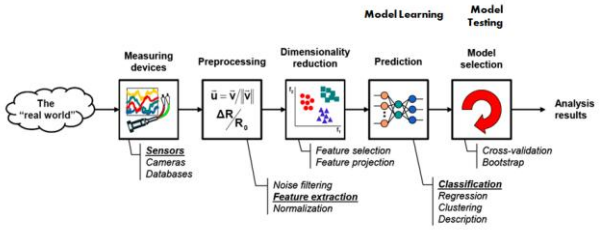
UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

NHẬN DẠNG
PATTERN RECOGNITION

ĐÁNH GIÁ MÔ HÌNH- EVALUATION

1

QUÁ TRÌNH HOC MÁY CƠ BẢN



2

Nội dung

1. Tầm quan trọng của Evaluation?
2. Các tiêu chí đánh giá.
3. Các phương pháp đánh giá.
4. Một số độ đo tương ứng với bài toán.

3

Tại sao phải đánh giá ?

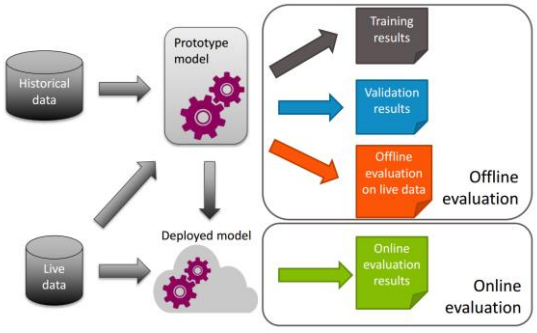
1. Biết được **khi nào huấn luyện mô hình thành công** ?
2. Biết được **mức độ thành công** của mô hình
3. Biết được **thời điểm dừng quá trình huấn luyện**
4. Biết được **khi nào cần cập nhật mô hình** ?

4

Một số câu hỏi căn bản khi evaluation

1. Đánh giá **khi nào** ?
2. **Các tiêu chí** đánh giá là gì ?
3. Dữ liệu – **Phương pháp đánh giá** ?
4. **Độ đo nào** được sử dụng ?

When to evaluation

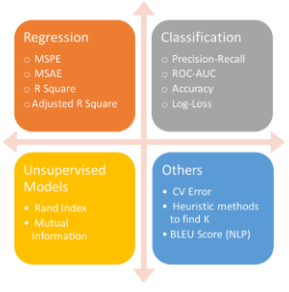


2. Các tiêu chí đánh giá

1. Tính chính xác (Accuracy)
2. Tính hiệu quả (Efficiency)
3. Khả năng xử lý nhiễu (Robustness).
4. Khả năng mở rộng (Scalability).
5. Khả năng diễn giải(Interpretability)
6. Mức độ phức tap (complexity)

2. 1 Accuracy – chính xác

➔ Tùy vào **bài toán, dữ liệu** sẽ có độ đo tương ứng.

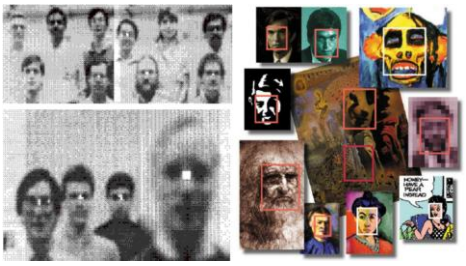


2.2 Efficiency – hiệu quả

➔ Chi phí về **thời gian và tài nguyên** (bộ nhớ) cần thiết cho việc huấn luyện và kiểm thử hệ thống.

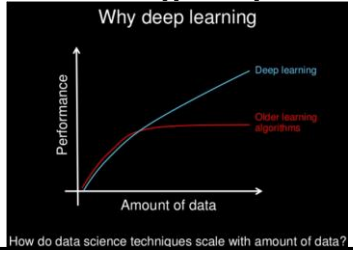
2.3 Robustnesss – xử lý nhiễu

➔ Khả năng xử lý của hệ thống đối với các ví dụ **nhiễu (lỗi)** hoặc **thiếu giá trị**.



