Metrics and Evaluation
Anett Seeland

DFKI Bremen & Universität Bremen
Robotics Innovations Center
Director: Prof. Dr. Frank Kirchner
www.dfki.de/robotics
robotics@dfki.de

Universität Bremen

Outline

(1) Performance metrics

(2) Sampling limited data

Outline

(1) Performance metrics

(2) Sampling limited data

Performance metrics

- A classifier C was built on a large training set an then tested on a set with m = 100 instances. Of the 50 true instances C predicted 40 to be true. And on the 50 false instances C said that 45 were false.

1. Set up the confusion matrix.

Performance metrics

- A classifier C was built on a large training set an then tested on a set with m = 100 instances. Of the 50 true instances C predicted 40 to be true. And on the 50 false instances C said that 45 were false.

1. Set up the confusion matrix.

		predicted class	
		yes	no
actual class	yes	40	10
	no	5	45

Performance metrics

- A classifier C was built on a large training set an then tested on a set with m = 100 instances. Of the 50 true instances C predicted 40 to be true. And on the 50 false instances C said that 45 were false.

1. Set up the confusion matrix.
2. Calculate the performance metrics: TP rate, FP rate, Specitivity, Precision, F-measure, Accuracy and Error Rate

Common metrics derived from the confusion matrix

		predicted class	
		yes	no
actual class	Yes	40	10
	No	5	45

- TP rate = Recall = Sensitivity = 4/5 = 0.8
- FP rate = FP / (FP + TN) = 0.1
- Specitivity = TN / (FP + TN) = 1 – FP rate = 0.9
- Precision = Positiv predictive value = TP / (TP + FP) = 8/9 = 0.88
- F-measure = $\dfrac{2}{1/\text{Recall} + 1/\text{Precision}}$ = 16/19 = 0.8421
- Accuracy = success rate = (TP + TN) / (TP + FN + FP + TN) = 0.85
- Error rate = 1 – success rate = 0.15

Performance metrics

- A classifier C was built on a large training set an then tested on a set with m = 100 instances. Of the 50 true instances C predicted 40 to be true. And on the 50 false instances C said that 45 were false.

1. Set up the confusion matrix.
2. Calculate the performance metrics: TP rate, FP rate, Specitivity, Precision, F-measure, Accuracy and Error Rate.
3. Calculate the normalized mutual information.

Normalized mutual information

$$I(A;P) = \sum_{a \in A} \sum_{g \in P} p(a,g) \log_2 \frac{p(a,g)}{p(a) \cdot p(g)}$$

		predicted class P	
		yes	no
actual class A	Yes	40	10
	No	5	45

$$p(a,g) = p(A=a, P=g) = \frac{c_{ag}}{m}$$

$$p(a) = p(A=a) = 1/m \sum_g c_{ag}$$

$$p(g) = p(P=g) = 1/m \sum_a c_{ag}$$

- p(A=Y,P=Y) = 40/100 = 0.4
- p(A=Y,P=N) = 10/100 = 0.1
- p(A=N,P=Y) = 5/100 = 0.05
- p(A=N,P=N) = 45/100 = 0.45

- p(A=Y) = (40+10)/100 = 0.5
- p(A=N) = (5+45)/100 = 0.5
- p(P=Y) = (40+5)/100 = 0.45
- p(P=N) = (10+45)/100 = 0.55

Normalized mutual information

$$I(A;P) = \sum_{a \in A} \sum_{g \in P} p(a,g) \log_2 \frac{p(a,g)}{p(a) \cdot p(g)}$$

		predicted class P	
		yes	no
actual class A	Yes	40	10
	No	5	45

- p(A=Y,P=Y) = 40/100 = 0.4
- p(A=Y,P=N) = 10/100 = 0.1
- p(A=N,P=Y) = 5/100 = 0.05
- p(A=N,P=N) = 45/100 = 0.45
- p(A=Y) = (40+10)/100 = 0.5
- p(A=N) = (5+45)/100 = 0.5
- p(P=Y) = (40+5)/100 = 0.45
- p(P=N) = (10+45)/100 = 0.55

$$I(A,P) = 0.4 \log_2 (0.4/(0.5*0.45)) + 0.1 \log_2 (0.1/(0.5*0.55)) + 0.05 \log_2 (0.05/(0.5*0.45)) + 0.45 \log_2 (0.45/(0.5*0.55)) \approx 0.3973$$

Normalized mutual information

		predicted class P	
		yes	no
actual class A	Yes	40	10
	No	5	45

$$I(A;P) = \sum_{a \in A} \sum_{g \in P} p(a,g) \log_2 \frac{p(a,g)}{p(a) \cdot p(g)}$$

I(A;P) = 0.4 $\log_2$ (0.4/(0.5*0.45)) + 0.1 $\log_2$ (0.1/(0.5*0.55)) +

0.05 $\log_2$ (0.05/(0.5*0.45)) + 0.45 $\log_2$ (0.45/(0.5*0.55))

≈ 0.3973

$$NI(A;P) = \frac{I(A;P)}{H(A)} \qquad H(A) = -\sum_{a \in A} p(a) \log_2 p(a)$$

NI(A;P) ≈ 0.3973 / (-0.5*log2(0.5)-0.5*log2(0.5))

≈ 0.3973 / 1 ≈ 0.3973

ROC graphs

Inst #	Class	Score	Inst #	Class	Score
1	Yes	0.95	11	No	0.10
2	Yes	0.28	12	Yes	0.59
3	Yes	0.90	13	No	0.57
4	No	0.62	14	No	0.29
5	No	0.85	15	Yes	0.56
6	Yes	0.64	16	No	0.36
7	Yes	0.53	17	No	0.80
8	No	0.38	18	No	0.43
9	Yes	0.67	19	Yes	0.76
10	No	0.44	20	Yes	0.71

1. Draw a ROC graph.
2. How should you choose the treshold to obtain highest accuracy?

   → 0.53

Outline

(1) Performance metrics

(2) Sampling limited data

Sampling methods – repetition

- Holdout methods
- Cross validation
- Leave-one-out
- Bootstrap

Leave-one-out – a special example

- Assume you have a random data set with exactly the same number of each of two classes. A classifier is used that predicts the majority class.

1. True error rate?

   → 50%

2. Problem with Leave-one-out?

   → In each fold, the opposite class to the test instance is in the mayority

   → Prediction will always be incorrect, sample error = 100%

0.632 Bootstrap – a special example

- Assume you have a random data set with two classes and use a classifier that memorizes the training set.

1. True error rate?

   → 50%

2. Sample error rate?

   → $e_{final} = 0.632 \cdot e_{test} + 0.368 \cdot e_{training}$

   $= 0.632 \cdot 50\% + 0.368 \cdot 0\% = 31.6\%$

   → misleadingly optimistic

Thank you!

DFKI Bremen & Universität Bremen
Robotics Innovations Center
Director: Prof. Dr. Frank Kirchner
www.dfki.de/robotics
robotics@dfki.de