University of Science, VNU-HCM

Faculty of Information Technology

Introduction to Data Science Course

Evaluation

Le Ngoc Thanh
lnthanh@fit.hcmus.edu.vn
Department of Computer Science

Ho Chi Minh City

Assessment

- Accuracy: The predicate capability of the classifier
- Effectiveness:
  - Cost to create the model
  - Cost to use the model
- Robustness: ability to solve noise or missing value
- Scale: Efficiency with Big Data
- Understandable
- Other properties: Tree size, number of laws, quality of law...

=>Hàm Loss chỉ dùng để Traning or đánh giá trên tập valid
=> Accuracy chỉ dùng trên tập Testing

Accuracy

- The dataset is divided into two completely independent parts.
  - Training set
  - Test set
- Measures to evaluate accuracy: confusion matrix, fault rate,...
- The method of estimating the accuracy of the classifier:
  - Holdout method, Random Subsampling
  - Cross-validation
  - Bootstrap

Some concepts (1/2)

- Let’s:
  - Positive tuples are samples belonging to a major class that is concerned
  - Negative tuples are the models that belong to the remaining classes
- P is the number of positive sample, N is number of negative samples in the test set.

=> P là mẫu dương, là mẫu mà chúng ta quan tâm
=> Những mẫu còn lại là mẫu âm

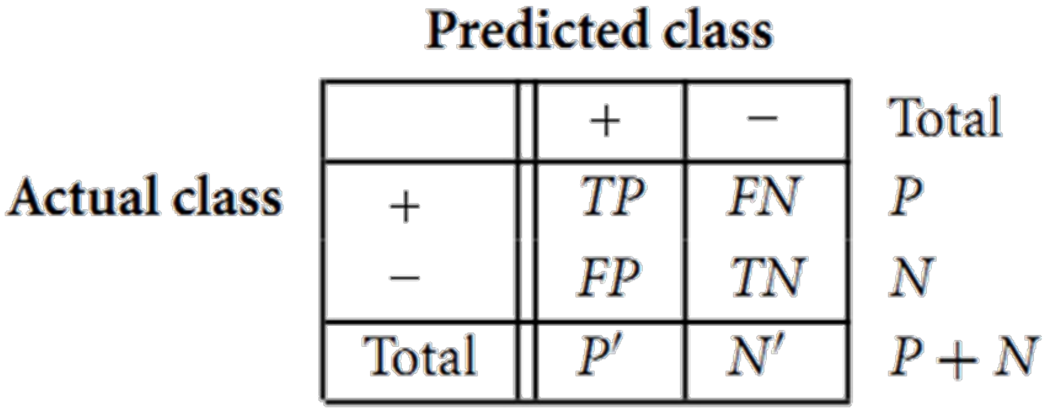
- TP (True Positives): number of positive samples are classified correctly
- TN (True Negatives): number of negative samples are classified correctly

Some concepts (2/2)

- FP (False Positives): number of negative samples are classified incorrectly to positive samples
- FN (False Negatives): number of positive samples are classified incorrectly to negative samples.