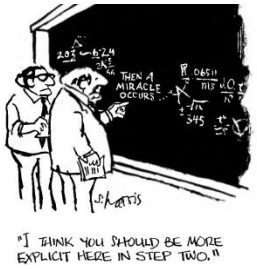
2.4 Scalability – mở rộng

➔ **Hiệu năng** của hệ thống (ví dụ: tốc độ học, độ chính xác) **thay đổi** như thế nào đối với **kích thước của tập dữ liệu**



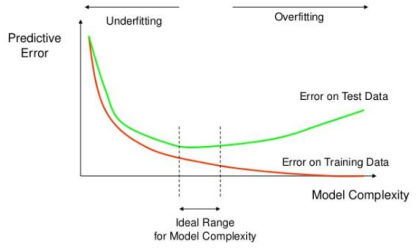
2.5 Interpretability – diễn giải

➔ **Mức độ dễ hiểu** (đối với người sử dung) của các **kết quả** và **hoạt động** của hệ thống.

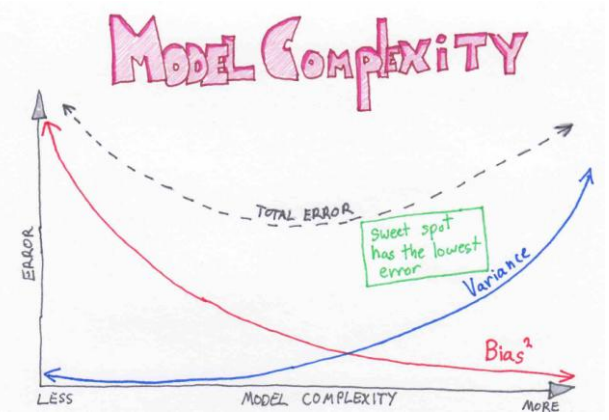


2.6 Complexity – mức độ phức tạp

➔ **Mức độ phức tạp** của hệ thống (hàm hyperthesis mục tiêu) học được.



13



14

3. Các phương pháp đánh giá

1. Hold-out / Repeated hold-out
2. Stratified sampling
3. Cross- validation
   1. K-Fold
   2. Leave one out
4. Bootrap sampling.

15

3. 1 Hold-Out (Splitting)

➔ **Toàn bộ dữ liệu D được chia thành 2 tập con không giao nhau**

D_Train: dùng để huấn luyện hệ thống

D_Test: để đánh giá hiệu năng hệ thống sau khi học

Một số yêu cầu:

- Dữ liệu đã sử dụng ở D_Train thì không được ở trong D_Test
- Các tỉ lệ thường sử dụng D_Tran =2/3 D, D_Train 80% D_Test 20% , , D_Train 70% D_Test 30%

➔ **Thường phù hợp cho tập D có kích thước lớn.**

16

3.1 Repeat Hold-Out

➔ **Áp dụng hold-out nhiều lần**

- Trong **mỗi lần lặp một tỉ lệ nhất định của D được lựa chọn ngẫu nhiên để tạo tập dữ liệu D'**
- Các giá trị lỗi (hoặc các giá trị đối với các tiêu chí đánh giá khác) được ghi nhận trong các bước lặp này **được lấy trung bình cộng để xác định lỗi tổng thể**

Hạn chế:

➔ Mỗi bước lặp **sử dụng một tập kiểm thử khác nhau**

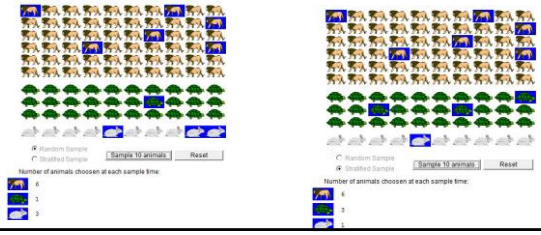
➔ Có trường hợp **một ví dụ trùng lặp trong các kiểm thử**.

17

3. 2 Statified sampling – lấy mẫu phân tầng

- Tập **dataset có kích thước nhỏ**.
- Dữ liệu không cân xứng (**unbalanced dataset**)

➔ **Mục tiêu**: Phân lớp (class distribution) trong tập huấn luyện và kiểm thử phải xấp xỉ như trong toàn tập Dataset.



3. 2 Statified sampling – lấy mẫu phân tầng

➔ Lấy mẫu phân tầng có tác dụng:

- Làm cân xứng (về phân bố lớp)
- Đảm bảo tỉ lệ phân lớp (tỉ lệ các ví dụ giữa các lớp) trong tập huấn luyện và tập kiểm thử là xấp xỉ nhau.

