HAP/LAP

Introduction to Machine Learning

Evaluation

Olatz Arbelaitz: olatz.arbelaitz@ehu.eus
www.aldapa.eus

Erasmus Mundus

HAP/LAP

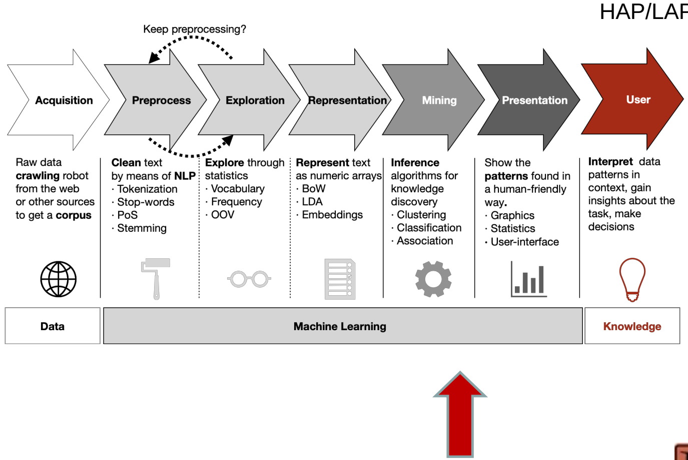
Topics

1.- Introduction. Machine Learning for LNP
2.- Learning with WEKA software:
- 2.1.-Introduction
- 2.2.-Preprocessing
  - Attribute (feature) selection
- **2.3.-Evaluation**
- 2.4.- Basic ML algorithms: Naive Bayes, K-NN, Decision Trees, Rules, …

Erasmus Mundus

Evaluation

HAP/LAP



Erasmus Mundus

Evaluation

HAP/LAP

Classification process

- **Division of the corpora** (*Test options*)
  - train / test
  - Cross-validation
- Classifier (*Classify*)
  - Set parameters
- **Evaluation** (*Classifier Output*)
  - Confusion matrix
  - Precision/recall
  - Microaveraging/macroaveraging

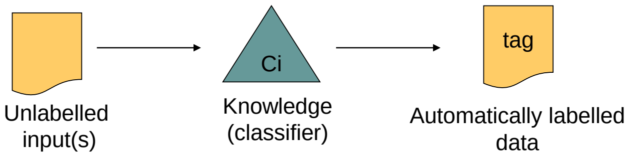
Erasmus Mundus

Evaluation

Division of the corpora

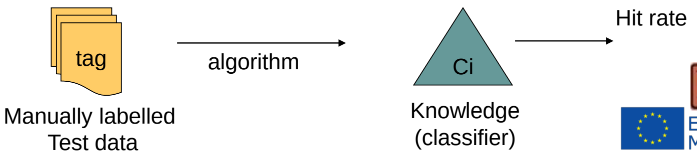
HAP/LAP

After learning → Make decisions, **obtain results**



Is the obtained result good?

Test→ to measure the quality of the built classifier



Erasmus Mundus

Evaluation

Division of the Corpus

HAP/LAP

Division of the corpus:

- **Train / Test**
  - Train: % 80 - % 66 of the corpus → for training (learning)
  - Develop/validate part of the trainining → for tunning
  - Test: % 20 - % 33 of the corpus → for testing (error rate of the last model)

Train (%66)		Test (% 33)
Training set	Develop/validation set	Test set

Once: hold out

Erasmus Mundus

Evaluation

Learning Process

HAP/LAP

Learning process train/test:

1. Divide the available data into **training**, **validation** and **test** set
2. Select architecture and training parameters
3. Train the model using the **training set**
4. Evaluate the model using the **validation set**
5. Repeat steps 2 through 4 using different architectures and training parameters
6. Select the best model and train it using data from the training and validation set
7. Assess this final model using the **test set**

The error rate obtained with the validation set is often smaller than the real error because this set has been used to select the model.
If Cross Validation is used in the learning process, steps 3 and 4 must be repeated in K folds.

