

KHOA CNTT & TRUYỀN THÔNG
BM KHOA HỌC MÁY TÍNH

Đánh giá hiệu quả của giải thuật học

PGS. TS. Đỗ Thanh Nghị
TS. Trần Nguyễn Minh Thư
tnmthu@ctu.edu.vn

1

1

Nội dung

- Nghi thức kiểm tra
- Các chỉ số đánh giá

2

2

Nghi thức kiểm tra

- nếu dữ liệu có **1 tập học** và **1 tập kiểm tra** sẵn dùng
 - dùng dữ liệu học để xây dựng mô hình,
 - dùng tập kiểm tra để đánh giá hiệu quả của giải thuật
- nếu dữ liệu **không có 1 tập kiểm tra** sẵn?

3

3

Nghi thức kiểm tra

- nếu dữ liệu không có 1 tập kiểm tra sẵn
 - sử dụng nghi thức **k-fold** :
chia tập dữ liệu thành k phần (fold) bằng nhau, lặp lại k lần, mỗi lần sử dụng k-1 folds để học và 1 fold để kiểm tra, sau đó tính trung bình của k lần kiểm tra
 - nghi thức **hold-out** : lấy ngẫu nhiên 2/3 tập dữ liệu để học và 1/3 tập dữ liệu còn lại dùng cho kiểm tra, có thể lặp lại quá bước này k lần rồi tính giá trị trung bình

4

4

Nghi thức kiểm tra

- nếu dữ liệu có số phần tử lớn hơn 300
sử dụng nghi thức k-fold với $k = 10$
- nếu dữ liệu có số phần tử nhỏ hơn 300
sử dụng nghi thức leave-1-out (k-fold với $k = \text{số phần tử}$)
 \Rightarrow Vd leave 1 out

5

5

➤ Chỉ số đánh giá

6

6

Confusion matrix (C) cho k lớp

dự đoán =>	1	...	k
1			
...			
k			

- $C[i, j]$: số phần tử lớp **i (dòng)** được giải thuật dự đoán là lớp **j (cột)**
- $C[i, i]$: số phần tử phân lớp đúng
- Độ chính xác lớp **i**: $C[i, i] / C[i,]$
- Độ chính xác tổng thể: $\sum C[i, i] / C$

7

7

Confusion matrix (C) cho k lớp

dự đoán =>	Setosa	vesicolor	virginica
Setosa	15	0	0
vesicolor	0	16	2
virginica	0	3	14

- ❑ Độ chính xác lớp **i**: $C[i, i] / C[i,]$
 - ❑ Setosa = ?
 - ❑ Vesicolor = ?
 - ❑ Virginica = ?
- ❑ Độ chính xác tổng thể: $\sum C[i, i] / C = ?$

8

8

Confusion matrix (C) cho k lớp

dự đoán =>	Setosa	vesicolor	virginica
Setosa	15	0	0
vesicolor	0	16	2
virginica	0	3	14

- $C[i, j]$: số phần tử lớp **i** (dòng) được giải thuật dự đoán là lớp **j** (cột)
- $C[i, i]$: số phần tử phân lớp đúng
- Độ chính xác lớp **i**: $C[i, i] / C[i,]$
 - Setosa = 15/15
 - Vesicolor = 16/18
 - Virginica = 14/17
- Độ chính xác tổng thể: $\sum C[i, i] / C = 45/50$

9

9

Confusion matrix (C) cho 2 lớp (+/-)

dự đoán =>	dương	âm
dương	TP	FN
âm	FP	TN

TP: true positive

tổng số phần tử lớp dương được giải thuật dự đoán lớp dương

TN: true negative

tổng số phần tử lớp âm được giải thuật dự đoán là lớp âm

FP: false positive

tổng số phần tử lớp âm được giải thuật dự đoán là lớp dương

FN: false negative

tổng số phần tử lớp dương được dự đoán là lớp âm

10

10

Confusion matrix (C) cho 2 lớp (+/-)

- **Precision**
- **Recall**
- **Accuracy**
- **F1**

dự đoán =>	dương	âm
dương	TP	FN
âm	FP	TN

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

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

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

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

dự đoán =>	dương	âm
dương	10 (TP)	5 (FN)
âm	8 (FP)	22 (TN)

11

11

Confusion matrix (C) cho 2 lớp (+/-)

dự đoán =>	dương	âm
dương	10 (TP)	5 (FN)
âm	8 (FP)	22 (TN)

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

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

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

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

$$\text{Precision} = 10 / (10 + 8) = 0.56$$

$$\text{Recall} = 10 / (10 + 5) = 0.67$$

$$\begin{aligned} \text{Accuracy} &= (10 + 22) / (10 + 5 + 8 + 22) \\ &= 32 / 45 = 0.71 \end{aligned}$$

$$\begin{aligned} \text{F1} &= 2 \times \text{precision} \times \text{recall} / (\text{precision} + \text{recall}) \\ &= 0.75 / 1.23 = 0.61 \end{aligned}$$

12

12

Bài kiểm tra 15 phút

Giả sử tập dữ liệu có 40000 mẫu tin trong đó có 8 mẫu tin thuộc lớp dương (+1) và 39992 mẫu tin thuộc lớp âm (-1), có hai mô hình phân lớp M1 và M2 cho kết quả tương ứng trong bảng 1, 2 như bên dưới. Anh/chị hãy cho biết mô hình nào thích hợp để xử lý tập dữ liệu trên? Hãy giải thích lý do cho lựa chọn đó.

Ma trận confusion thu được từ mô hình M1 (bảng bên trái) và M2 (bảng bên phải)

dự báo =>	dương	âm
dương	1	7
âm	1	39991

dự báo =>	dương	âm
dương	8	0
âm	32	39960