,632 Err(M, B)**

Content

- Introduction
- Protocol
- **Performance metrics**

Confusion matrix (C) for *k* classes

prediction=>	1	…	k
1			
…			
k			

- C[i, j]: #examples in class **i** are predicted in class **j**
- C[i, i]: #examples in class **i** are correctly predicted
- Accuracy for class **i**: C[i,i] / C[i,]
- Global accuracy: ∑C[i,i] / C

Confusion matrix (C) for *2* classes (+1/-1)

Prediction =>	**+1**	**-1**
+1	TP	FN
-1	FP	TN

$$prec = \frac{tp}{tp + fp}$$

$$rec = \frac{tp}{tp + fn} = \frac{tp}{pos}$$

Confusion matrix (C) for *2* classes (+1/-1)

Prediction =>	**+1**	**-1**
+1	TP	FN
-1	FP	TN

$$acc = \frac{tp + tn}{tp + fn + tn + fp} = \frac{tp + tn}{pos + neg}$$

$$F1 = \frac{2 \times prec \times rec}{prec + rec}$$

$$bep = \frac{prec + rec}{2}$$

Classification of imbalanced datasets

- Imbalanced dataset
  - Proportion of positive class < 10% (minority class)
  - Proportion of negative class > 90% (majority class)
- Classification of imbalanced datasets
  - Model misclassifies minority class
  - Performance metrics: F1, ROC (Receiver Operating Characteristic)

ROC

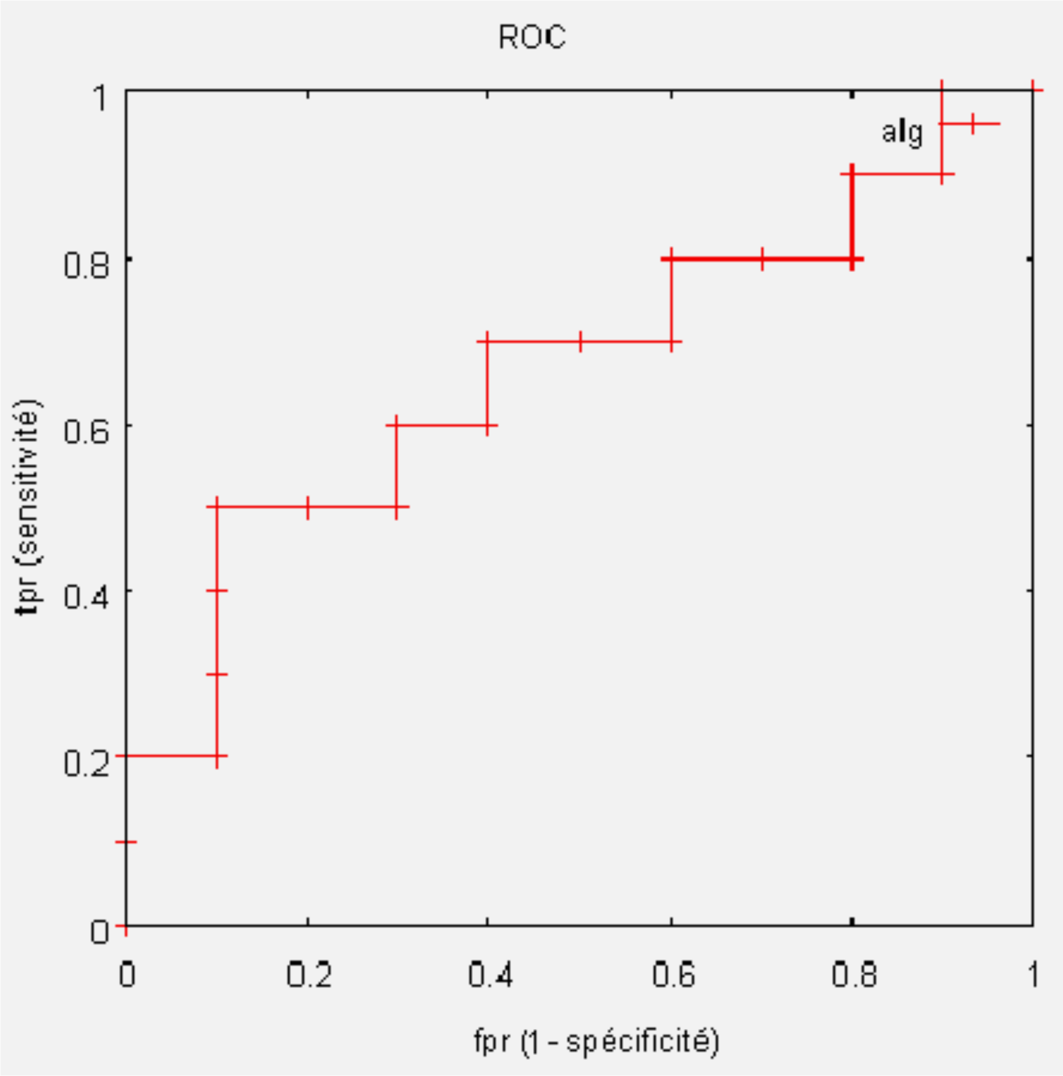
- **R**eceiver **O**perating **C**haracteristic
  - How much model is capable of distinguishing between classes
  - ROC curve (2D plot): fpr (false positive rate) on x-axis, tpr (true positive rate) on y-axis
  - O(0, 1): ideal point idéal (perfect classification)
  - Diagonal (tpr = fpr): random guess
  - Performance: Area under the ROC curve (AUC)