=> FP: bản chất là mẫu âm nhưng máy tính nhận nhầm thành mẫu dương => dương sai

=> FN: tương tự => âm sai

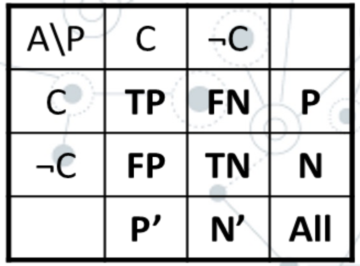


Predicted class

		+	−	Total
Actual class	+	TP	FN	P
	−	FP	TN	N
	Total	P'	N'	$P+N$

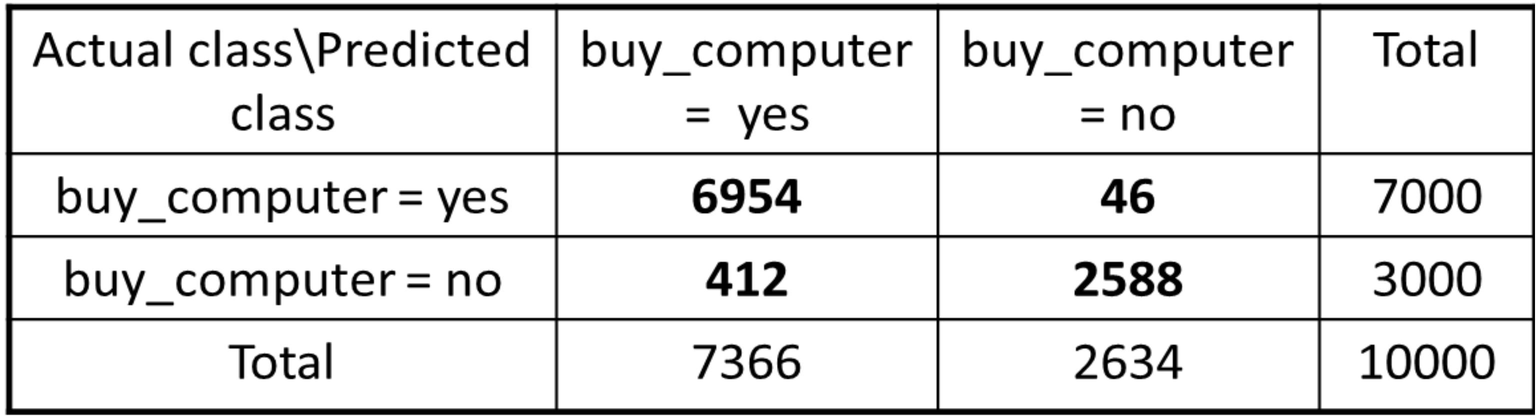
Confusion Matrix

Confusion Matrix



A\P	C	¬C	
C	**TP**	**FN**	**P**
¬C	**FP**	**TN**	**N**
	P’	**N’**	**All**

- Data in computer store, positive samples P are samples with buys_computer = yes



Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	**6954**	**46**	7000
buy_computer = no	**412**	**2588**	3000
Total	7366	2634	10000

- Determine TP, TN, FP, FN?
- Ideally, the sub diagonal should be 0 or approximately 0

=> Đánh giá trên toàn tập dự liệu là Accuracy
=> Đánh giá trên mẫu thì là Predicted

Accuracy

- Accuracy: sample rate in test set is classified correctly.

$$accuracy = \frac{TP + TN}{P + N}$$

- Example:
  - accuracy = (6954 + 2588)/ 10000 = 0.95

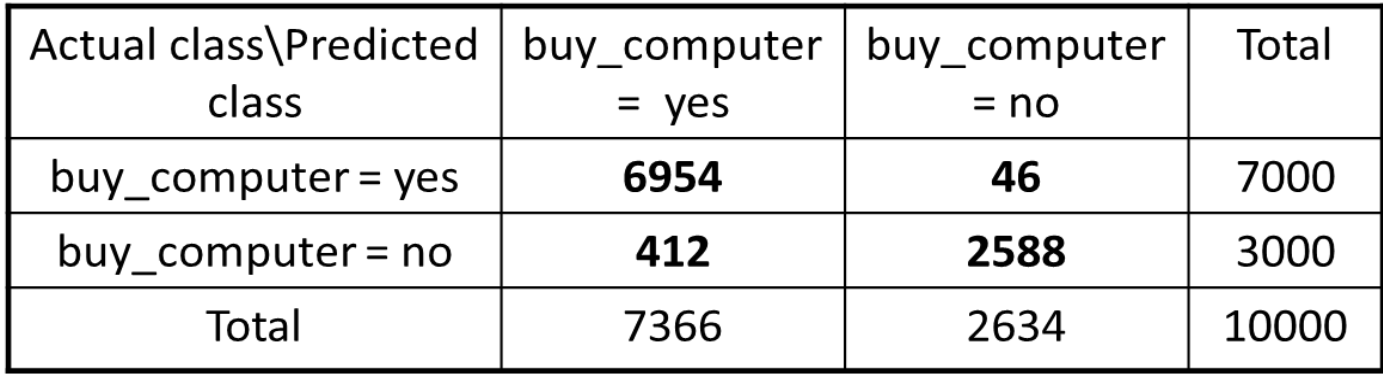
Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	**6954**	**46**	7000
buy_computer = no	**412**	**2588**	3000
Total	7366	2634	10000