➔ **Phương pháp này không áp dụng cho bài toán hồi quy.**

19

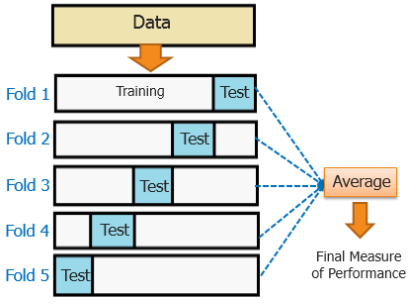
3.3 Cross-Validation

Cross validation - Phương pháp phù hợp khi tập dữ liệu **vừa và nhỏ**

- **K-Fold Cross Validation**
- **Stratified K-fold Cross Validation**
- **Leave One Out Cross Validation**

20

3.3 Cross-Validation

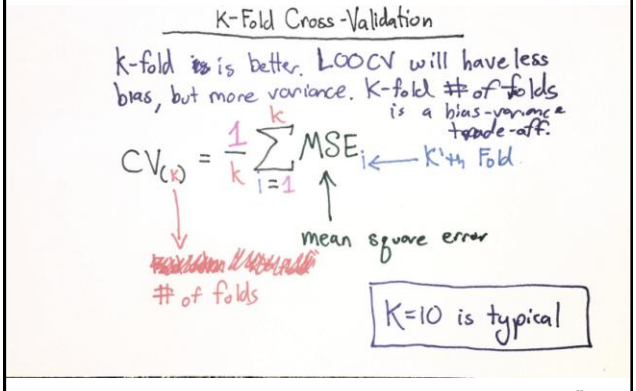


21

3.3 Cross-Validation

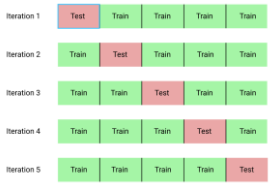
- Tập dữ liệu D được **chia thành k tập con không giao nhau** (gọi là fold) **có kích thước xấp xỉ nhau**
- Mỗi lần (trong k lần lặp), **một tập con được sử dụng làm tập kiểm thử**, và **k-1 tập con còn lại được dùng làm tập huấn luyện**.
- K giá trị lỗi (mỗi giá trị tương ứng với fold) được **tính trung bình cộng để thu được độ lỗi tổng thể**.
- Các lựa chọn cho K thường là **5 hoặc 10**

22

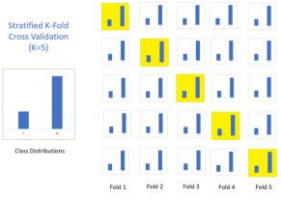


23

3.3 Cross-Validation



K-Fold Cross Validation



Stratified K-Fold Cross Validation

Các lớp trong tập có **phân bố lớp xấp xỉ bằng nhau**

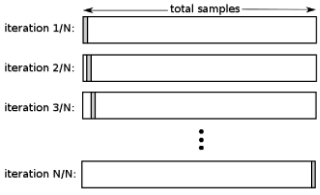
24

3.3 Cross-Validation

Leave-one-out cross-validation – Phù hợp cho tập D **(rất) nhỏ**

Là một dạng của cross-validation trong đó:

- Số lượng các folds **bằng kích thước của tập dữ liệu**.
- **Mỗi fold chỉ bao gồm một ví dụ**.



25

3.4 Bootstrap sampling

- **Cross-validation –** lấy mẫu không lặp lại (sampling without replacement)
- **Bootstrap sampling –** lấy mẫu có lặp lại (sampling with replacement)

26

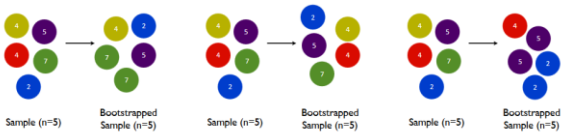
3.4 Bootstrap sampling

- **Cross-validation –** lấy mẫu không lặp lại (sampling without replacement)
- **Bootstrap sampling –** lấy mẫu có lặp lại (sampling with replacement)