$$tpr = \frac{tp}{tp + fn} = \frac{tp}{pos}$$

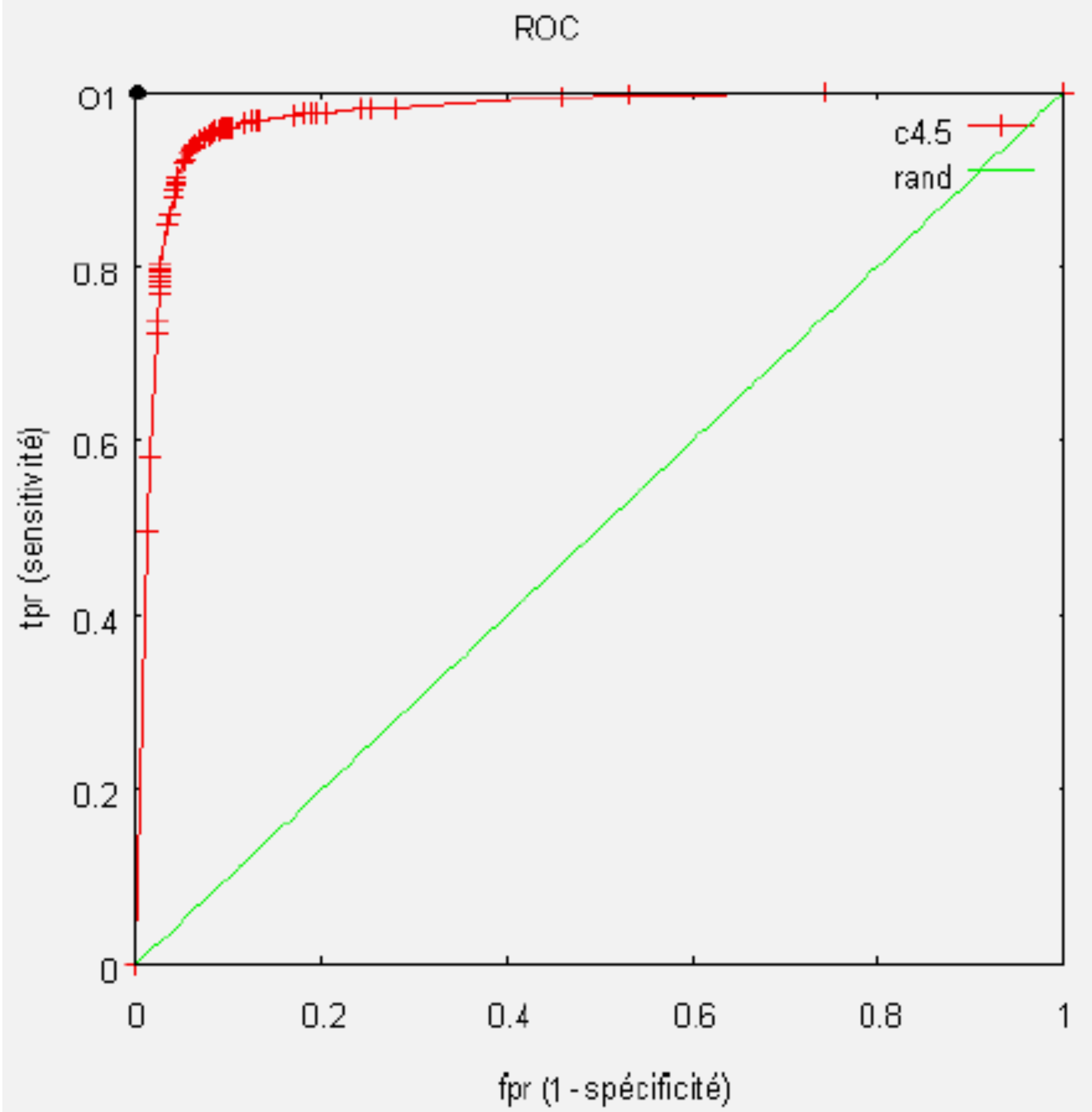
$$fpr = \frac{fp}{fp + tn} = \frac{fp}{neg}$$

ROC

Order	Classes	Discriminant threshold
1	+1	0.9
2	+1	0.8
3	-1	0.7
4	+1	0.6
5	+1	0.55
6	+1	0.54
7	-1	0.53
8	-1	0.55
9	+1	0.51
10	-1	0.505
11	+1	0.4
12	-1	0.39
13	+1	0.38
14	-1	0.37
15	-1	0.36
16	-1	0.35
17	+1	0.34
18	-1	0.33
19	+1	0.3
20	-1	0.1



ROC


ROC
O1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
tpr (sensitivité)
fpr (1 - spécificité)
c4.5
rand

Merci !