Boosting: AdaBoost

Hunter Glanz

OUTLINE

Foundational Machine Learning

- You've learned about:
  - Traditional Regression
  - Logistic Regression
  - K-Nearest Neighbors
  - Discriminant Analysis
  - Support Vector Machines
  - Tree-Based Methods

Foundational Machine Learning

- You've learned about:
  - Traditional Regression
  - Logistic Regression
  - K-Nearest Neighbors
  - Discriminant Analysis
  - Support Vector Machines
  - Tree-Based Methods

Remember *there's no free lunch*!

Ensemble Learning Strategies

- Ensemble learning refers to algorithms that combine the predictions from two or more models:
  - Let's team up!

Ensemble Learning Strategies

- Ensemble learning refers to algorithms that combine the predictions from two or more models:
  - Let's team up!
  - Near infinite number of ways to do this so we'll talk generally about three broad strategies:
    1. Bagging
    2. Stacking
    3. Boosting

Ensemble Learning Strategies

- Ensemble learning refers to algorithms that combine the predictions from two or more models:
  - Let's team up!
  - Near infinite number of ways to do this so we'll talk generally about three broad strategies:
    1. Bagging
    2. Stacking
    3. Boosting

Today we will focus on **AdaBoost**

Motivation

A procedure that combines the outputs of many "weak" learners to produce a powerful "committee."

Motivation

A procedure that combines the outputs of many "weak" learners to produce a powerful "committee."

Most people cite **Adaptive Boosting (AdaBoost)** by Freund and Schapire (1997) as the big emergence of boosting.

AdaBoost.M1

- Consider a two-class problem, with the output variable coded as -1 and 1, and a single vector of predictor variables, X.

AdaBoost.M1

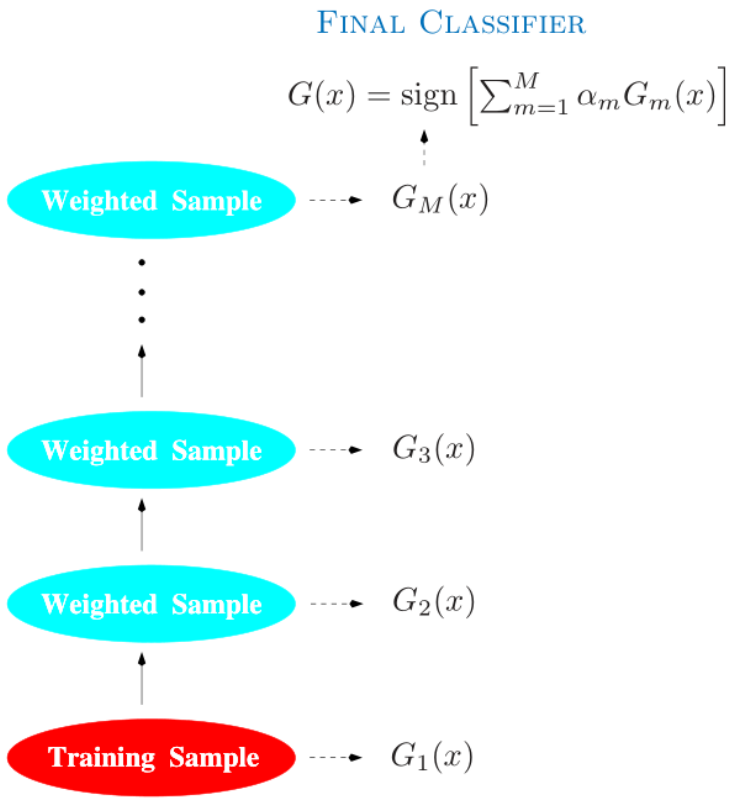
- Consider a two-class problem, with the output variable coded as -1 and 1, and a single vector of predictor variables, X.
- A weak classifier is one whose error rate is only slightly better than random guessing (i.e. a coin flip).

AdaBoost.M1

- Consider a two-class problem, with the output variable coded as -1 and 1, and a single vector of predictor variables, X.
- A weak classifier is one whose error rate is only slightly better than random guessing (i.e. a coin flip).

Sequentially apply the weak classification algorithm to repeatedly modified versions of the data.

The AdaBoost.M1 Visual



Two Sets of Weights

$$G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$$

- The α_m:

Two Sets of Weights

$$G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$$

- The α_m:
  - Computed by the boosting algorithm
  - Weight the contribution of each respective $G_m(x)$
  - **Giver higher influence to the more accurate classifiers in the sequence**

Two Sets of Weights

$$G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$$

- The α_m:
  - Computed by the boosting algorithm
  - Weight the contribution of each respective $G_m(x)$
  - **Giver higher influence to the more accurate classifiers in the sequence**
- The weighted samples/data (w_i):

Two Sets of Weights

$$G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$$

- The α_m:
  - Computed by the boosting algorithm
  - Weight the contribution of each respective $G_m(x)$
  - **Giver higher influence to the more accurate classifiers in the sequence**
- The weighted samples/data (w_i):
  - Weights applied to each of the training observations
  - Initially all set to $w_i = 1/N$
  - **Observations that are misclassified have their weights increased, whereas weights are decreased for those correctly classified**

AdaBoost Notes

- When and How to use?!

AdaBoost Notes

- When and How to use?!
  - You can technically **boost** most, if not all, machine learning algorithms!
  - Implementations have converged on some very popular and successful versions

AdaBoost Notes

- When and How to use?!
  - You can technically **boost** most, if not all, machine learning algorithms!
  - Implementations have converged on some very popular and successful versions
- Elements of Statistical Learning example (pg. 339, not PDF page):
  - Each weak learner is a "stump": two-terminal node classification tree
  - Boosting reduces the prediction error rate by almost a factor of four
  - It outperforms a single large classification tree

Small trees are popular choices for the weak learners