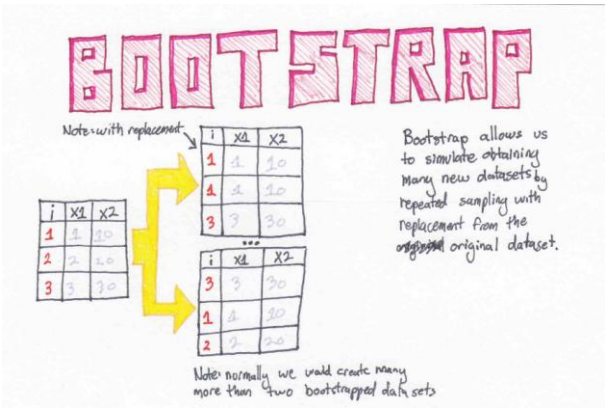
27

3.4 Bootstrap sampling

- **Phù hợp với tập dữ liệu có kích thước (rất nhỏ)**



28



29

4. Một số độ đo

1. Accuracy/ Error
2. Precision/Recall
3. F-Score
4. AP/MAP

30

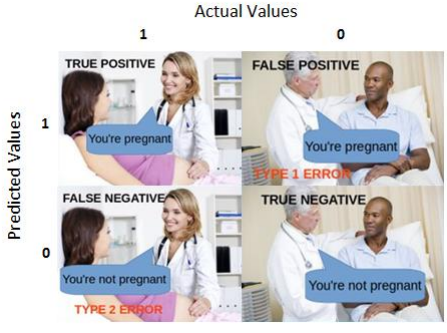
Confusion matrix (ma trận nhầm lẫn)

- TP_i (true positive): Số lượng các ví dụ thuộc lớp c_i được phân loại chính xác vào lớp c_i
- FP_i (false positive): Số lượng các ví dụ không thuộc lớp c_i bị phân loại nhầm vào lớp c_i
- TN_i (true negative): Số lượng các ví dụ không thuộc lớp c_i được phân loại (chính xác)
- FN_i (false negative): Số lượng các ví dụ thuộc lớp c_i bị phân loại nhầm (vào các lớp khác c_i)

Lớp c_i		Được phân lớp bởi hệ thống	
		Thuộc	Ko thuộc
Phân lớp thực sự (đúng)	Thuộc	TP_i	FN_i
	Ko thuộc	FP_i	TN_i

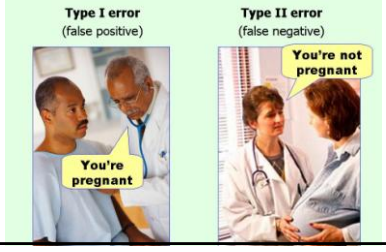
31

Confusion matrix (ma trận nhầm lẫn)



32

NHẦM BỎ SÓT



33

Error Types

Null Hypothesis		Decision Based on test: Accept	Decision Based on test: Reject
In Actuality	TRUE	✓	✗ Type I error
In Actuality	FALSE	✗ Type II error	✓

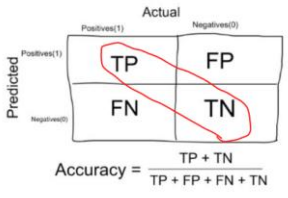
Type 1: **Loại bỏ** ví dụ mà đúng ra **không nên loại bỏ**

Type 2: **Chấp nhận** ví dụ mà đúng ra **không nên chấp nhận**

34

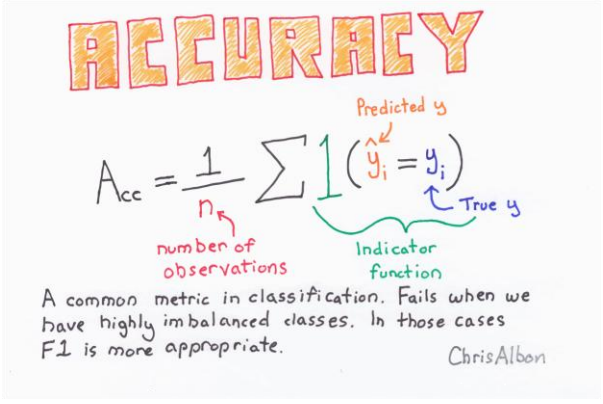
4. 1 Accuracy – độ chính xác

➔ Mức độ dự đoán (phân lớp) **chính xác** của hệ thống (đã được huấn luyện) đối với ví dụ kiểm chứng (test data).



Error = 1 - accuracy

35



36

4. 1 Accuracy – độ chính xác

- Là độ đo tính toán đơn giản nhất.
- Phù hợp cho các bài toán bộ dữ liệu cân bằng trong đó tỉ lệ FP (*nhầm*) và FN (*bỏ sót*) cân bằng nhau.

