ensemble learning

Carlos Soares
(based on materials kindly provided
by João Mendes Moreira)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

plan & goals

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- introduction
- categories of methods
- popular methods
- issues

- understand the basic principles of ensemble learning
- understand the intuition and high-level algorithm of some of the most common ensemble methods

definition

- multiple models
  - **base models**
- … each of them obtained by applying a learning process to a given problem
  - e.g. same algorithm applied to different samples of the data
- … combined to make a single prediction
  - e.g. in classification, each model makes a prediction
  - … then combined to obtain the final prediction of the ensemble

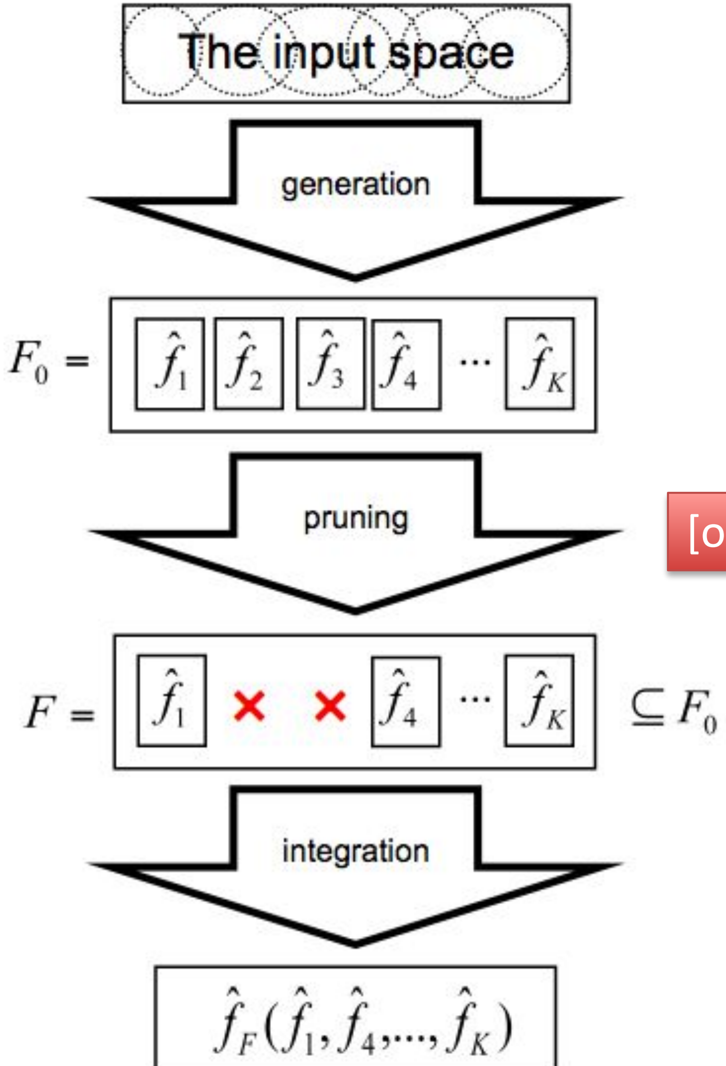
intuition

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

- aggregation of multiple learned models with the goal of improving model quality
  - e.g. expert panel in a human decision-making process
  - … or the popular concept of “the wisdom of the crowds”

ensemble learning process

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



[optional]

discussion

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

why should ensemble methods work?
[even better, when...?]

what's the catch?

pros & cons

+

- accuracy
  - majority compensates for individual errors
- diversity is key
  - individual models specialize in different areas of the data space
    - how?
  - ... but must be reasonably accurate
    - ... and by “reasonable” we mean...?

-

- complexity
  - understanding the global model
  - explaining decisions
  - computational
- remember Occam’s Razor
  - simplicity leads to greater accuracy
  - identifying the best model requires identifying the proper "model complexity"

gps

- introduction
- categories of methods
    - homogeneous
- popular methods
- issues

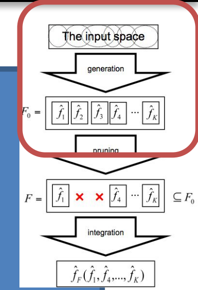
ensembles methods for...