Erasmus Mundus

Evaluation

Division of the Corpus

HAP/LAP

- **Cross-validation:** for small corpora
  - Division in *k* folds (10)
  - *k*-1 train and 1 test
  - Learn *k* times
  - Average error rate: errors in the *k* iterations/*k*

  10-folds or 2 times 5-folds

  5-folds (when the corpus is small)

Exp. 1					Test set
Exp. 2				Test set	
Exp. 3			Test set		
Exp. 4		Test set			
Exp. 5	Test set				

Erasmus Mundus

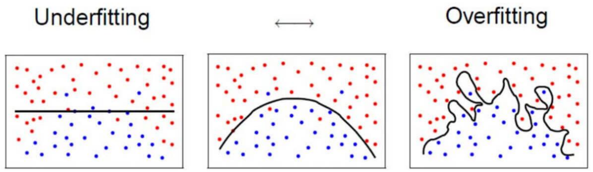
Evaluation

Division of the Corpus

HAP/LAP

- **Leave-one-out** cross-validation:
  - Similar to the previous but in each of the parts an example is left out for testing and the rest are used for training
  - K = number of examples → number of iterations
  - All the examples are used for training and for testing
  - Used in sparse databases

Building a single classifier with all the training data can lead to overfitting. The use of cross-validation can reduce it (train classifiers with different data)



Evaluation

Division of the Corpus

HAP/LAP

How many cross-validation folds

When the corpus is divided in many folds:
+ Error bias is small
- Error variance is big
- High computational cost

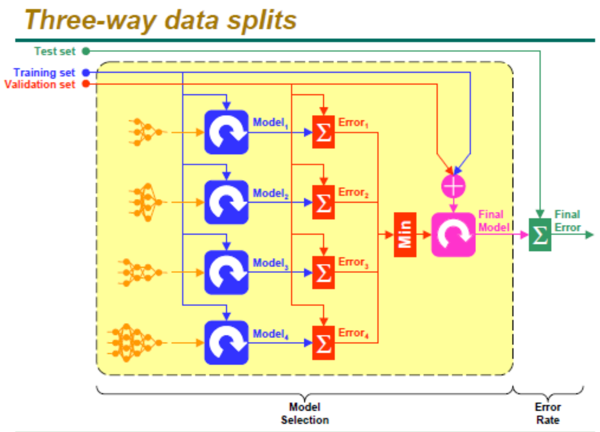
When the corpus is divided in few folds:
+ Low computational cost
+ Error variance is small
- Error bias is big

In practice, the number of folds is selected according to the dataset
- In big datasets 3-Fold Cross Validation can be enough
- In sparse datasets leave-one-out can be used to train for most of the examples
- **Usually: K-Fold Cross Validation K=10**

Evaluation

Learning Process

HAP/LAP



Evaluation

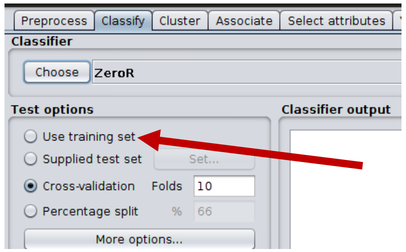
Division of the Corpora

HAP/LAP

To divide the corpora:

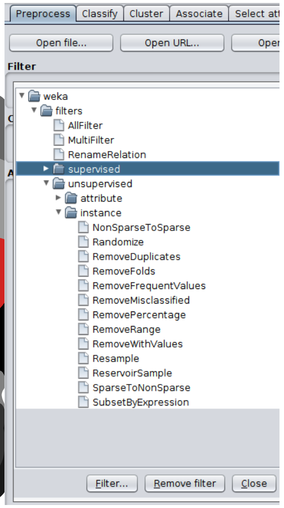
Test options:

- Use training set: use all the examples used for learning → **NO**
- Supplied test set: provide test file
  - press *Set* and select file
- Cross - validation (*fold* indicate number of folds)
- Percentage split (percentage used for learning)



WEKA: preprocessing
Filters

HAP/LAP



Unsupervised

Instances

Remove (delete)

RemoveFolds: delete a fold

RemoveMisclassified: delete wrongly classified

(requires selecting classifier)