Hạn chế:

- Chỉ thể hiện độ chính xác không thể hiện loại lỗi trong mô hình.

37

4. 2 Precision/Recall

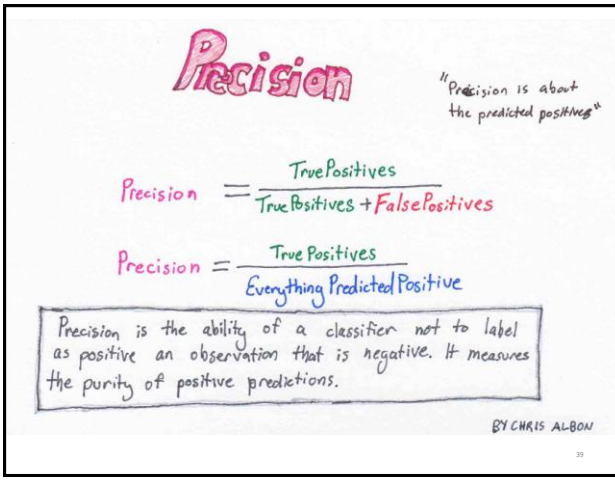
Thực tế (Actual)

(Predicted)	1	0
1	True Positive	False Positive
0	False Negative	True Negative

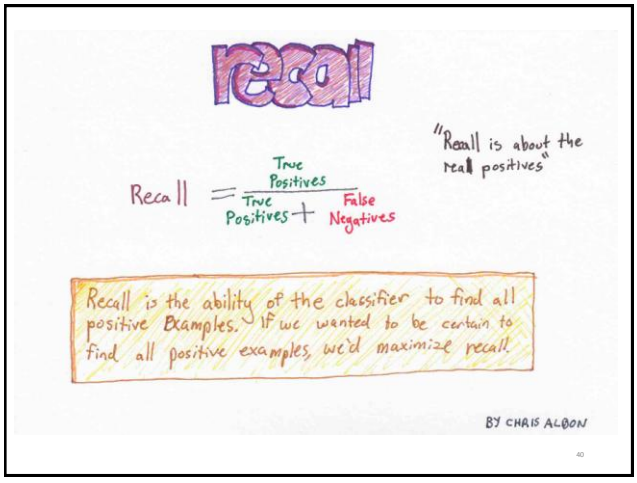
$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} = \frac{\textit{True Positive}}{\textit{Total Predicted Positive}}$$

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} = \frac{\textit{True Positive}}{\textit{Total Actual Positive}}$$

38



39



40

4. 2 Precision/Recall

- **Precision** đối với lớp c_i
  - → Tổng số các ví dụ thuộc lớp c_i được phân loại chính xác chia cho tổng số các ví dụ được phân loại vào lớp c_i

$$Precision(c_i) = \frac{TP_i}{TP_i + FP_i}$$

- **Recall** đối với lớp c_i
  - → Tổng số các ví dụ thuộc lớp c_i được phân loại chính xác chia cho tổng số các ví dụ thuộc lớp c_i

$$Recall(c_i) = \frac{TP_i}{TP_i + FN_i}$$

41

4. 2 Precision/Recall

- Làm thế nào để tính toán được giá trị Precision và Recall (một cách tổng thể) cho toàn bộ các lớp $C=\{c_i\}$?
- Trung bình vi mô (Micro-averaging)

$$Precision = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FP_i)} \qquad Recall = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FN_i)}$$

- Trung bình vĩ mô (Macro-averaging)

$$Precision = \frac{\sum_{i=1}^{|C|} Precision(c_i)}{|C|} \qquad Recall = \frac{\sum_{i=1}^{|C|} Recall(c_i)}{|C|}$$

42

4. 2 Precision/Recall

		Actual	
		Spam	Not Spam
Predict	Spam	8	32
	Not Spam	2	8

- Prec = 8/(8+32) = 20%
- Rec = 8/10 = 80%

➔Tỷ lệ xác suất bộ lọc chính xác khi **xác định 1 mail là thư rác** là 20%.

➔ Tỷ lệ xác suất **một thư rác bị bộ lọc phát hiện** là 80%.

43

4. 2 Precision/Recall

- Một mô hình tốt mong muốn khi Precision và Recall **đều cao**.
- Chọn Precision hay Recall tùy thuộc vào bài toán.