- classification
- regression
- clustering
  - aka consensual clustering
- label ranking
- …
  - anything, really

our focus

types of ensembles: how to generate models

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

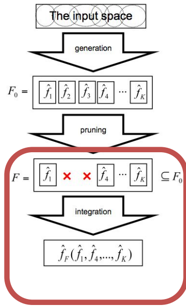


- homogeneous
  - single induction algorithm
- heterogeneous
  - multiple induction algorithms

our focus

where does diversity come from?

types of ensembles: how to combine models



regression

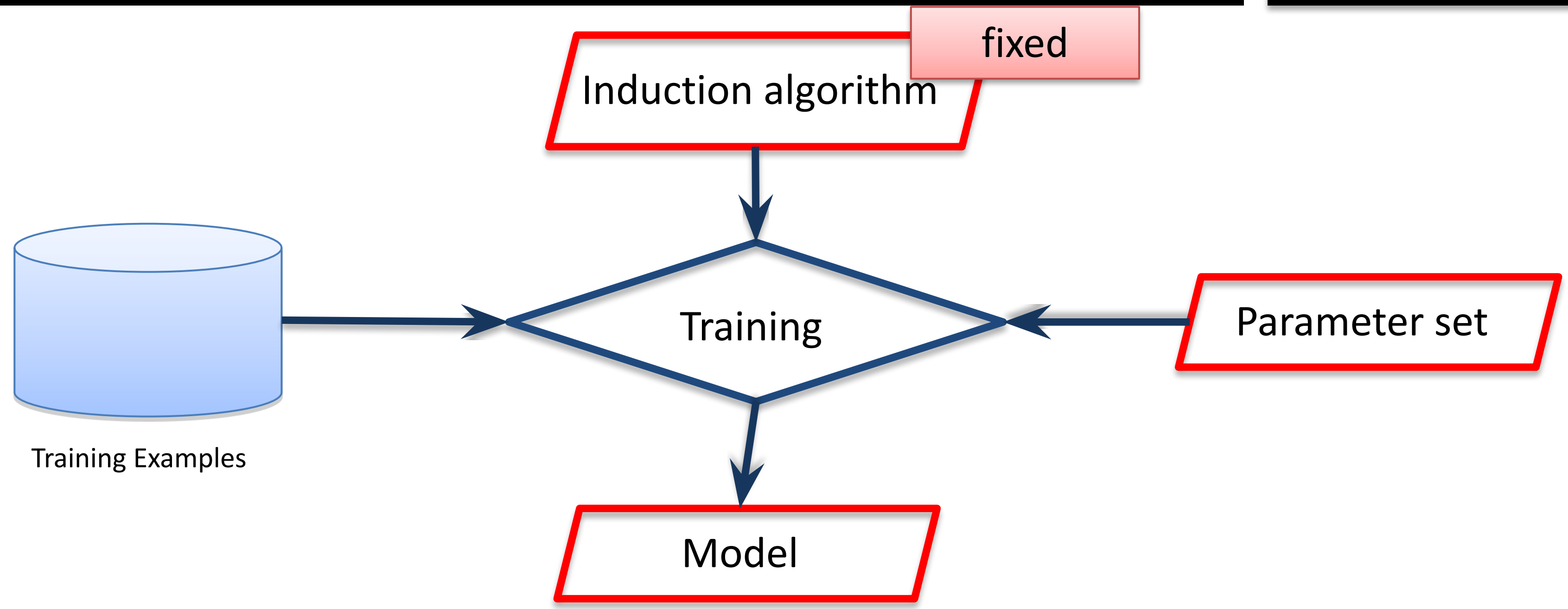
- **average**
- **weighted average**
- sum
- weighted sum
- product
- maximum
- minimum
- median

classification

- **majority voting**
- **weighted majority voting**
- borda count
  - base models rank candidates in order of preference
    - e.g. remember scoring?
  - points assigned to each position
  - prediction is class with more points

homogeneous ensembles: how to generate different models?