RemovePercentage: delete %

→ for **train/test division**

RemoveRange: which instances to delete

Erasmus Mundus

Evaluation

HAP/LAP

For train/test division

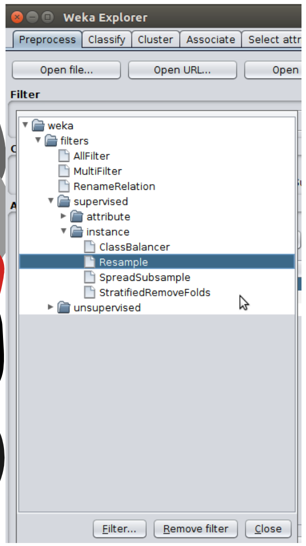
– RemovePercentage: for dividing train/test
- training set:
  – Open complete file
  – Select **RemovePercentage** filter
  – Write percentage desired for division
  – Apply filter
  – Save the generated dataset as a new file
- test set:
  – Open complete file (or use undo)
  – Select **RemovePercentage** filter
  – Activate **invertSelection** in the filter
  – Apply filter
  – Save the generated dataset as a new file

The class is not taken into account in the division!

Erasmus Mundus

WEKA: preprocessing
Filters

HAP/LAP



Supervised

Instances

(re)sampling

Delete instances to maintain distribution when the number of examples in the categories is unbalanced, *sampling*

- **StratifiedRemoveFolds**

  (for crossvalidation)

Erasmus Mundus

Evaluation

HAP/LAP

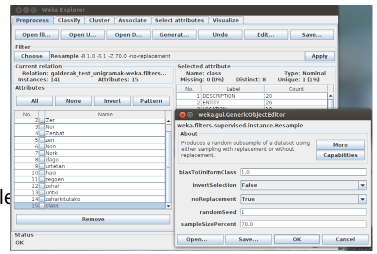
For train/test division

To divide taking into account the class

Resample

training set:

- Open complete file
- Select **Resample (supervised)** filter
- Write desired percentage
- *Select BiasToUniformClass*
- *noReplacement: true*
- Apply filter
- Save the generated dataset as a new file



test set:

- Open complete file (or undo)
- Select **invertSelection** in filter
- Apply filter
- Save the generated dataset as a new file

Erasmus Mundus

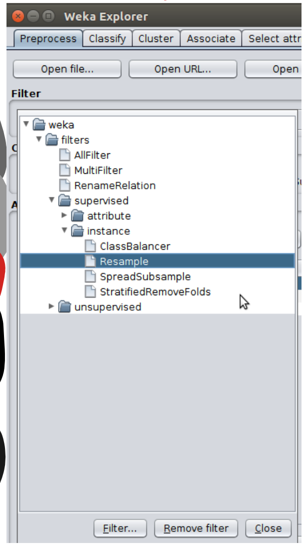
WEKA: preprocessing

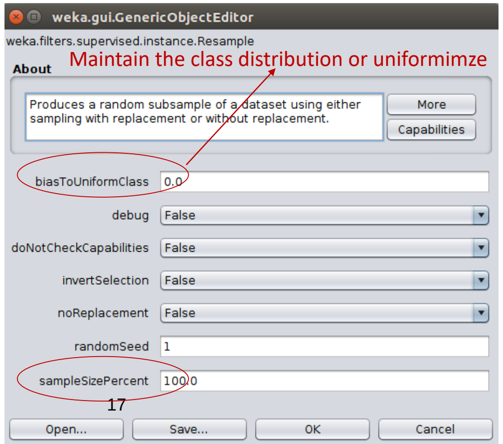
Filters. Resample

HAP/LAP

resample to generate subsets of data

Delete instances to maintain the distribution by categories (when the number of examples in unbalanced)