Hạn chế:

- Precision và Recall **thường mất cân bằng nhau**.

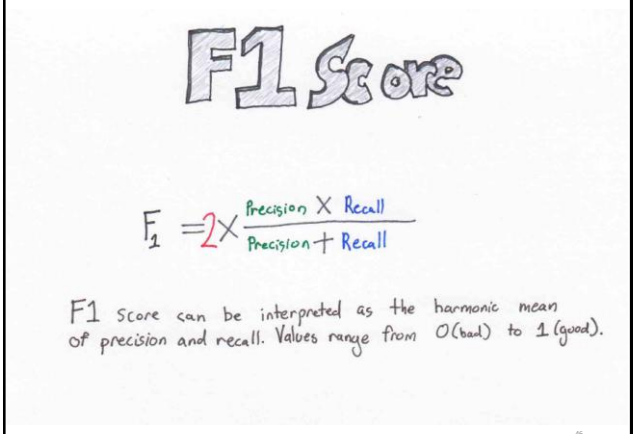
44

4. 3 F- Score

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

- Khi β>1, recall được coi trọng hơn precision
- Khi β<1, precision được coi trọng hơn.
- Khi β=1, precision và recall coi trọng như nhau.
- β thường được sử dụng là β=2 và β=0.5

45



46

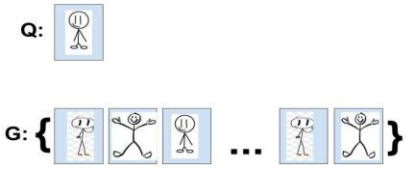
4. 3 F1 -Score

- F là một **trung bình điều hòa (harmonic mean)** của các tiêu chí Precision à Recall. Nó có xu hướng **lấy giá trị gần với giá trị nào nhỏ hơn giữa 2 tiêu chí** này.
- F1 **có giá trị lớn nếu cả 2 giá trị Precision và Recall đều lớn ➔ F1 càng cao độ phân lớp càng tốt.**

47

4. 4 Average Precision

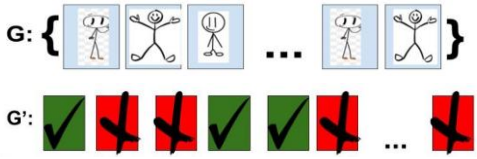
Information Retrieval



- Q to be the user query
- G to be a set of labeled data in the database

48

4. 4 Average Precision



- Ground truth positives (GTP) – number of **True of query Q**.
- d(i,j) to be a score function to show how similar object i is to j
- G' which an ordered set of G according to score function d(,)

49

4. 4 Average Precision

$$\text{AP@k} = \frac{1}{\text{GTP}} \sum_{i=1}^{k} \frac{\text{TP seen}}{i}$$

- K to be the index of G'
- GTP refers to the total number of ground truth positives for the query
- TP seen refers to the number of true positives seen till k

50

4. 4 Average Precision

Overall AP = ⅓ (1/1 + 0/2 + 0/3 + 2/4 + ⅖ + 0 … + 0) = 0.7

Calculation of a AP for a given query, Q, with a GTP=3

51

4. 4 Average Precision

Overall AP = ⅓ (1/1 + 2/2 + 3/3 + 0/4 + 0/5 + 0 … + 0) = 1.0

Calculation of a pefect AP for a given query, Q, with a GTP=3

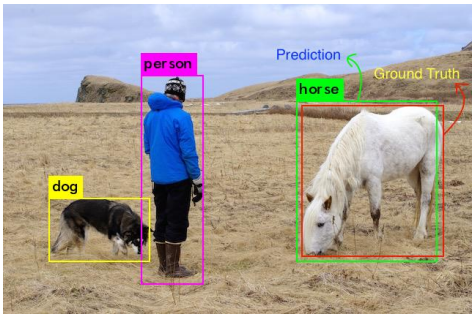
52

4. 4 Mean Average Precision - mAP

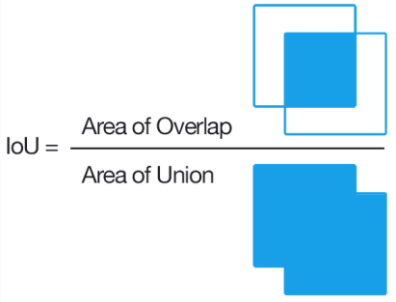
$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP(q)}}{Q}$$

