



# Evaluation

Albert Bifet



April 2012

# COMP423A/COMP523A Data Stream Mining

## Outline

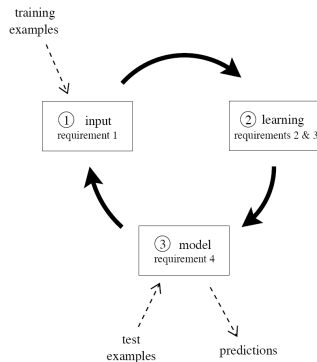
1. Introduction
2. Stream Algorithmics
3. Concept drift
4. **Evaluation**
5. Classification
6. Ensemble Methods
7. Regression
8. Clustering
9. Frequent Pattern Mining
10. Distributed Streaming



## Big Data & Real Time

# Data stream classification cycle

1. Process an example at a time, and inspect it only once (at most)
2. Use a limited amount of memory
3. Work in a limited amount of time
4. Be ready to predict at any point



# Evaluation

1. Error estimation: *Hold-out or Prequential*
2. Evaluation performance measures: *Accuracy or  $\kappa$ -statistic*
3. Statistical significance validation: *MacNemar or Nemenyi test*

Evaluation Framework

# Error Estimation

## Data available for testing

- ▶ Holdout an independent test set
- ▶ Apply the current decision model to the test set, at regular time intervals
- ▶ The loss estimated in the holdout is an unbiased estimator

Holdout Evaluation

# 1. Error Estimation

## No data available for testing

- ▶ The error of a model is computed from the sequence of examples.
- ▶ For each example in the stream, the actual model makes a prediction, and then uses it to update the model.

Prequential or  
Interleaved-Test-Then-Train

# 1. Error Estimation

## Hold-out or Prequential?

Hold-out is more accurate, but needs data for testing.

- ▶ Use prequential to approximate Hold-out
- ▶ Estimate accuracy using sliding windows or fading factors

Hold-out or Prequential or  
Interleaved-Test-Then-Train



## 2. Evaluation performance measures

	Predicted Class+	Predicted Class-	Total
Correct Class+	75	8	83
Correct Class-	7	10	17
Total	82	18	100

Table : Simple confusion matrix example

- ▶ Accuracy =  $\frac{75}{100} + \frac{10}{100} = \frac{75}{83} \frac{83}{100} + \frac{10}{17} \frac{17}{100} = 85\%$
- ▶ Arithmetic mean =  $(\frac{75}{83} + \frac{10}{17})/2 = 74.59\%$
- ▶ Geometric mean =  $\sqrt{\frac{75}{83} \frac{10}{17}} = 72.90\%$

## 2. Performance Measures with Unbalanced Classes

	Predicted Class+	Predicted Class-	Total
Correct Class+	75	8	83
Correct Class-	7	10	17
Total	82	18	100

Table : Simple confusion matrix example

	Predicted Class+	Predicted Class-	Total
Correct Class+	68.06	14.94	83
Correct Class-	13.94	3.06	17
Total	82	18	100

Table : Confusion matrix for chance predictor

## 2. Performance Measures with Unbalanced Classes

### Kappa Statistic

- ▶  $p_0$ : classifier's prequential accuracy
- ▶  $p_c$ : probability that a chance classifier makes a correct prediction.
- ▶  $\kappa$  statistic

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

- ▶  $\kappa = 1$  if the classifier is always correct
- ▶  $\kappa = 0$  if the predictions coincide with the correct ones as often as those of the chance classifier

### Forgetting mechanism for estimating prequential kappa

Sliding window of size  $w$  with the most recent observations

### 3. Statistical significance validation (2 Classifiers)

	Classifier A Class+	Classifier A Class-	Total
Classifier B Class+	c	a	c+a
Classifier B Class-	b	d	b+d
Total	c+b	a+d	a+b+c+d

$$M = |a - b - 1|^2 / (a + b)$$

The test follows the  $\chi^2$  distribution. At 0.99 confidence it rejects the null hypothesis (the performances are equal) if  $M > 6.635$ .

McNemar test

### 3. Statistical significance validation (> 2 Classifiers)

Two classifiers are performing differently if the corresponding average ranks differ by at least the critical difference

$$CD = q_{\alpha} \sqrt{\frac{k(k+1)}{6N}}$$

- ▶  $k$  is the number of learners,  $N$  is the number of datasets,
- ▶ critical values  $q_{\alpha}$  are based on the Studentized range statistic divided by  $\sqrt{2}$ .

Nemenyi test

### 3. Statistical significance validation (> 2 Classifiers)

Two classifiers are performing differently if the corresponding average ranks differ by at least the critical difference

$$CD = q_{\alpha} \sqrt{\frac{k(k+1)}{6N}}$$

- ▶  $k$  is the number of learners,  $N$  is the number of datasets,
- ▶ critical values  $q_{\alpha}$  are based on the Studentized range statistic divided by  $\sqrt{2}$ .

# classifiers	2	3	4	5	6	7
$q_{0.05}$	1.960	2.343	2.569	2.728	2.850	2.949
$q_{0.10}$	1.645	2.052	2.291	2.459	2.589	2.693

Table : Critical values for the Nemenyi test

## Cost Evaluation Example

	Accuracy	Time	Memory
Classifier A	70%	100	20
Classifier B	80%	20	40

Which classifier is performing better?

# RAM-Hours

## RAM-Hour

Every GB of RAM deployed for 1 hour

## Cloud Computing Rental Cost Options





## Cost Evaluation Example

	Accuracy	Time	Memory	<b>RAM-Hours</b>
Classifier A	70%	100	20	<b>2,000</b>
Classifier B	80%	20	40	<b>800</b>

Which classifier is performing better?

# Evaluation

1. Error estimation: *Hold-out or Prequential*
2. Evaluation performance measures: *Accuracy or  $\kappa$ -statistic*
3. Statistical significance validation: *MacNemar or Nemenyi test*
4. **Resources needed:** *time and memory or RAM-Hours*

Evaluation Framework