Error Rate

- Error Rate: The sample rate was incorrectly classified in the test set (= 1 - accuracy)

$$error\ rate = \frac{FP + FN}{P + N}$$

- Example:
  - error rate = (412 + 46)/ 10000 = 0.05



Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	**6954**	**46**	7000
buy_computer = no	**412**	**2588**	3000
Total	7366	2634	10000

Class imbalance (1/2)

- Classes of interest can be rare compared to other classes
- Example:
  - In the phishing detection application, the class of interest is "fraud" but occurs much less than those of the class "Nonfraudulant".
  - In diagnosis, the class of interest is “cancer”, the sample rate is labeled "yes" much lower than the label "no".

Class imbalance (2/2)

- Classifier is correct in negative samples but completely incorrect in positive samples
- Example:
  - A classifier of accuracy 99% shows a very high probability of prediction. However, if the wrong 1% belongs to the positive sample, 99% becomes pointless

→ Resolution by measure sensitivity and specificity

Sensitivity và Specificity

- Sensitivity: correct positive sample recognition ratio

=> Độ nhạy: tôi rất nhạy với mẫu dương => đc qtam hơn

$$sensitivity = \frac{TP}{P}$$

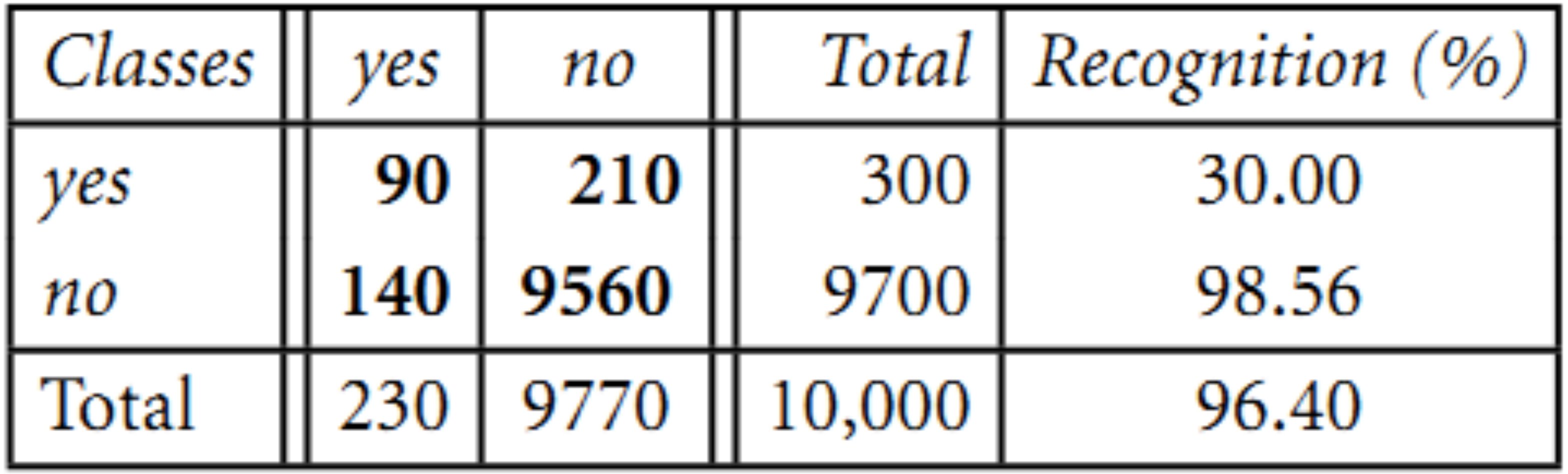
- Specificity: correct negative sample recognition ratio

=> Độ nhạy với mẫu âm

$$specificty = \frac{TN}{N}$$

Sensitivity vs. Specificity

- The classifier has a high accuracy of 96.40%.
- However, the ability to identify positive samples is quite low because of low sensitivity.