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

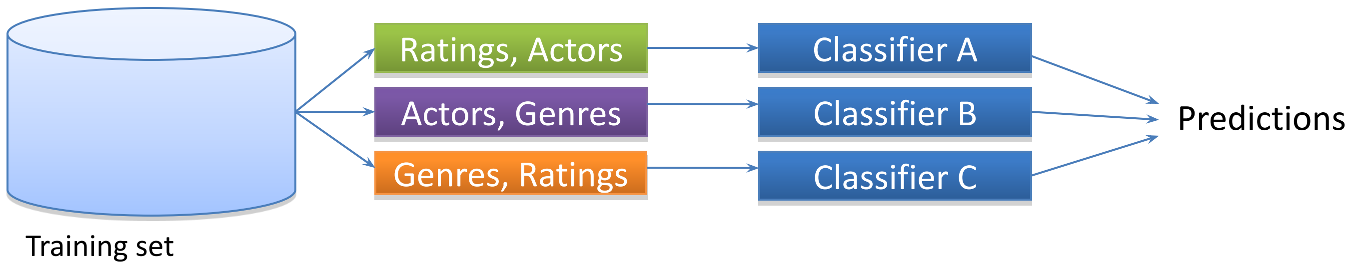


- Data manipulation
  - training set
- Modeling process manipulation
  - induction algorithm
  - parameter set
  - model
    - uncommon

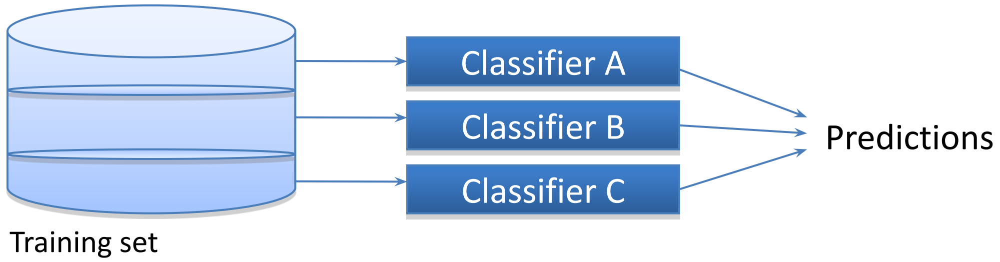
data manipulation

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Manipulating the input features



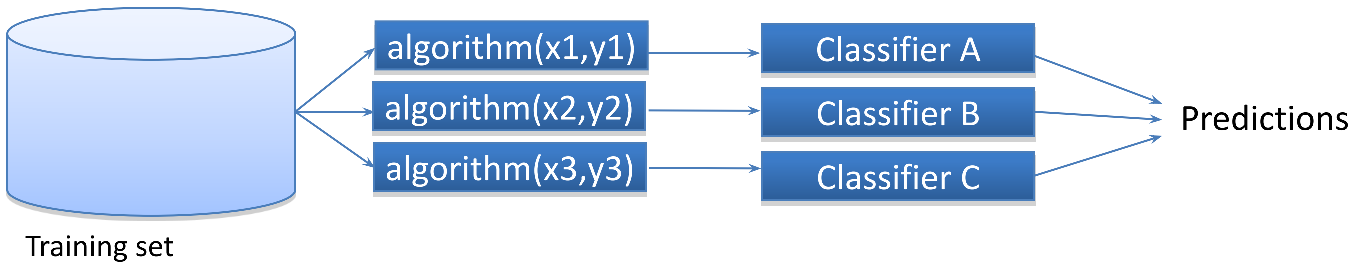
Sub-sampling from the training set



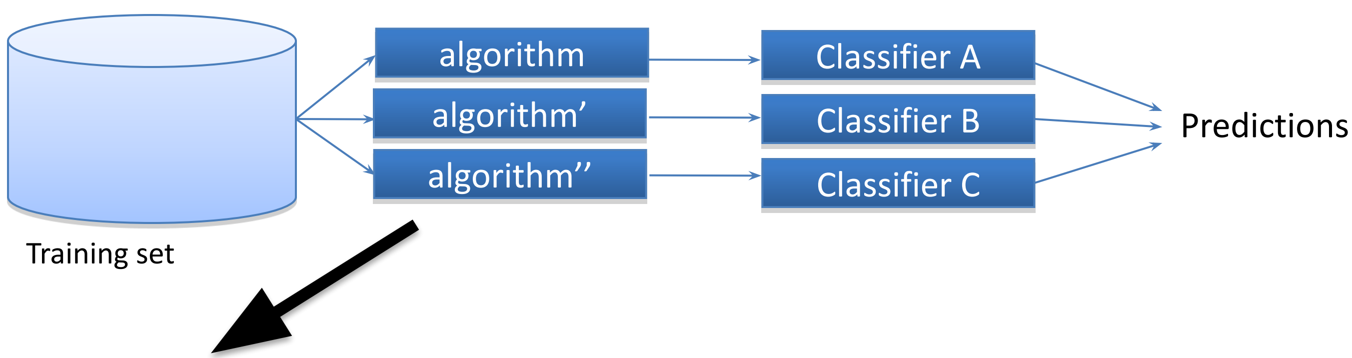
modeling process manipulation

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Manipulating the parameter sets



