ESERCITAZIONE PIATTAFORMA WEKA

Giuseppe Castellucci – Simone Filice

Web Mining & Retrieval 2013/2014

18/03/2014

Outline

- Intro Weka
- ARFF Format
- Performance measures
  - Decision Trees
  - Confusion Matrix
  - Precision, Recall, F1, Accuracy
- Parameter Tuning
  - Knn
- Error diagnostics
  - Knn
  - High bias and High Variance

Intro WEKA

- Collection of ML algorithms - open-source Java package
  - http://www.cs.waikato.ac.nz/ml/weka/
- Documentation
  - http://www.cs.waikato.ac.nz/ml/weka/index_documentation.html
- Schemes for classification include:
  - Decision trees, rule learner
  - Naive bayes
  - KNN
  - SVM
- For classification, Weka allows train/test split or Cross-fold validation

ARFF File

- Require declarations of @RELATION, @ATTRIBUTE and @DATA
- @RELATION declaration associates a name with the dataset
  - @RELATION <relation-name>
- @ATTRIBUTE declaration specifies the name and type of an attribute
  - @ATTRIBUTE <attribute-name> <datatype>
    - Datatype can be numeric, nominal, string or date

```
@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
@ATTRIBUTE class {Setosa,Versicolor,Virginica}
```

- @DATA declaration is a single line denoting the start of the data segment

```
@DATA
1.4, 0.2, Setosa
1.4, ?, Versicolor
```

Performance measures

- Visualize IRIS dataset, its attributes and classes
  - What can we say about it?
- Execute a Decision Tree (J48) algorithm on the IRIS dataset
- In output notice:
  - Confusion matrix
  - True positive, true negative, false positive, false negative
  - Precision, recall, f1-measure, accuracy
- Visualize the tree
- And if we remove some feature?

Parameter Tuning

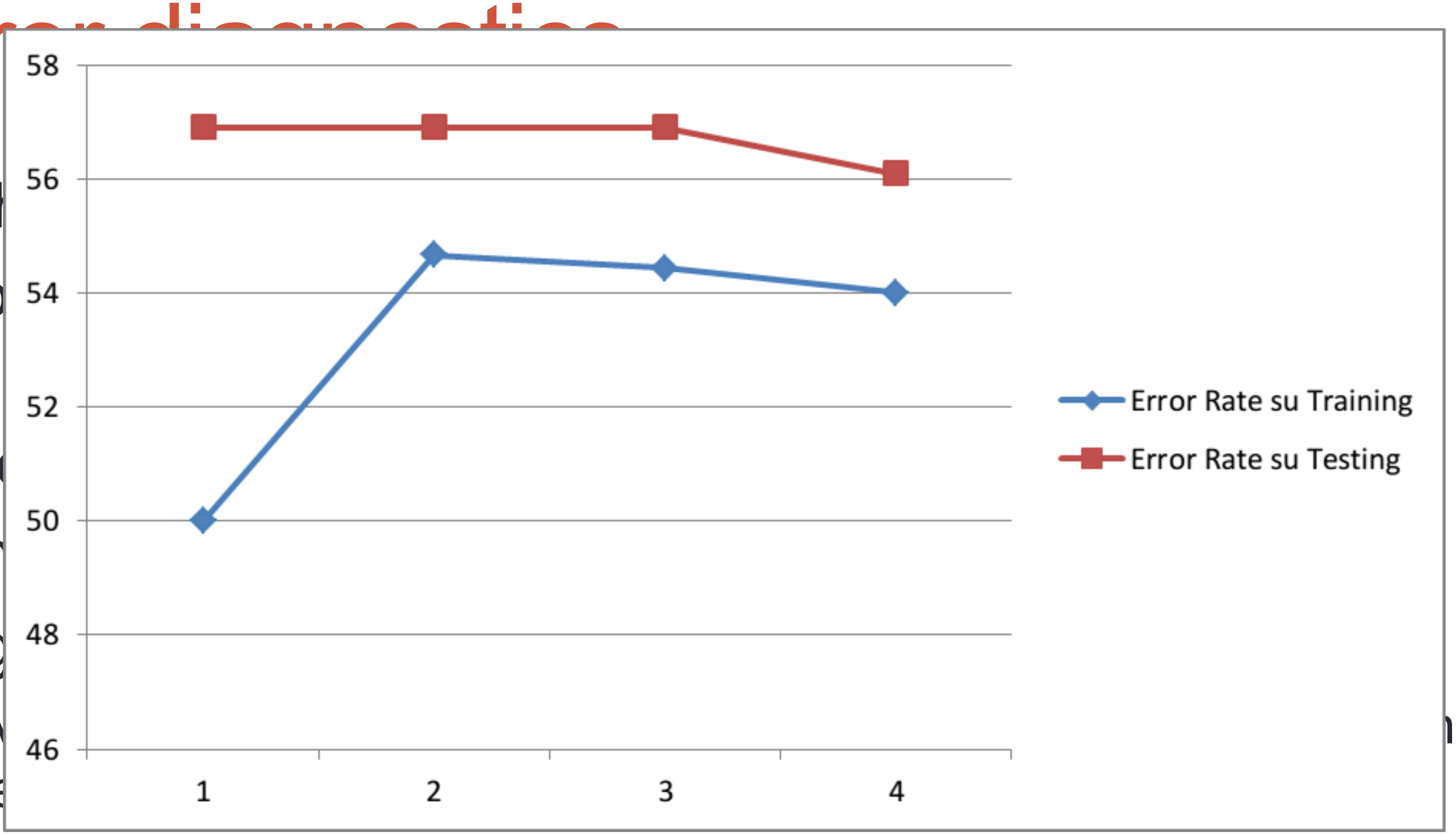
- Diabets datasets
  - First of all, visualize statistics on it
- Are classes imbalanced?
- What performance measure are suitable in this case?
- Execute the KNN learning algorithm
  - We have to choose the best K!
  - Execute a 5-fold cross validation on the training set with different values of K
  - K=1,2,5,10,15,30
- Final test measure on the test dataset