Classes	*yes*	*no*	*Total*	*Recognition (%)*
yes	**90**	**210**	300	30.00
no	**140**	**9560**	9700	98.56
Total	230	9770	10,000	96.40

Precision vs. Recall

- Precision: is the proportion of the class that assigns a label to positive is actually positive. => số mẫu dương đúng trong tổng số những mẫu máy tính cho là dương

=> Mẹo để tăng: trả về 1 mẫu thôi (đc ăn cả ngã về 0)

$$precision = \frac{TP}{TP + FP}$$

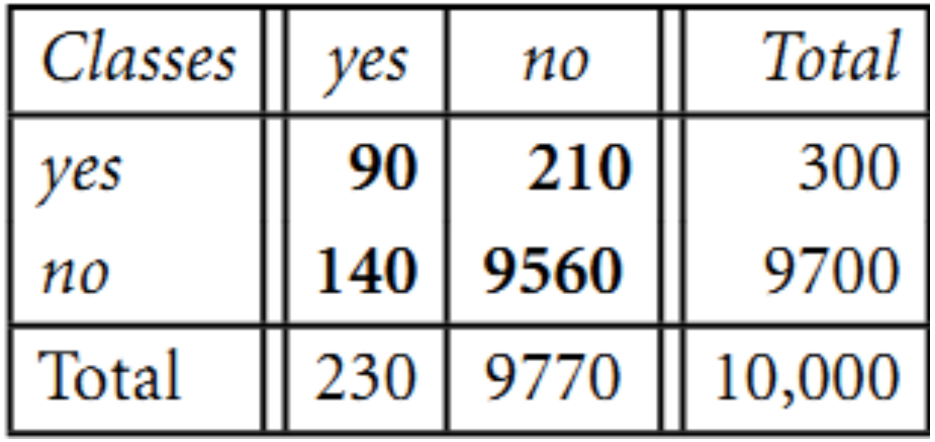
- Recall: is the positive sample rate assigned by the classifier. => số lượng mẫu dương đúng trên toàn bộ mẫu dương => độ phủ

=> Mẹo tăng: trả về hết

$$recall = \frac{TP}{TP + FN} = \frac{TP}{P}$$

Precision and Recall

- Precision(yes) = 90/230 = 39.13%
- Recall(yes) = 90/300 = 30.00%



Classes	*yes*	*no*	*Total*
yes	**90**	**210**	300
no	**140**	**9560**	9700
Total	230	9770	10,000

Precision and Recall

- Highest precision is 1.0:
  - Showing each sample that the marking class belongs to the positive is actually positive.
  - Unable to show the number of positive samples is classified incorrectly
- Highest recall is 1.0:
  - Showing all positive samples is labeled properly.
  - Unable to present how many other samples are mislabeled in the positive.

=> Trung bình nghich đảo của precision và recall để cân bằng (vì chỉ sắp xếp đc trong KG 1 chiều, có tính thứ tự)

F-Score

- F-score: A combination of precision and recall

2 là độ chính xác và độ phủ là như nhau

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

- F-score is harmonic mean between precision and recall
- Equally weighted between precision and recall (β=1)
- If you want to one over another, you can set β=2, β=0.5