Manipulating the induction algorithm



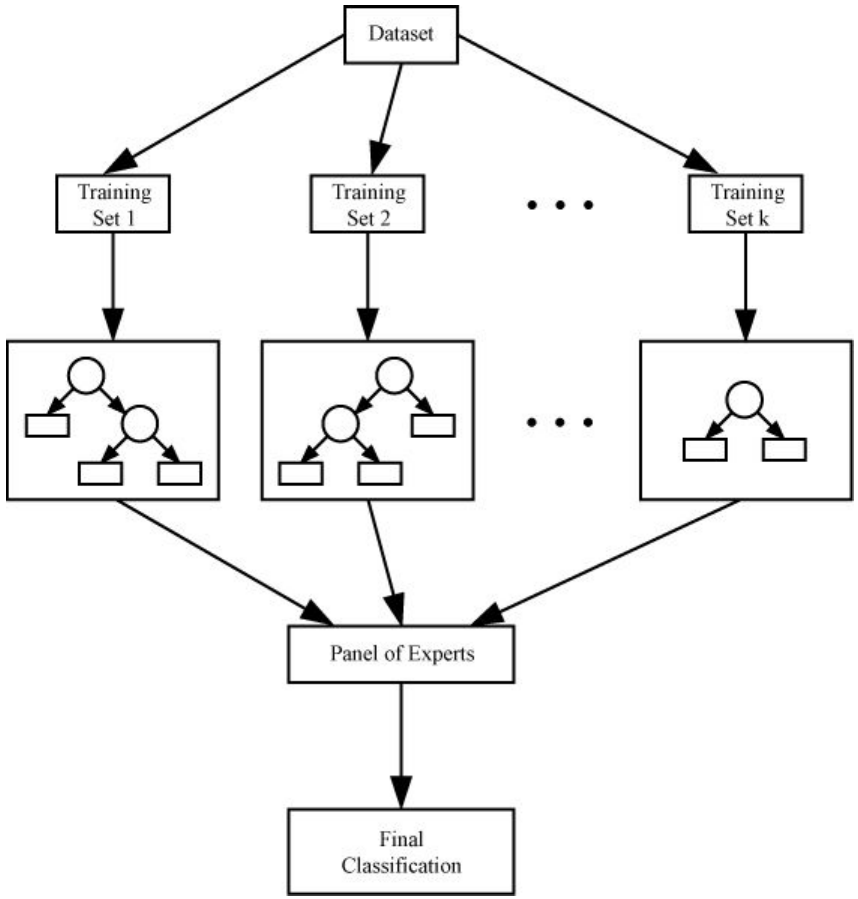
still homogeneous: *algorithm'* and *algorithm''* are variations of *algorithm*

gps

- introduction
- categories of methods
- popular methods
  - bagging
  - boosting
  - random forest
  - negative correlation
- issues

bagging: Bootstrap AGGregatING

- diagnosis analogy
  - diagnosis based on the majority vote of multiple doctors
  - trained in slightly different contexts
- training
  - given a set *D* of *d* tuples
  - at each iteration *i*
    - training set D_i of d tuples is sampled with replacement from *D*
      - i.e. bootstrap
    - model M_i is learned for training set D_i
- prediction
  - given an observation *X*
  - for each classifier M_i
    - make a prediction
  - an aggregation of the predictions is the prediction of the bagged model *M** for *X*



http://en.wikibooks.org/wiki/File:DTE_Bagging.png

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

bagging

- accuracy
  - often significantly better than a single classifier derived from *D*
  - robust to noise
- … if classifier is unstable!
  - unstable means a small change to the training data may lead to major decision changes
    - decision trees
    - neural networks

boosting

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

- training
  - equal weights are assigned to each training example
  - learn model M_1
  - learn additional k-1 models
    - give more weight to the examples that were incorrectly predicted by M_i
    - learn model M_{i+1}
- prediction
  - given an observation X
  - for each classifier M_i
    - make a prediction
  - an aggregation of the predictions is the prediction of the bagged model M^* for X
    - the weight of each classifier's vote is a function of its accuracy