- Q is the number of queries
- **AveP(q)** is the average precision (AP) for a given query, q

IoU - Intersection over union



IoU - Intersection over union



IoU - Intersection over union

IoU={0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1}

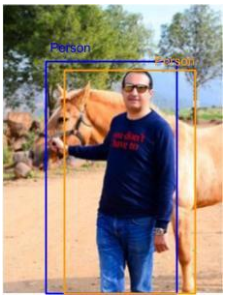
IoU - Intersection over union

Passcal VOC Challenge : based on 50% IOU

COCO Challenge: ranging from 5% to 95%

57

IoU - Intersection over union



True Positive

Example (IoU > 0.5)

58

IoU - Intersection over union

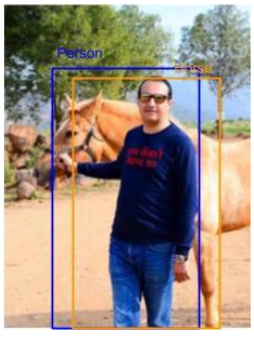
False Positive

59

IoU - Intersection over union

False Negative

IoU > 0.5 but has the wrong classification



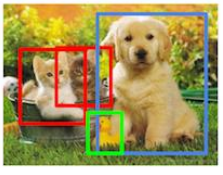
60

Detection

Classification

CAT

Object Detection



CAT, DOG, DUCK

61

Detection

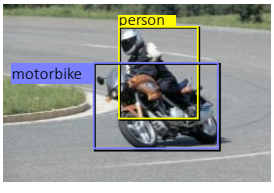
Class	X coordinate	Y coordinate	Box Width	Box Height
Dog	100	600	150	100
Horse	700	300	200	250
Person	400	400	100	500

62

Problem formulation

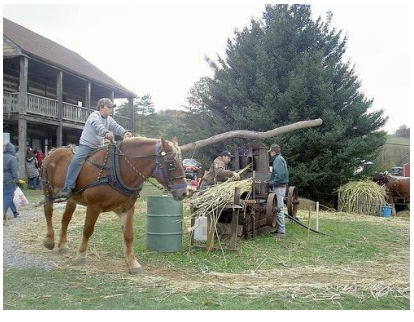
{ airplane, bird, motorbike, person, sofa }

Input



Desired output

Evaluating a detector

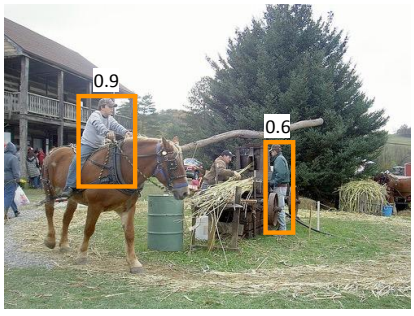


Test image (previously unseen)

First detection ...

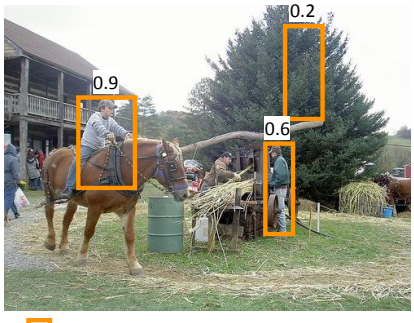
‘person’ detector predictions

Second detection ...



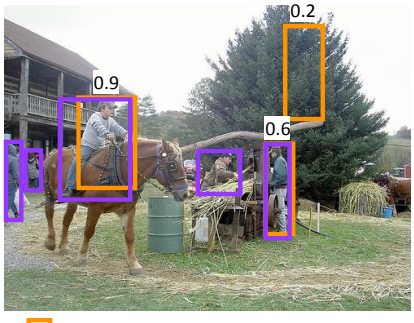
‘person’ detector predictions

Third detection ...