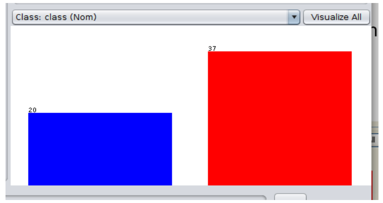
WEKA: preprocessing

HAP/LAP

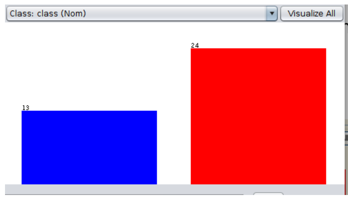
Filters. Resample

Labor

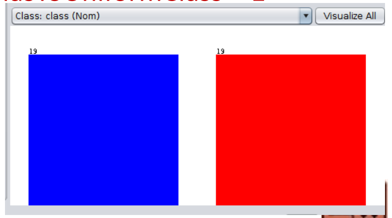
- Original: 57 instances
- Resample (% 67): 38 instances



BiasToUniformClass = 0



BiasToUniformClass = 1



Erasmus Mundus

For Crossvalidation

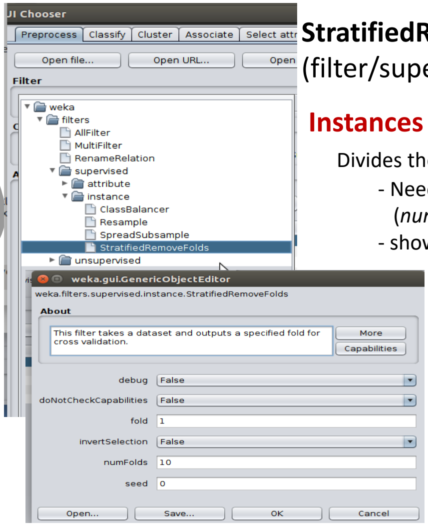
HAP/LAP

StratifiedRemoveFolds
(filter/supervised/instances)

Instances

Divides the dataset in sets

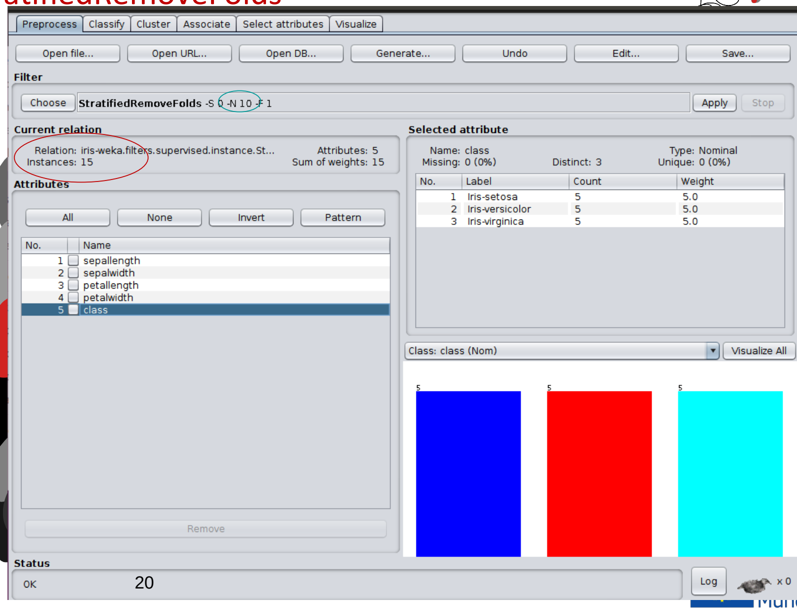
- Need to indicate number of folds (*numFolds*)
- shows one of the folds(*fold*)



Try iris:
- numFolds
- fold

Erasmus Mundus

StratifiedRemoveFolds



Evaluation

For train/test division Assignment

HAP/LAP

-Divide ReutersGrain-train.arff into **train** (70%) and **development** (30%) sets maintaining the original class distribution.

	N. Inst	N. Feat	Class(0)	Class(1)
Train				
Dev				

-Transform into BoW

	N. Inst	N. Feat	Class(0)	Class(1)
Train				
Dev				

Erasmus Mundus

Evaluation

For train/test division Assignment

HAP/LAP

- Do both options have the same number of features?
- Is it possible to train/test with different features?
- How can the development or test set be generated with the same dictionary?

Erasmus Mundus

Evaluation

For train/test division Assignment

HAP/LAP

- How can the development or test set be generated with the same dictionary?