$$F_\beta = \frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$$

=> đánh trọng, B=2 thì tăng độ quan trọng của precision hơn

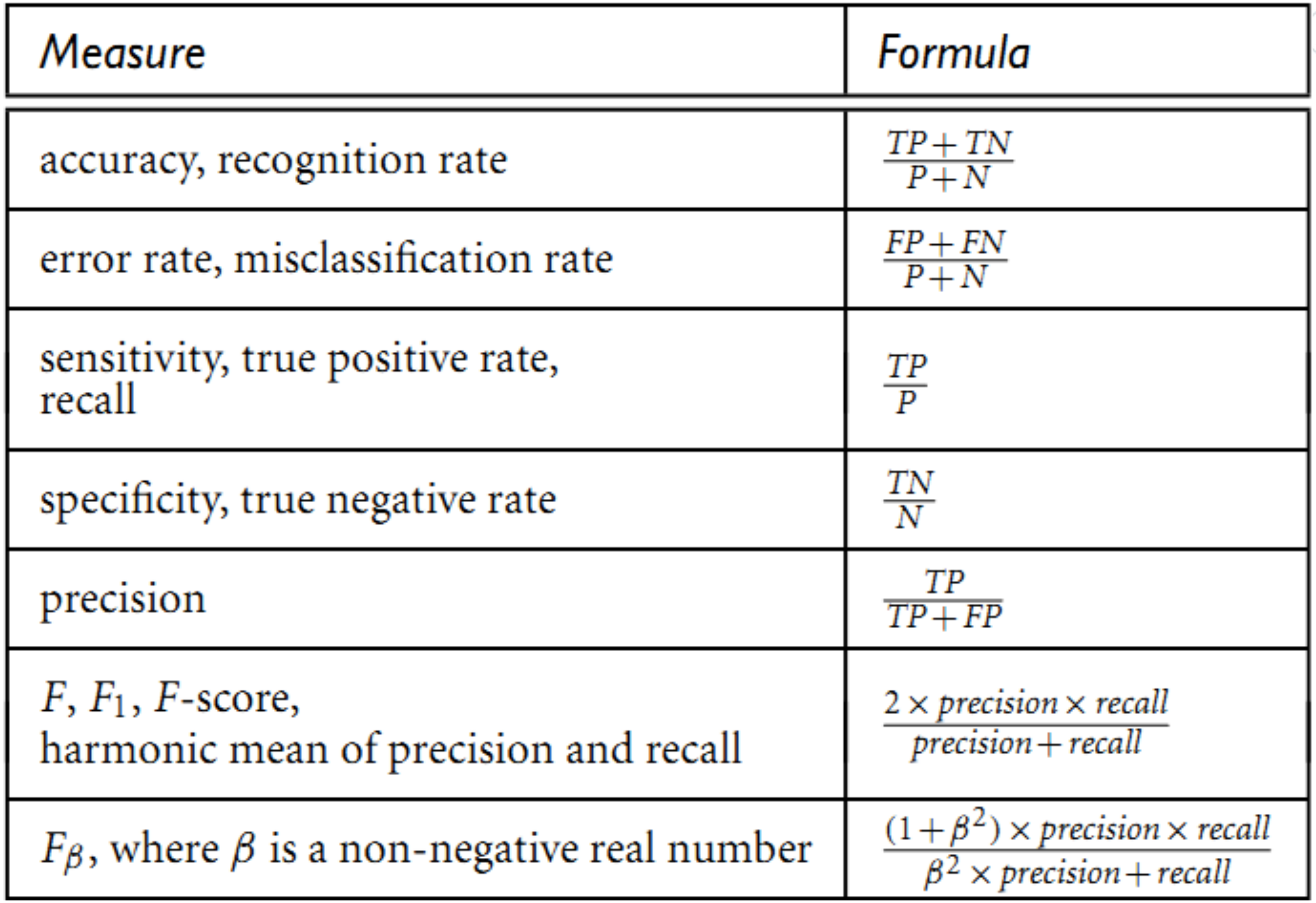
Example

classes	buy_computer = yes	buy_computer = no	total	recognition(%)
buy_computer = yes	6954	46	7000	99.34
buy_computer = no	412	2588	3000	86.27
total	7366	2634	10000	95.42