Error diagnostics

- Vehicle dataset
- Plot a learning curve on different training set size
  - 25%,50%,75%,100%
- Use Naive Bayes learning algorithm
- What can we say from the results?
- High Bias
  - After a certain value of *m*, the learning process saturates and the testing error becomes similar to the training error
- Solutions?
  - Add new informative features
  - Use a more sophisticated algorithm (or the same algorithm with a more complex parameterization)

Error diagnostics

- Ve
- Plo
  - 2
- Us
- Wh
- Hig
  - A ne
    te



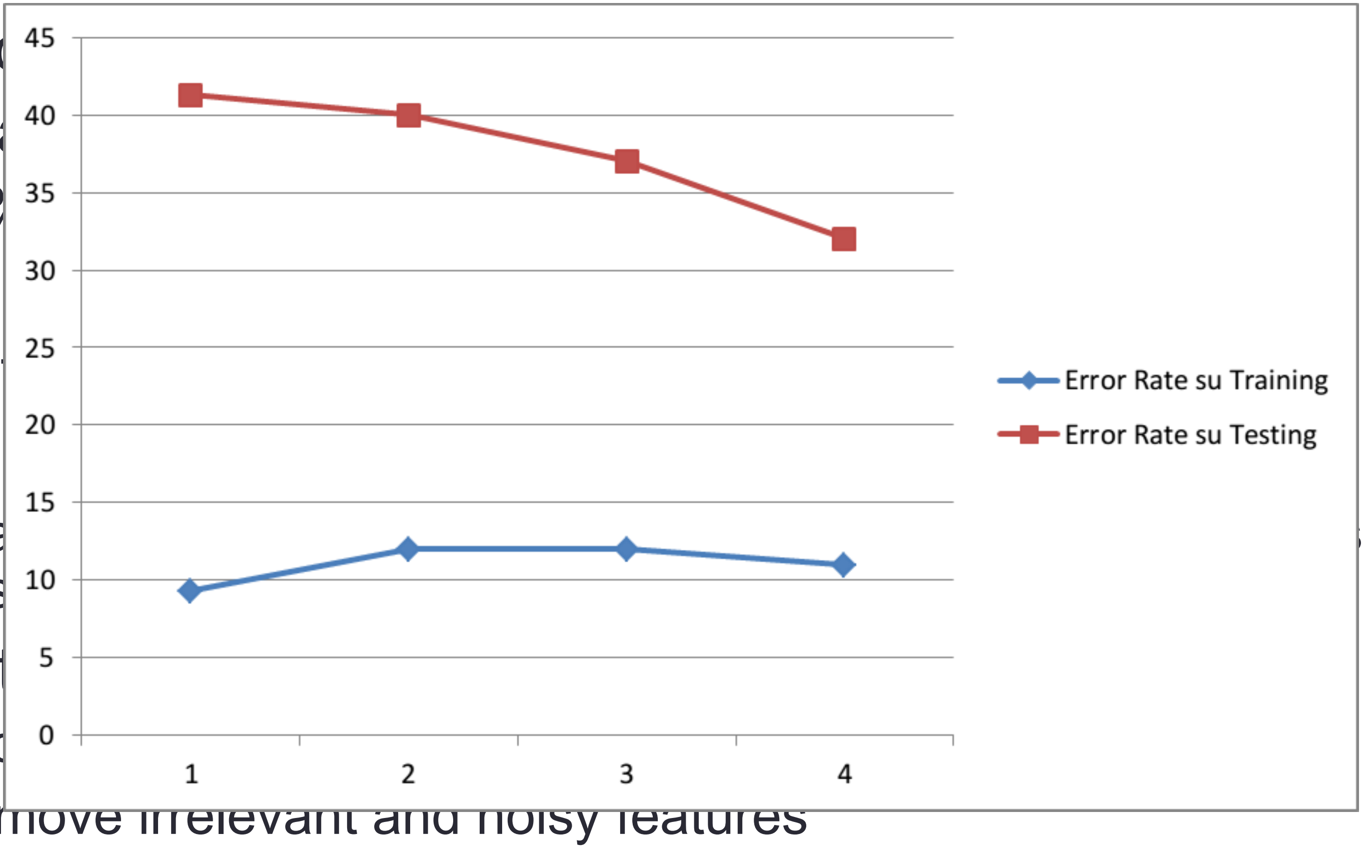
- Solutions?
  - Add new informative features
  - Use a more sophisticated algorithm (or the same algorithm with a more complex parameterization)

Error diagnostics

- Vehicle dataset
- Plot a learning curve on different training set size
  - 25%,50%,75%,100%
- Use KNN with K=2
- What can we say from the results?
- High Variance
  - A large gap between the training error and the testing error is observed. The saturation point is still not reached!
- Solutions?
  - Add new examples
  - Remove irrelevant and noisy features
  - Use a less complicated parameterization (example simpler polynomial function in regression)

Error diagnostics

- Vehi
- Plot
  - 25%
- Use
- Wha
- High
  - A la
    obs
- Solu
  - Add
  - Remove irrelevant and noisy features
  - Use a less complicated parameterization (example simpler polynomial function in regression)


Error Rate su Training
Error Rate su Testing