-Save dictionary in :

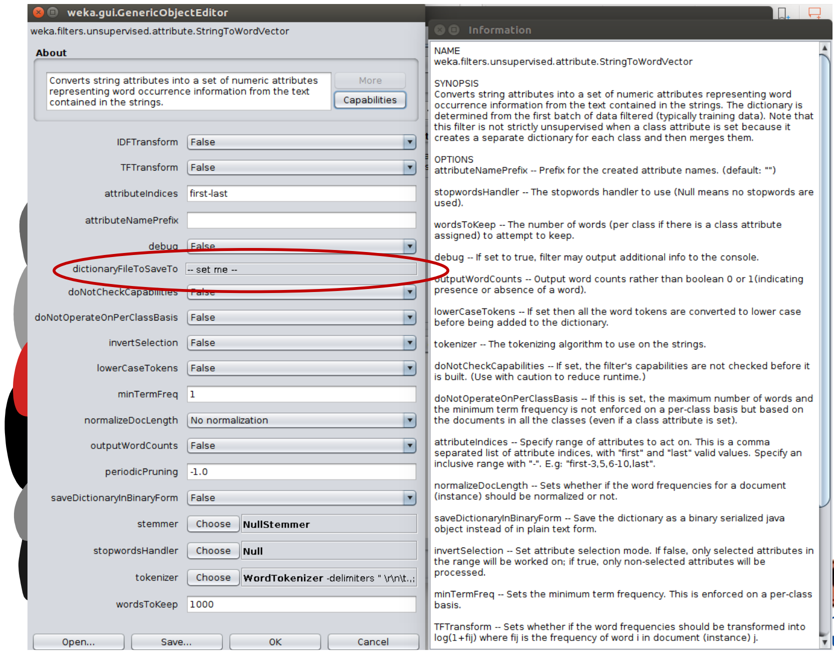
Filter/Unsupervised/Attribute/StringtoWordVector

- Use the dictionary and same options in

Filter/Unsupervised/Attribute/FixedDictionaryStringtoWordVector

- And then, what happens if we select features?

Erasmus Mundus



Evaluation

Feature selection and test

HAP/LAP

How to select features and then test? AttributeSelectedClassifier

- selection based on *train* file
- Compatible *test* file required

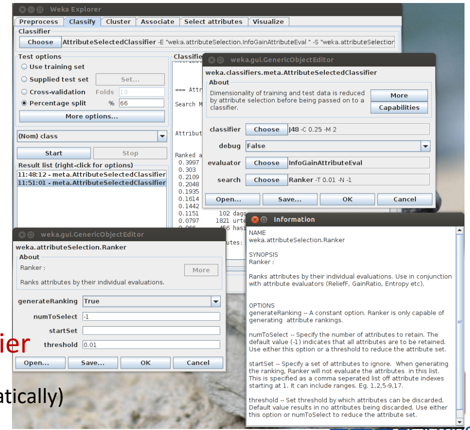
Done when classifying:

AttributeSelectedClassifier

(meta)

- Classifier
- Evaluator (features)
- Search

Try with iris



misc.InputMappedClassifier

(last versions of Weka do it automatically)

Erasmus Mundus

Evaluation

How to evaluate?

Confusion Matrix

HAP/LAP

Binary classifier:

Real→ C = 1 positive class C = 0 negative class

Prediction→ C_M = 1 positive class predicted and C_M = 0 negative class predicted

(Confusion Matrix)

		Real class	
		C = 1	C = 0
Predicted Class	C_M = 1	TP	FP
	C_M = 0	FN	TN

TP = True Positive. The example was postive and classified as postive

TN = True Negative. The example was negative and classified as negative

FP = False Positive. The example was negative and classified as postive

FN = False Negative. The example was postive and classified as negative

Erasmus Mundus

Evaluation

How to evaluate?

HAP/LAP

Binary classifier:

Hit rate (accuracy): proportion of examples with correct prediction among the tested examples.

$$Accuracy(AC) = \frac{TP+TN}{TP+TN+FP+FN}$$

Precision taking into account the examples classified as being of the positive class, number of hits among them

$$Precision = \frac{TP}{TP+FP}$$

Recall: number of hits among the examples of the positive class which where tested,

$$Recall = \frac{TP}{TP+FN}$$

Erasmus Mundus

Evaluation

How to evaluate?

