An Introduction to Ensemble and Boosting Methods

Amir Saffari

Institute for Computer Graphics and Vision (ICG)
Graz University of Technology, Austria
http://www.ymer.org/amir/
saffari@icg.tugraz.at, amir@ymer.org

PASCAL Bootcamp 2007
Vilanova i la Geltrú, Spain
Outline

Ensemble Methods
  Introduction
  Model Averaging
  Bagging

Stagewise Additive Modeling
  Stagewise Additive Modeling
  Boosting
  Practical Example
Outline

**Ensemble Methods**
- Introduction
- Model Averaging
- Bagging

**Stagewise Additive Modeling**
- Stagewise Additive Modeling
- Boosting
- Practical Example
Choosing your operating system
Majority voting scheme

- Ask experts for their opinion and choose the option with majority vote.
- Let’s say we have a set of M experts:
  \[ H = \{ f_1, f_2, \ldots, f_M \}, \quad f_m(budget) \in \{ \text{Linux}, \text{Windows} \} \]
- Assume Linux = +1, Windows = −1, then the majority vote decision will be:
  \[ F(budget) = \text{sign}(\frac{1}{M} \sum_{m=1}^{M} f_m(budget)) \]
- This is the main concept behind ensemble methods.
- Diversity is just more than great.
Majority voting scheme

- Ask experts for their opinion and choose the option with majority vote.
- Let’s say we have a set of $M$ experts: $H = \{f_1, f_2, \ldots, f_M\}$, $f_m(budget) \in \{\text{Linux, Windows}\}$
- Assume \text{Linux} = +1, \text{Windows} = -1, then the majority vote decision will be:
  
  $F(budget) = \text{sign}(\frac{1}{M} \sum_{m=1}^{M} f_m(budget))$

- This is the main concept behind ensemble methods.
- \text{Diversity} is just more than great.
Majority voting scheme

- Ask experts for their opinion and choose the option with majority vote.
- Let's say we have a set of $M$ experts:
  \[ H = \{f_1, f_2, \ldots, f_M\}, \quad f_m(budget) \in \{Linux, Windows\} \]
- Assume $Linux = +1$, $Windows = -1$, then the majority vote decision will be:
  \[ F(budget) = \text{sign}(\frac{1}{M} \sum_{m=1}^{M} f_m(budget)) \]
- This is the main concept behind ensemble methods.
- Diversity is just more than great.
Majority voting scheme

- Ask experts for their opinion and choose the option with majority vote.
- Let’s say we have a set of $M$ experts: 
  \[ H = \{f_1, f_2, \ldots, f_M\}, \quad f_m(budget) \in \{\text{Linux}, \text{Windows}\} \]
- Assume Linux = +1, Windows = −1, then the majority vote decision will be:
  \[ F(budget) = \text{sign}\left(\frac{1}{M} \sum_{m=1}^{M} f_m(budget)\right) \]
- This is the main concept behind ensemble methods.
- Diversity is just more than great.
Majority voting scheme

- Ask experts for their opinion and choose the option with majority vote.
- Let’s say we have a set of $M$ experts:
  $H = \{f_1, f_2, \ldots, f_M\}, \quad f_m(budget) \in \{Linux, Windows\}$
- Assume $Linux = +1$, $Windows = -1$, then the majority vote decision will be:
  $F(budget) = \text{sign}(\frac{1}{M} \sum_{m=1}^{M} f_m(budget))$
- This is the main concept behind ensemble methods.
- Diversity is just more than great.
Notations

- $D = \{(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)\}$
- $x_n \in \mathbb{R}^d, t_n \in \{-1, +1\}$
- $H = \{f_1(x), f_2(x), \ldots, f_M(x)\}$
- $y_m = f_m(x) \in \{-1, +1\}$
- $F(x) = \sum_{m=1}^{M} \alpha_m f_m(x)$
- $\alpha_m \in \mathbb{R}^+, \sum_{m=1}^{M} \alpha_m = 1$
Why to use ensemble methods?

Better performance
Assume that: $\forall j : p(y_m \neq t) \leq \mu < 1/2$, and the decisions of different models are independent, then the chance of a wrong decision by the ensemble, $p(F \neq t) = 1 - Pr(k \leq M/2)$, where $Pr(k \leq K)$ is the cumulative distribution function of a binomial distribution.
This upper bound is pretty much better than the original error rate.
Performance of ensemble of classifiers

For $\mu = 0.3$ and $M = 21$, the chance of misclassification is around 0.026 (T. G. Diettrich 2000).
Why to use ensemble methods?

