Bagging and boosting

Séance « bb »

de l'UE « apprentissage automatique »

Bruno Bouzy

bruno.bouzy@parisdescartes.fr

www.mi.parisdescartes.fr/~bouzy

Outline

- Bootstrap
- Cross-validation
- Bagging
- Boosting
- Conclusion

Bootstrap

- Goal: estimation of a statistical stuff
  - Standard error
  - Confidence interval

- Interests
  - Any statistics without direct formula
  - Sampling is expensive or hard to get

Bootstrap

- Sample of size n
- Do B times
  - Random sample with replacement
  - Compute your statistics
- Average your statistics on the B iterations

Bootstrap

- References
  - Johnson 2001 an intro to the boostrap, Teaching statistics
  - Efron & Tibshirani 1986, boostraps methods... Statistical science vol 1 n 1 pages 54-77
  - Efron & Tibshirani 1993, an intro to the boostrap, book

Cross-Validation (CV)

- Predicting model to be tested

- Split the sample into a test set and a learning set

- Learning set
  - The set of examples on which the model is learnt
- Test set
  - The set of examples on which the model is tested

Cross-Validation (CV)

- K-fold cross-validation
- The sample is split into k subsets
- For each subset Sk do
  - Learn on subsets ≠ Sk
  - Test on Sk

- Complete CV
  - For any k < n perform a k-fold CV

Cross-Validation (CV)

- A reference:
  - Kohavi, A study of CV and boostrap for accuracy estimation and model selection, ijcai 1995

Bagging

- = Bootstrap + AGGreggatING

- Do B (=50) times
  - Random sample with replacement
  - Build a decision tree T_B
- class of an example:
  - class with the most votes given by the T_B.

Bagging

- References
  - Breiman, random forests, machine learning 2001, (fourth rank top cited paper).
  - Breiman, bagging predictors, machine learning 1996.

Boosting

- A general method for improving the accuracy of any given learning algorithm (called the weak learner).
- Background: the PAC model

Boosting

- Training set (x_i, y_i) with y_i=+-1
- T=1, 2, ..., T rounds
- $D_t(i)$ = weight of example i at time t

  (initialized with equal values)
- At time t, the learner learns a « weak hypothesis » h_t.
- $\varepsilon_t = \mathrm{Proba}_{i \text{ in } Dt}[h_t(x_i) \neq y_i] = \sum_{i:ht(xi)\neq yi} D_t(i)$

Boosting

- $\alpha_t = ½ \ln((1-\varepsilon_t)/\varepsilon_t)$
- Update:

  $D_{t+1}(i) = (D_t(i)/Z_t) \exp(-\alpha_t)$ if $h_t(x_i) = y_i$

  $= (D_t(i)/Z_t) \exp(+\alpha_t)$ if $h_t(x_i) \neq y_i$

  (Z_t normalisation factor over i)
- $H(x) = \text{sign}(\sum_{t=1,\ldots,T} \alpha_t h_t(x))$

Boosting

- References
  - Freund Shapire, a short intro to boosting, journal of japanese soc for ai, 1999.
  - Shapire, the strength of weak learnability, machine learning 1990.
  - Freund, boosting a weak learning algorithm by majority, info & computation, 1995.

Boosting

- Comments
  - Bad learner ($\varepsilon_t \approx 1$): focus on the successes
    - $h_t(x_i) = y_i$ then $D_{t+1}(i)$ increases
    - $h_t(x_i) \neq y_i$ then $D_{t+1}(i)$ decreases
  - Good learner ($\varepsilon_t \approx 0$): focus on the losses
    - $h_t(x_i) = y_i$ then $D_{t+1}(i)$ decreases
    - $h_t(x_i) \neq y_i$ then $D_{t+1}(i)$ increases

Boosting

Questions:

- What happens when $\varepsilon_t \approx 1/2$? Then $\alpha_t = 0$, and no update may occur? and the weak learner may return the same hyopthesis...

- What happens if the weak learner cannot manage D ? Drawing a sample with D and learn on it ?

Boosting

- Training error:

$$\prod_t [2\sqrt{(\varepsilon_t)(1-\varepsilon_t)}] = \prod_t \sqrt{1-4\gamma_t^2}) \leq \exp(-2\sum_t \gamma_t^2)$$

++ drops very fast

- Bound on the generalisation error:

$$Pr[H(x) \neq y] + O(\sqrt{Td/m})$$

d: VC-dimension of the weak hypothesis

m: size of the sample

T: number of iterations: overfitting ? Not in practice.

Summary

- Cross-validation
- Bootstrap
- Bagging =
  - Bootstrap + decision trees + AGGregatING
- Boosting