HAP/LAP

Binary classifier:

F_1-score (F-measure): Harmonic mean of the precision and

$$F_1 = \frac{2\times precision \times recall}{precision + recall}$$

Error: proportion of examples with wrong prediction among the tested examples

$$Error(1-AC) = \frac{FP+FN}{TP+TN+FP+FN}$$

Erasmus Mundus

Evaluation

How to evaluate?

HAP/LAP

- Example: *weather.arff*
  - 4features, 100 instances, class={**good, bad**}
  - Classifier: *rules* **JRip**
  - Test options: Percentage split (% 66) → 34 instan.

- Classifier Output

=== Run information ===

Scheme: weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1

Relation: weather

Instances: 100

Attributes: 5 outlook, temperature, humidity, windy, play

Test mode: split 66.0% train, remainder test

Erasmus Mundus

Evaluation

How to evaluate?

HAP/LAP

Correctly Classified Instances	28	82.3529 %
Incorrectly Classified Instances	6	17.6471 %

percentage 28*100/34
6*100/34

=== Confusion Matrix ===

a	b	classified as
18	1	a = good
5	10	b = bad

Category		classified as	
		YES	NO
real	YES	TP	FN
	NO	FP	TN

category; **good**

TP = true positive = 18 FN = false negative = 1

FP = false positive = 5 TN = true negative = 10

Correctly Classified Instances = **TP + TN** = 18 + 10 = 28

Incorrectly Classified Instances = **FN + FP** = 1 + 5 = 6

Erasmus Mundus

How to evaluate?

HAP/LAP

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.947	0.333	0.783	0.947	0.857	0.811	good
	0.667	0.053	0.909	0.667	0.769	0.811	bad
Weighted Avg.	0.824	0.209	0.838	0.824	0.818	0.811	

=== Confusion Matrix ===

a	b	classified as
18	1	a = good
5	10	b = bad

Category		classified as	
		YES	NO
real	YES	TP	FN
	NO	FP	TN

TP Rate = TP/(TP+FN) = 18/(18+1) = 0.947 → good

FP Rate = FP/(FP+TN) = 5/(5+10) = 0.333

↓ the probability of falsely rejecting the null hypothesis for a particular test.

Erasmus Mundus

Evaluation

How to evaluate?

HAP/LAP

Category **good**		classified as	
		YES	NO
real	YES	18	1
	NO	5	10

Category **bad**		classified as	
		YES	NO
real	YES	10	5
	NO	1	18

TP = true positive = 18
FN = false negative = 1
FP = false positive = 5
TN = true negative = 10

TP = true positive = 10
FN = false negative = 5
FP = false positive = 1
TN = true negative = 18

Weighted Average:

TP Rate = [TP Rate (good) * 19 + TP Rate (bad) * 15] / 34

FP Rate = [FP Rate (good) * 19 + FP Rate (bad) * 15] / 34

Erasmus Mundus

Evaluation

How to evaluate?

HAP/LAP

Category **good**		classified as	
		YES	**NO**
real	**YES**	18	1
	NO	5	10

TP = true positive = 18
FN = false negative = 1
FP = false positive = 5
TN = true negative = 10

Category **bad**		classified as	
		YES	**NO**
real	**YES**	10	5
	NO	1	18

TP = true positive = 10
FN = false negative = 5
FP = false positive = 1
TN = true negative = 18

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.947	0.333	0.783	0.947	0.857	0.811	good
	0.667	0.053	0.909	0.667	0.769	0.811	bad
Weighted Avg.	0.824	0.209	0.838	0.824	0.818	0.811	

Precision (good) = [TP / (TP + FP)] = 18 / (18 + 5) = 0,783
Recall (good) = [TP / (TP + FN)] = 18 / (18 + 1) = 0,947
F-measure (good) = [2·Pr·Rc / (Pr + Rc)] = 0,857

Erasmus Mundus

Evaluation

How to evaluate?

