

Boosting and Bagging

Luís Alexandre

lfbaa@di.ubi.pt

NNIG

May 4, 2004



Voltar

Fechar

Bagging and boosting

- These methods create a set or ensemble of classifiers from a given data set.
- Each classifier is generated with a different training set obtained from the original using resampling techniques.
- The final output is obtained by voting.



Voltar

Fechar

What is a bootstrap sample ?

- Consider a data set D with m data points.
- A bootstrap sample D^i can be create from D by choosing m points from D randomly, with replacement.
- On average, 37% of the points in D will be in D^i .



Voltar

Fechar

Bagging

- Bagging (= **B**ootstrap **agg**regating) produces replications of the training set by sampling with replacement.
- Each replication of the training set has the same size as the original set, but some examples can appear more than once while others don't appear at all.
- A classifier is generated from each replication.
- All classifiers are used to classify each sample from the test set using a voting scheme.



Voltar

Fechar

Bagging

- The complete procedure is now described.
 - Consider a training set D with m cases.
 - Define the probability of the n th sample in the training set as $P(n) = 1/m$.
 - Now sample m times from the distribution $P(n)$.
 - Sample from D with replacement. This way a re-sampled training set D^i is built. D^i is normally called a bootstrap sample from D .
 - Repeat this procedure to construct a sequence of several independent bootstrap training sets.
 - Construct a corresponding sequence of classifiers by using the same classification algorithm applied to each of the bootstrap training sets.
 - For the final classification each one of these classifiers votes for each class [2].



Voltar

Fechar

When is bagging usefull

- Bagging should only be used if the learning machine is **unstable**.
- A learning machine is said to be unstable if a small change in the training set yields large variations in the classification.
- Examples of unstable learning machines: neural networks, decision trees.
- Example of a stable learning machine: k -nearest neighbors.



Voltar

Fechar

Bagging properties

- Improves the estimate if the learning algorithm is unstable.
- Reduces the variance of predictions without changing the bias.
- Degrades the estimate if the learning algorithm is stable.



Voltar

Fechar

What is boosting ?

- It is a family of methods for accelerating a learning algorithm:
 - AdaBoost (two-class problem) [4]
 - AdaBoost.M1 and M2 (multiple-class problem) [4]
 - AdaBoostR (regression) [3]
- The idea is to boost a **weak learning algorithm** into a strong learning algorithm.
- A weak learning algorithm can be, i.e., inaccurate rules of thumb that are slightly better than random guessing.
- It produces several classifiers.
- The training set chosen at a given time depends on the performance of earlier classifiers.
- Harder to classify points are chosen more often than easier points so that the algorithm is concentrating on the most difficult points.



Voltar

Fechar

AdaBoost

- Consider a training set with m points. At each repetition or trial $t = 1, 2, \dots, T$, point i has weight w_i^t , where $w_i^1 = 1/m$ for all i .
- For each trial do:
 - construct classifier g^t from the training instances using weights w_i^t ;
 - the error rate of g^t on the training data (ϵ^t) is the sum of the weights w_i^t of the misclassified instances;
 - if $\epsilon^t = 0$ or $\epsilon^t \geq 1/2$ the process terminates;
 - otherwise the weights w_i^{t+1} for the next trial are set so that the error rate of g^t on the training set using the weights w_i^{t+1} is exactly $1/2$.
- The final composite classifier g^* is built combining the g^t classifiers using voting.



Comments on AdaBoost

- Freund and Schapire proved that the error rate of g^* on the unweighted training instances approaches zero exponentially quickly as T increases [6].
- In [5] the author concludes that much of the benefits of boosting are caused by over-fitting the training data set. He also says that although boosting generally increases accuracy, it leads to a deterioration on some data sets.
- Breiman [1] notes that boosting forces the classifier to have a 50% error rate on the re-weighted training instances which causes the learning system to produce a quite different classifier on the following trial. This leads to an extensive exploration of the classifier space.
- Quinlan [6] did some experiments with boosting C4.5 (a decision tree [7]) and found that the advantages of boosting are lost if there is non-trivial classification noise in the learning sets. This is unexpected since as boosting uses a combination of classifiers it should be robust in the



Volar

Fechar

presence of noise. He also believes that the reason why boosting increases the classification accuracy may be unrelated to its convergence properties on the training data.

- Breiman [1] also says that boosting failure happens more likely with relatively small data sets.



Voltar

Fechar

Referências

- [1] L. Breiman. Arcing classifiers. Technical Report 460, Statistics Department, University of California, Berkeley, 1996.
- [2] L. Breiman. Bagging predictors. *Machine Learning*, 26(2):123–140, 1996.
- [3] H. Drucker. Improving regressors using boosting techniques. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 107–115. Morgan-Kaufmann, 1997.
- [4] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.
- [5] J.R. Quinlan. Bagging, Boosting, and C4.5. In *Proceedings AAAI-96 Fourteenth National Conference on Artificial Intelligence*, Portland, OR, 1996.



Voltar

Fechar

- [6] J.R. Quinlan. Boosting first-order learning. In *Algorithmic Learning Theory, 7th International Workshop, ALT'96*, Sydney, Australia, 1996. Lecture Notes in Artificial Intelligence 1160, Springer-Verlag.
- [7] R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Inc., 1993.

