Data Analysis and Model Classification

Miniproject 1

**GUIDELINES – SUM UP**

OBJECTIVE

* Apply a suitable feature selection method
* Establish a classifier which should be able to generalize to new data from the same problem
* Estimate the performance of your classifier

analysis and computations : feature selection, cross validation, classifier construction

GUIDESHEET 1

DATA EXPLORATION :

* plots and statistical parameters (to get a feeling of the data)
* statistical tests :
* histogram : find similar and diff distributions between classes
* boxplot : comparison between features
* boxplot Notch parameter
* t-test : analysis of p-values
* which test for samples without normal distributions ?
* FINAL GOAL of this section : « Often, classes look different for a certain feature, but we want to objectively validate if there is a real diff or not = if the diff is statistically significant »
* Scores : classification error or classification accuracy (discussion etc)
* FINAL GOAL : choice of the best score & justification

GUIDESHEET 2

(Rmq : demande de choisir un subset de features avant de commencer cette discussion sur les classifiers)

LDA/QDA CLASSIFIERS (linear/quadratic discriminant classifiers) & TRAINING AND TESTING ERROR

* Training/testing error comparison
* Robustness of the classifier
* Complexity of the classifier
* Variability of the performance
* FINAL GOAL : choice of the best classifier & justification

CROSS VALIDATION FOR PERFORMANCE VALIDATION

* Choose the number of samples/class/fold
* Cross validation error & std
* Stable estimation of accuracy
* FINAL GOAL : Choose best model/classifier

GUIDESHEET 3

*This guidesheet should help you understand better how to select the best features for your classifier and crucially, how to correctly estimate the performance of a model (classifier) on an unseen and independent dataset*

MODEL SELECTION AND NESTED CROSS VALIDATION

* FINAL GOAL : Optimal number of features

GUIDESHEET 4

REPORT

* Report + motivates the choices you made
* discussing the results you obtained from your implementation

This report should cover :

* dataset exploration over feature selection
* classification
* role of (nested) cross-validation in your analysis

Structure:

* introduction
* methods
* results
* discussion