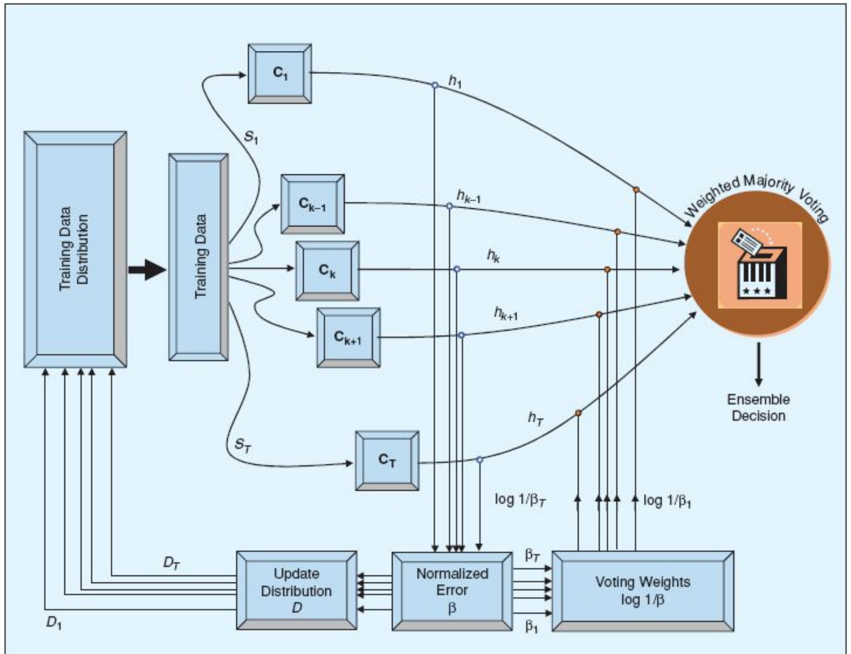


Figure 9. The AdaBoost.M1 algorithm.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

boosting: discussion

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

- boosting vs. bagging
  - differences
    - independent sampling vs. error-dependent sampling
    - uniform aggregation vs. weighted aggregation
- … so
  - boosting tends to achieve greater accuracy
  - … but it also risks overfitting the model to misclassified data

random forest

- training
  - learn k models
  - … with changed algorithm
    - at each split
      - randomly select a subset of the original features during the process of tree generation
- prediction
  - given an observation X
  - for each classifier M_i
    - make a prediction
  - an aggregation of the predictions is the prediction of the bagged model M^* for X

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.
https://doi.org/10.1023/A:1010933404324

random forest: discussion

- RF vs adaboost
  - comparable in accuracy
  - more robust to errors and
  - … outliers
- … vs bagging and adaboost
  - RF is insensitive to the number of attributes selected for consideration at each split and
  - faster

negative correlation learning

U. PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

- training
  - learn *k* models
  - … with changed algorithm
    - trained to minimize the error function of the ensemble
      - i.e., it adds to the error function a penalty term with the average error of the models already trained
- prediction
  - given an observation *X*
  - for each classifier M_i
    - make a prediction
  - an aggregation of the predictions is the prediction of the bagged model *M** for *X*

negative correlation learning: discussion

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

- only regression
  - algorithms that try to minimize/maximize a given objective function
    - e.g., neural networks, support vector regression
- models negatively correlated with the averaged error of the previously generated models

popular ensemble methods: summary

- bagging
  - base models: train algorithm on different bootstrap samples
  - prediction: average/majority
  - task: classification and regression
- boosting
  - base models: sequence of training processes, with more weight given to instances incorrectly classified by previous model
  - prediction: weighted vote
  - task: classification
- random forest
  - base models: train algorithm on different samples of attributes
  - prediction: average/majority
  - task: classification and regression
- negative correlation learning
  - base models: sequence of training processes, with new models forced to be more negatively correlated with the existing ones
  - prediction: average
  - task: regression

gps

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- introduction
- categories of methods
- popular methods
- issues

characteristics of the base models: classification

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

- base classifiers should be as accurate as possible and
  - although there is “the strength of weak classifiers”
    - R.E. Schapire. 1990. The Strength of Weak Learnability. *Mach. Learn.* 5, 2 (July 1990), 197-227
- having diverse errors
  - Brown, G. & Kuncheva, L., “Good” and “Bad” Diversity in Majority Vote Ensembles, *Multiple Classifier Systems, Springer,* **2010**, *5997*, 124-133