F-measure(B-Yes)= 96.81%

Summary of the measurements

A\P	C	¬C	
C	**TP**	**FN**	**P**
¬C	**FP**	**TN**	**N**
	P'	**N'**	**All**



Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$
F, F_1, F-score, harmonic mean of precision and recall	$\frac{2 \times precision \times recall}{precision + recall}$
F_β, where β is a non-negative real number	$\frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$

Exercise 3

=> accuracy: 0,84 => sensitive: 0.75 => precision: 0.6875 => F-score: 0.72

=> error rate: 0.16 => specific: 0.88 => recall: 0.75

- Data relating to the customer classification is deceptive or non-deceptive by a bank before lending:

		Predict		
	Class	**Deceptive**	**Non-deceptive**	**Total**
Actual	**Deceptive**	44	15	59
	Non-deceptive	20	146	166
	Tổng	64	161	225

- Suppose the class of interest is deceptive, create confusion matrix
- Calculating the measurements accuracy, error rate, sensitivity, specificity, precision, F-Score

Estimation methods

Reliability when estimating

- Whether the figures are calculated from the measurements that are reliable?
  - Depends on the type of data
  - Depend on how the data is collected
  - Depends on how to divide data into training and test episodes.
  - …

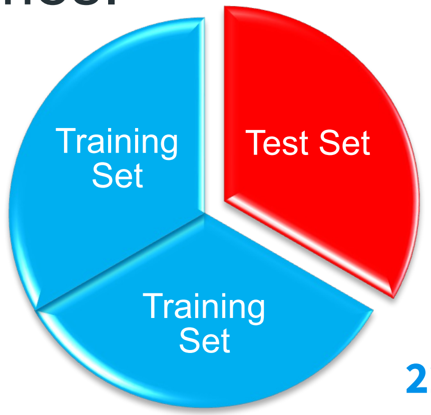
→ Method is required to reliably estimate the accuracy

Holdout Method

- Data is randomly divided into 2 independent sections
  - Training set up 2/3 to draw the model
  - Test set accounted for 1/3 to estimate accuracy

→ Samples may not represent all data, missing class in the experiment set

- Random sampling: is the variant of holdout
  - Repeat holdout k times, the accuracy is the average of all times.