‘person’ detector predictions

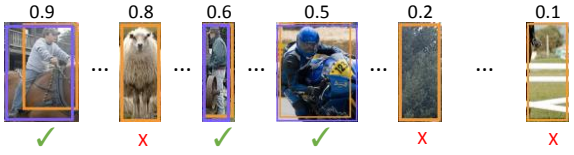
Compare to ground truth



‘person’ detector predictions

ground truth ‘person’ boxes

Sort by confidence



true positive
(high overlap)

false positive
(no overlap, low overlap, or duplicate)

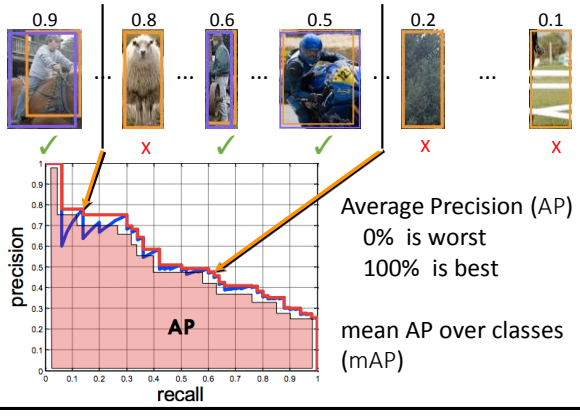
Evaluation metric

$$precision@t = \frac{\#true\ positives@t}{\#true\ positives@t + \#false\ positives@t}$$

$$\frac{\checkmark}{\checkmark + x}$$

$$recall@t = \frac{\#true\ positives@t}{\#ground\ truth\ objects}$$

Evaluation metric



Average Precision (AP)
0% is worst
100% is best

mean AP over classes (mAP)

COO- Dataset

```
Average Precision (AP):
   AP                  % AP at IoU=.50:.05:.95 (primary challenge metric)
   AP^IoU=.50          % AP at IoU=.50 (PASCAL VOC metric)
   AP^IoU=.75          % AP at IoU=.75 (strict metric)
AP Across Scales:
   AP^small            % AP for small objects: area < 32^2
   AP^medium           % AP for medium objects: 32^2 < area < 96^2
   AP^large            % AP for large objects: area > 96^2
Average Recall (AR):
   AR^max=1            % AR given 1 detection per image
   AR^max=10           % AR given 10 detections per image
   AR^max=100          % AR given 100 detections per image
AR Across Scales:
   AR^small            % AR for small objects: area < 32^2
   AR^medium           % AR for medium objects: 32^2 < area < 96^2
   AR^large            % AR for large objects: area > 96^2
```

Normalized Confusion matrix (chuẩn hóa ma trận nhầm lẫn)

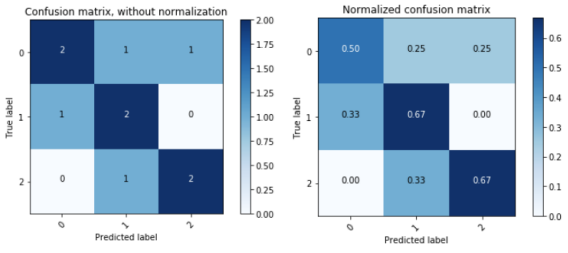
	Predicted as Positive	Predicted as Negative
Actual : Positive	True Positive (TP)	False Negative (FN)
Actual : Negative	False Positive(FP)	True Negative (TN)

	Predicted as Positive	Predicted as Negative
Actual : Positive	TPR=TP/(TP+FN)	FNR = FN/(TP+FN)
Actual : Negative	FPR=FP/(FP+TN)	TN=TN/(FP+TN)

➔ False Positive Rate (FNR) + True Positive Rate =1

73

Normalized Confusion matrix (chuẩn hóa ma trận nhầm lẫn)



74

Normalized Confusion matrix (chuẩn hóa ma trận nhầm lẫn)

- **False Positive Rate** còn được gọi là False Alarm Rate (tỉ lệ **báo động nhầm**).
- **False Negative Rate** còn được gọi là Miss Detection Rate (tỉ lệ **bỏ sót**)

	Predicted as Positive	Predicted as Negative
Actual : Positive	TPR=TP/(TP+FN)	FNR = FN/(TP+FN)
Actual : Negative	FPR=FP/(FP+TN)	TN=TN/(FP+TN)

➔ Trong một số bài toán việc tăng hay giảm **FNR, FPR** phụ thuộc vào ngưỡng nào đó

75

Tài liệu tham khảo

Slide được tham khảo từ:

- http://www.cs.virginia.edu/~hw5x/Course/IR2015/_site/lectures/
- https://nlp.stanford.edu/IR-book/newslides.html
- https://course.ccs.neu.edu/cs6200s14/slides.html

77