Khoa Công Nghệ Thông Tin
Trường Đại Học Cần Thơ

Đánh giá hiệu quả của giải thuật học

Trần Nguyễn Minh Thư
tnmthu@cit.ctu.edu.vn

Cần Thơ
02-2015

Nội dung

- Nghi thức kiểm tra
- Các chỉ số đánh giá

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

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

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 = số phần tử)

  => Vd leave 1 out

5

Chỉ số đánh giá

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: ∑C[i,i] / C

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: ∑C[i,i] / C = ?

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: ∑C[i,i] / C = 45/50

9

Confusion matrix (C) cho 2 lớp (+/-)

Ma trận contingency

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

Confusion matrix (C) cho 2 lớp (+/-)

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

Precsion
Recall
Accuracy
F1

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

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

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

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

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

11

Confusion matrix (C) cho 2 lớp (+/-)

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

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

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

Precision = 10/(10+8) = 0.56
Recall = 10/(10+5) = 0.67
Accuracy=(10+22)/10+5+8+22)
= 32/45 = 0.71
F1 = 2 x precision x recall /
(prec + recall) = 0.75/1.23 = 0.61

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

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

12

Dữ liệu không cân bằng

- nếu dữ liệu không cân bằng
  - lớp positive có số lượng rất nhỏ so với lớp negative
  - ví dụ : positive = 5%, negative = 95%
  - một giải thuật học có thể cho kết quả 95% độ chính xác khi phân loại, nhưng chúng ta có thể mất hoàn toàn lớp positive
- khả năng tách lớp positive từ lớp negative

13