Statistical reason

Why to use ensemble methods?

Computational reason

Why to use ensemble methods?

Representational reason

Why to use ensemble methods?

- **Computational efficiency** We are looking for a set of weak learners (classifiers, or hypotheses): \( p(y \neq t) < 1/2 \).
- **Different classes of base models** Choices could be: Trees (stumps, small, large), Naive Bayes, k-Nearest Neighbors, Neural Networks, Linear SVM, YOUR-MAGICAL-MODEL, ...
Why to use ensemble methods?

- **Computational efficiency** We are looking for a set of weak learners (classifiers, or hypotheses): $p(y \neq t) < 1/2$.
- **Different classes of base models** Choices could be: Trees (stumps, small, large), Naive Bayes, k-Nearest Neighbors, Neural Networks, Linear SVM, YOUR-MAGICAL-MODEL, ...
How to find the base models?

- Train a diverse set of models on the same datasets.
- Train a set of models from a specific class of learners by using diversity in the datasets, parameters, or initial conditions.
- Cross-validated committees
- Bagging
- Boosting
How to find the base models?

- Train a diverse set of models on the same datasets.
- Train a set of models from a specific class of learners by using diversity in the datasets, parameters, or initial conditions.
- Cross-validated committees
- Bagging
- Boosting
How to find the base models?

- Train a diverse set of models on the same datasets.
- Train a set of models from a specific class of learners by using diversity in the datasets, parameters, or initial conditions.
- Cross-validated committees
  - Bagging
  - Boosting
How to find the base models?

- Train a diverse set of models on the same datasets.
- Train a set of models from a specific class of learners by using diversity in the datasets, parameters, or initial conditions.
- Cross-validated committees
- Bagging
- Boosting
Outline

**Ensemble Methods**
- Introduction
- Model Averaging
- Bagging

**Stagewise Additive Modeling**
- Stagewise Additive Modeling
- Boosting
- Practical Example
Bagging

- Create subsets of the training samples, called bootstrap replicates, each containing examples drawn randomly with replacement from the original training dataset, and train learning algorithms over them.

- The method is called bootstrap aggregation.

- Originally developed to reduce the variance of the learning algorithms.

Bagging

- Create subsets of the training samples, called bootstrap replicates, each containing examples drawn randomly with replacement from the original training dataset, and train learning algorithms over them.

- The method is called bootstrap aggregation.

- Originally developed to reduce the variance of the learning algorithms.

Bagging

- Create subsets of the training samples, called bootstrap replicates, each containing examples drawn randomly with replacement from the original training dataset, and train learning algorithms over them.
- The method is called bootstrap aggregation.
- Originally developed to reduce the variance of the learning algorithms.

Outline

Ensemble Methods
  Introduction
  Model Averaging
  Bagging

Stagewise Additive Modeling
  Stagewise Additive Modeling
  Boosting
  Practical Example

Amir Saffari

An Introduction to Ensemble and Boosting Methods
Stagewise additive modeling

\[ F(x) = \sum_{m=1}^{M} \alpha_m f_m(x) \]

General Forward Stagewise Additive Modeling

- Set \( F^{(0)}(x) = 0 \)
- for \( m = 1 \) to \( M \), do
  - \( \{f_m(x), \alpha_m\} = \underset{f,\alpha}{\text{argmin}} \sum_{n=1}^{N} L(t_n, F^{(m-1)}(x_n) + \alpha f(x_n)) \)
  - \( F^{(m)}(x) = F^{(m-1)}(x) + \alpha_m f_m(x) \)

Stagewise additive modeling

\[ F(x) = \sum_{m=1}^{M} \alpha_m f_m(x) \]

General Forward Stagewise Additive Modeling

- Set \( F^{(0)}(x) = 0 \)
- for \( m = 1 \) to \( M \), do
  - \( \{f_m(x), \alpha_m\} = \arg\min_{f,\alpha} \sum_{n=1}^{N} L(t_n, F^{(m-1)}(x_n) + \alpha f(x_n)) \)
  - \( F^{(m)}(x) = F^{(m-1)}(x) + \alpha_m f_m(x) \)

Stagewise additive modeling

\[ F(x) = \sum_{m=1}^{M} \alpha_m f_m(x) \]

General Forward Stagewise Additive Modeling

- Set \( F^{(0)}(x) = 0 \)
- for \( m = 1 \) to \( M \), do
  - \( \{ f_m(x), \alpha_m \} = \arg\min_{f,\alpha} \sum_{n=1}^{N} L(t_n, F^{(m-1)}(x_n) + \alpha f(x_n)) \)
  - \( F^{(m)}(x) = F^{(m-1)}(x) + \alpha_m f_m(x) \)