HAP/LAP

Category **good**		classified as	
		YES	**NO**
real	**YES**	18	1
	NO	5	10

TP = true positive = 18
FN = false negative = 1
FP = false positive = 5
TN = true negative = 10

Category **bad**		classified as	
		YES	**NO**
real	**YES**	10	5
	NO	1	18

TP = true positive = 10
FN = false negative = 5
FP = false positive = 1
TN = true negative = 18

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.947	0.333	0.783	0.947	0.857	0.811	good
	0.667	0.053	0.909	0.667	0.769	0.811	bad
Weighted Avg.	0.824	0.209	0.838	0.824	0.818	0.811	

Precision (bad) = [TP / (TP + FP)] =
Recall (bad) = [TP / (TP + FN)] =
F-measure (bad) = [2·Pr·Rc / (Pr + Rc)] =

Erasmus Mundus

Evaluation

How to evaluate?

HAP/LAP

Multiclass classifier:

Real → C = {y_1, y_2... y_e}

	Real class	
Predicted class	$C = y_j$	$C = y_i$
$C_M = y_j$	TP_j	FP_j
$C_M = y_i$	FN_j	TN_j

Confusion matrix for class y_j

$$Precision_j = \frac{TP_j}{TP_j + FP_j}$$

$$Recall_j = \frac{TP_j}{TP_j + FN_j}$$

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}$$

Erasmus Mundus

Evaluation

How to evaluate?

HAP/LAP

Example: iris.arff NaiveBayes

Time taken to build model: 0.02 seconds

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
1	0	1	1	1	Iris-setosa
0.96	0.04	0.923	0.96	0.941	Iris-versicolor
0.92	0.02	0.958	0.92	0.939	Iris-virginica

All correct

=== Confusion Matrix ===

a	b	c	classified as
50	0	0	a = Iris-setosa
0	48	2	b = Iris-versicolor
0	4	46	c = Iris-virginica

4 incorrect

Erasmus Mundus

Evaluation

How to evaluate?

HAP/LAP

=== Confusion Matrix ===

a	b	c	classified as
50	0	0	a = Iris-setosa
0	48	2	b = Iris-versicolor
0	4	46	c = Iris-virginica

Category Ci		Classified as	
		Yes	**No**
Real	**Yes**	TPi	FNi
	No	FPi	TNi

- Iris-setosa
  TP = 50, FP = 0, FN = 0, TN = 48 + 2 + 4 + 46 = 100
- Iris-versicolor
  TP = 48, FP = 4, FN = 2, TN = 50+46 = 96
- Iris-virginica
  TP = 46, FP = 2, FN = 4, TN = 50+48 = 98

Erasmus Mundus

Evaluation

How to evaluate?

HAP/LAP

Microaveraging global addition

$$\hat{\mathrm{Re}}^{\mu} = \frac{TP}{TP+FN} = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m}(TP_i + FN_i)}$$

$$\hat{\mathrm{Pr}}^{\mu} = \frac{TP}{TP+FP} = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m}(TP_i + FP_i)}$$

Macroaveraging: precision and recall are locally calculated for each category and then added

$$\hat{\mathrm{Pr}}^{M} = \frac{\sum_{i=1}^{m} \hat{\mathrm{Pr}}_i}{m} \qquad \hat{\mathrm{Re}}^{M} = \frac{\sum_{i=1}^{m} \hat{\mathrm{Re}}_i}{m}$$

Note: classifiers that are good for categories with many test instances will obtain good microaverage values

Erasmus Mundus

Evaluation

How to evaluate?

HAP/LAP

- When evaluating we take into account:
  - **Precision**: taking into account the examples classified into a category, number of hits among them
  - **Recall**: among the examples of a category, number of hits
  - Combination of both metrics (*F-score/F-measure*)
- Which is more important?
  - **precision**: what the system says is always correct (→ although it might say it fewer times)
    or,
  - **recall**: although with some mistakes to say it more

Erasmus Mundus

Evaluation

How to evaluate?