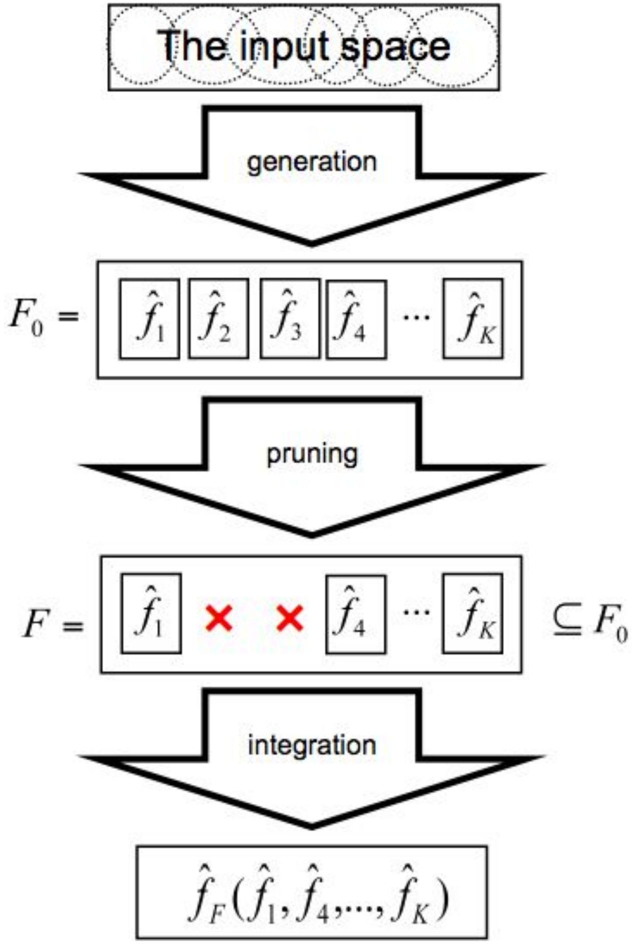
characteristics of the base models: regression

- more amenable to theoretical analysis
  - the error of an ensemble $\hat{f}_F$ with K base learners in relation to the true values given by f is:
    - $E(\hat{f}_F - f)^2 = \overline{bias}^2 + \frac{1}{K} \times \overline{var} + \left(1 - \frac{1}{K}\right) \times \overline{covar}$
    - ... assuming the integration function is the average
- the goal is to minimize $E(\hat{f}_F - f)^2$, so
  - the average bias of the base learners should be as small as possible
    - i.e. the base learners should be as accurate (on average) as possible
  - the average variance of the base learners should be as small as possible
    - i.e. the base learners should be as robust to small changes on the training data (on average) as possible
  - the average covariance of the base learners should be as low as possible
    - i.e. the base learners should have negative correlation

summary

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- combination of multiple models
  - majority compensates for individual errors
- individual models specialize in different areas of the data space
  - diversity is key
- today
  - focused on homogeneous
  - but essentially applicable to heterogeneous ensembles



Introductory References

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- *'Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations', Ian H. Witten and Eibe Frank, 1999*
- *'Data Mining: Practical Machine Learning Tools and Techniques second edition', Ian H. Witten and Eibe Frank, 2005*
- *Todd Holloway, 2008, "Ensemble Learning Better Predictions Through Diversity", power point presentation*
- *Leandro M. Almeida, "Sistemas Baseados em Comitês de Classificadores"*
- *Cong Li, 2009, "Machine Learning Basics 3. Ensemble Learning"*
- R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine, vol. 6, no. 3, pp. 21–45, Quarter 2006.*
- João Mendes-Moreira, Carlos Soares, Alípio Jorge, Jorge Freire de Sousa, "Ensemble approaches for regression: a survey", ACM Computing surveys, 45(1), article 10, 2012.

Core References

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- Wolpert, D. H., Stacked generalization, *Neural Networks,* **1992**, *5*, 241-259
- Breiman, L., Bagging predictors, *Machine Learning,* **1996**, *26*, 123-140
- Freund, Y. & Schapire, R., Experiments with a new boosting algorithm, *International Conference on Machine Learning,* **1996**, 148-156
- Breiman, L., Random forests, *Machine Learning,* **2001**, *45*, 5-32
- Liu, Y. & Yao, X., Ensemble learning via negative correlation, *Neural Networks,* **1999**, *12*, 1399-1404
- Rodríguez, J. J.; Kuncheva, L. I. & Alonso, C. J., Rotation forest: a new classifier ensemble, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **2006**, *28*, 1619-1630