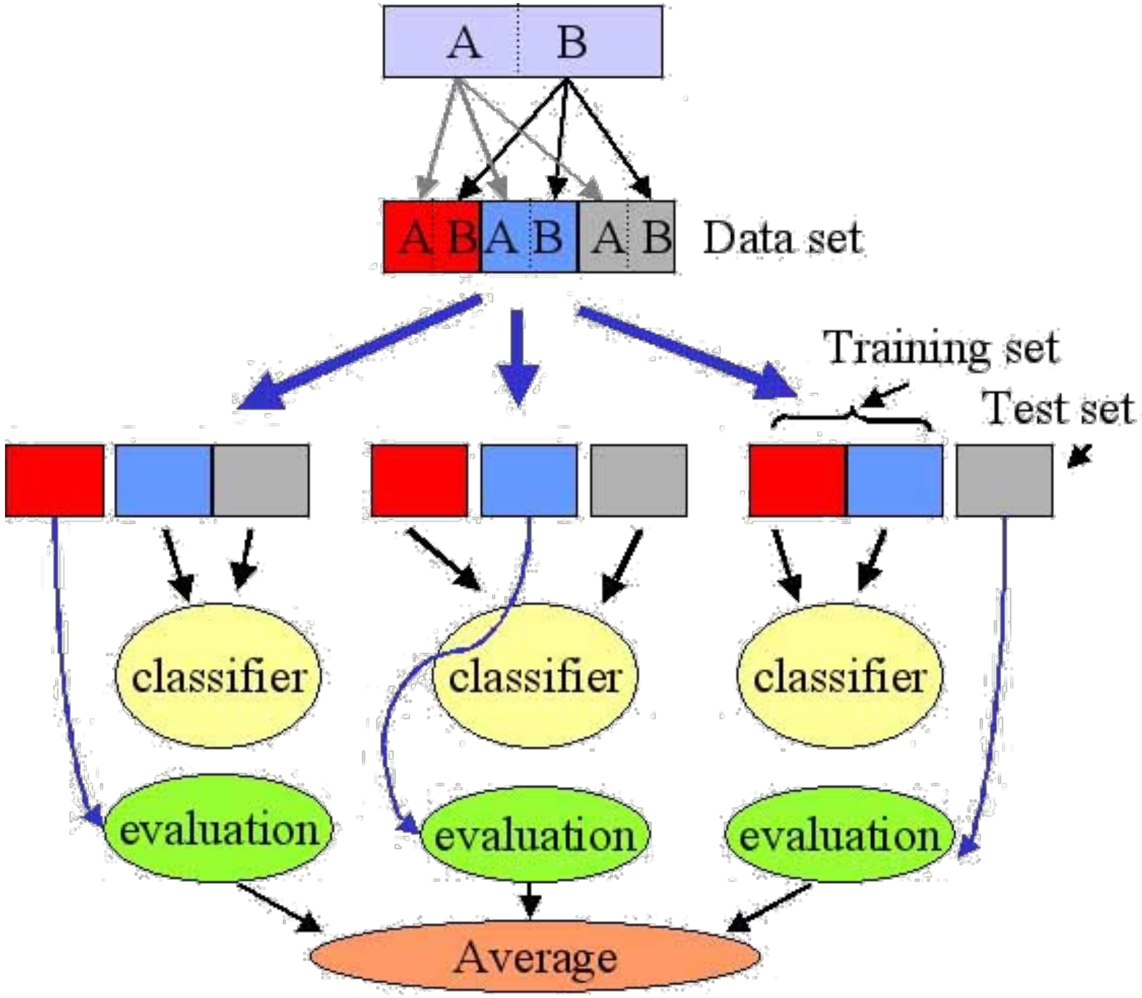


K-fold Cross-Validation

- Randomly divide the data into K-independent and roughly equal size. D={D1, D2, …, Dk}
  - Perform k evaluation.
  - For the time i, the Di episode was used as a test, the rest of the training
  - K = 10 commonly used
- Leave-one-out: is k fold with K is the number of samples (only applies when the data size is small)
- Stratified cross-validation: Distributing the classes of samples in each fold is roughly the same as the original data

K-fold Cross-Validation

=> Đánh giá k phần => bất kù DL nào cũng đc train và test
=> Cũng có những cái ko áp dụng đc k-fold: đồ thị



Bootstrap

=> Dành cho DL ko quá nhiều , quá ít
=> Dùng phương pháp bỏ vào, lấy ra
=> Hiếm gặp

- Usually apply to small datasets
- Each time a sample is selected, it is likely to be picked again and added to the training set
- There are a few bootstrap methods, which are common **.632 Bootstrap**
  - The d-size dataset will be sampled Bootstrap d times. So training set has d samples. Samples that do not include the training will be used to test. About 63.2% of data fall into training assignments and 36.8% for test episodes (because according to probability (1-1/d) d ≈ e^{-1} = 0.368)
  - Repeats the sampling k times and the accuracy:

$$Acc(M) = \frac{1}{k}\sum_{i=1}^{k}(0.632\times Acc(M_i)_{test\_set} + 0.368\times Acc(M_i)_{train\_set})$$

+ Thì 90'
+ Đc mang tài liệu giấy, điền tên vào
+ CÓ 2 dạng câu hỏi:
+ dạng hiểu, lý thuyết
+ dạng bài tập : thống kê, data visualization : cho DL thì visual như nào, cho visual hỏi ổn ko . Có 1 số ô thiếu , nhiễu thì điền như nào (trước khi làm phải cm or giả định cột đó theo phân phối chuẩn để dùng 1.5IQR??)
+ Cho đg xog hỏi độ lỗi
+ Ko hỏi mạng nơ ron , chỉ hỏi phần model tuyến tính
+ Đề thi tiếng anh nhưng trả lời 1 ngôn ngữ (Anh or Việt)

The End