Ensembles

Rodrigo Fernandes de Mello

Invited Professor at Télécom ParisTech

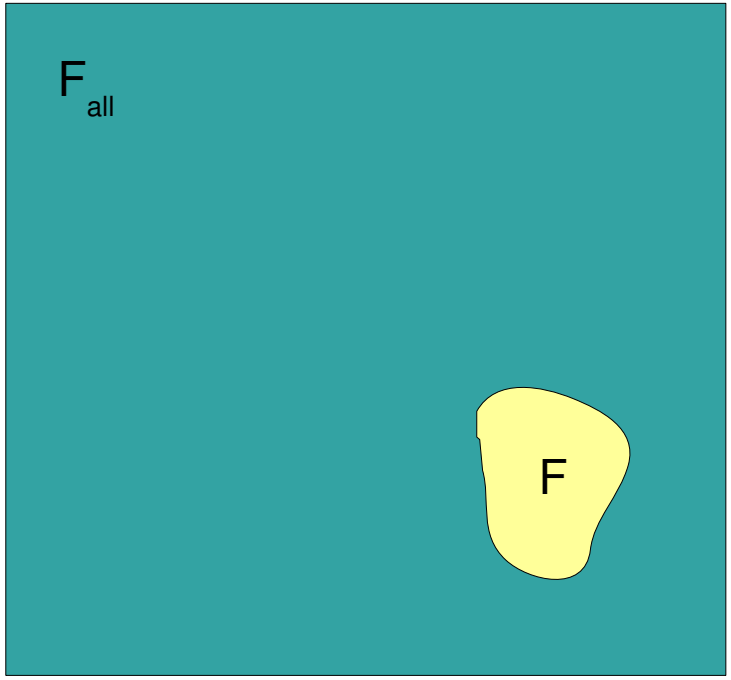
Associate Professor at Universidade de São Paulo, ICMC, Brazil

http://www.icmc.usp.br/~mello

mello@icmc.usp.br

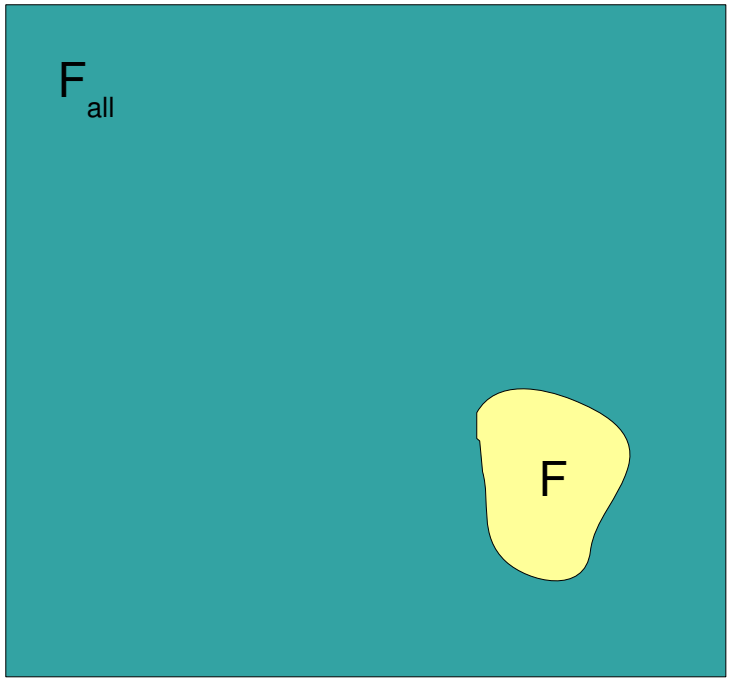
Ensembles

- Designed to reduce overfitting
  - It is central to remember the Bias-Variance Dilemma
    - The bias depends on the examples used for training
      - This is due to the Shattering Coefficient



Ensembles

- By producing several different datasets from the original
  - Based on uniform sampling with replacement
  - That brings perturbations to the input space
    - And as consequence to the distinct classifiers obtained
    - Can we formulate that and prove it?



Ensembles: Bagging

- Bagging or **B**ootstrap **Agg**regat**ing**
  - Proposed by Leo Breiman, 1994 to support the overfitting reduction
  - Given some dataset with **n** examples
    - It uses uniform sampling to form other **m** training sets with **n'** examples
    - If **n**=**n'** then, each new dataset is expected to contain approximately 63.2% of the examples of the original dataset
      - Observe this is a sampling with replacement
  - **m** models (proposed for trees) are build up:
    - Then models are combined to produce the outputs
      - By averaging, if regression
      - By voting, if classification

Ensembles: Boosting

- Boosting is based on the question posed by Kearns and Valiant (1988, 1989)
  - It attempts to convert weak learners to strong ones
    - A weak learner is defined to be a classifier that is only slightly correlated with the true classification (it can label examples better than random guessing)

- Boosting is not algorithmically constrained
  - There are several algorithms
    - AdaBoost, LPBoost, XGBoost, etc.
  - They build up some model
    - Test it to check if it is more effective than guessing
    - Given some weight for it according to its performance

Ensembles: Random Forests

- Random forests or random decision forests:
  - It constructs a multitude of decision trees at training time
    - Using different uniform samples from the original training set
  - It outputs the:
    - The main class label after a voting system, when the problem involves classification
    - Or it combines the results of individual trees (models) in case of regression problems

Summary

- Questions...