HAP/LAP

- What is more interesting, high Precision (Pr) or high Recall (Re)?
  - Both
  - Medicine: does she/he have cancer?
    - When we diagnose it as positive, to be true (Pr): precision
    - To diagnose cancer when it exists (Rc): recall
  - Classify texts from newspapers
    - The classifier is not used on its own. The classifier helps but final classification is manual
  - Web page classification in search engines
    - Wrong classifications are not that important
  - ...

Erasmus Mundus

Evaluation

Assignment

HAP/LAP

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
  5  0  0 |  a = soft
  0  3  1 |  b = hard
  1  2 12 |  c = none
```

-Which is the most trustful category?

-If we take randomly an example of each category which will the one correctly classified with higher probability?

Class	TP	FP	FN	TN	Pr	Rc	F-Measure
soft							
hard							
none							
				Micro			
				Macro			

Erasmus Mundus

Evaluation

HAP/LAP

How to know which are the done errors?

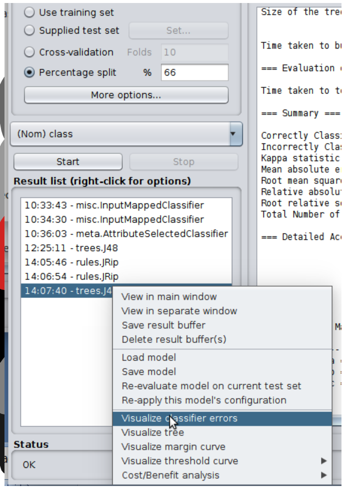
- Visualize classifier errors
- More Options: Output predictions

Erasmus Mundus

Evaluation

Visualize classifier errors

HAP/LAP



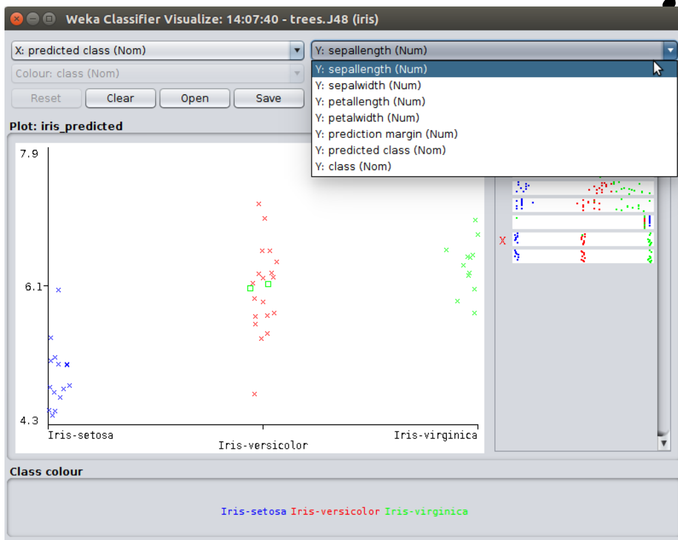
- Instance distribution in the space.
- It is possible to select attributes in axes X and Y.
- Information about errors

Erasmus Mundus

Evaluation

Visualize classifier errors

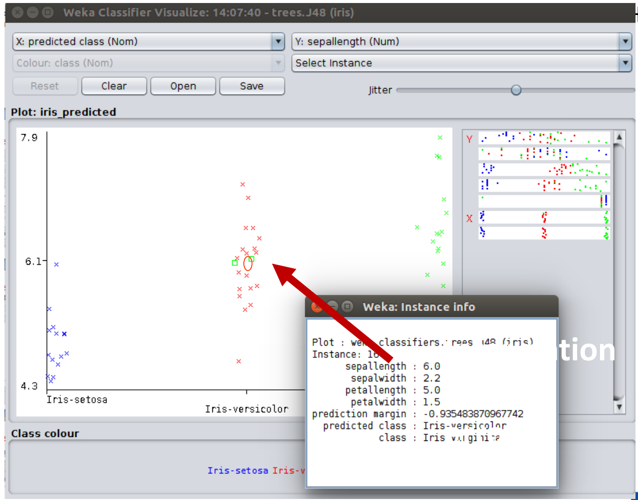
HAP/LAP



Erasmus Mundus

Evaluation

Visualize classifier errors



HAP/LAP

Erasmus Mundus