Stagewise additive modeling

\[ F(x) = \sum_{m=1}^{M} \alpha_m f_m(x) \]

General Forward Stagewise Additive Modeling

- Set \( F^{(0)}(x) = 0 \)
- for \( m = 1 \) to \( M \), do
  - \( \{ f_m(x), \alpha_m \} = \text{argmin}_{f, \alpha} \sum_{n=1}^{N} L(t_n, F^{(m-1)}(x_n) + \alpha f(x_n)) \)
  - \( F^{(m)}(x) = F^{(m-1)}(x) + \alpha_m f_m(x) \)

Outline

Ensemble Methods
  Introduction
  Model Averaging
  Bagging

Stagewise Additive Modeling
  Stagewise Additive Modeling
  Boosting
  Practical Example
AdaBoost

\[ F(x) = \sum_{m=1}^{M} \alpha_m f_m(x) \]
\[ l(t, y) = -t \cdot y \]

Discrete AdaBoost

1. Set \( W = \{ w_1, w_2, \ldots, w_N \}, \forall n : w_n = 1/N \)
2. for \( m = 1 \) to \( M \), do
   1. \( f_m(x) = \arg\min_f \sum_{n=1}^{N} w_n(t_n - f(x_n))^2 \)
   2. \( e_m = \sum_{n=1}^{N} w_n l(t_n, f_m(x_n)) \)
   3. \( \alpha_m = \log \frac{1-e_m}{e_m} \)
   4. \( w_n \leftarrow w_n \exp(\alpha_m l(t_n, f_m(x_n))) \)
   5. \( w_n \leftarrow \sum_{n=1}^{N} w_n \)

AdaBoost

$$F(x) = \sum_{m=1}^{M} \alpha_m f_m(x)$$
$$l(t, y) = -t \cdot y$$

Discrete AdaBoost

- Set \( W = \{w_1, w_2, \ldots, w_N\}, \forall n : w_n = 1/N \)
- for \( m = 1 \) to \( M \), do
  - \( f_m(x) = \arg\min_f \sum_{n=1}^{N} w_n (t_n - f(x_n))^2 \)
  - \( e_m = \sum_{n=1}^{N} w_n l(t_n, f_m(x_n)) \)
  - \( \alpha_m = \log \frac{1-e_m}{e_m} \)
  - \( w_n \leftarrow w_n \exp(\alpha_m l(t_n, f_m(x_n))) \)
  - \( w_n \leftarrow \sum_{n=1}^{N} w_n \)

AdaBoost

\[ F(x) = \sum_{m=1}^{M} \alpha_m f_m(x) \]
\[ l(t, y) = -t \cdot y \]

**Discrete AdaBoost**

- Set \( W = \{ w_1, w_2, \ldots, w_N \}, \forall n : w_n = 1/N \)
- for \( m = 1 \) to \( M \), do
  - \( f_m(x) = \arg\min_f \sum_{n=1}^{N} w_n (t_n - f(x_n))^2 \)
  - \( e_m = \sum_{n=1}^{N} w_n l(t_n, f_m(x_n)) \)
  - \( \alpha_m = \log \frac{1 - e_m}{e_m} \)
  - \( w_n \leftarrow w_n \exp(\alpha_m l(t_n, f_m(x_n))) \)
  - \( w_n \leftarrow \sum_{n=1}^{N} w_n \)

AdaBoost

\[ F(x) = \sum_{m=1}^{M} \alpha_m f_m(x) \]

\[ l(t, y) = -t \cdot y \]

Discrete AdaBoost

- Set \( W = \{ w_1, w_2, \ldots, w_N \} \), \( \forall n : w_n = 1/N \)
- for \( m = 1 \) to \( M \), do
  - \( f_m(x) = \text{argmin}_f \sum_{n=1}^{N} w_n(t_n - f(x_n))^2 \)
  - \( e_m = \sum_{n=1}^{N} w_n l(t_n, f_m(x_n)) \)
  - \( \alpha_m = \log \frac{1-e_m}{e_m} \)
  - \( w_n \leftarrow w_n \exp(\alpha_m l(t_n, f_m(x_n))) \)
  - \( w_n \leftarrow \sum_{n=1}^{N} w_n \)

AdaBoost

\[ F(x) = \sum_{m=1}^{M} \alpha_m f_m(x) \]
\[ I(t, y) = -t \cdot y \]

Discrete AdaBoost

- Set \( W = \{ w_1, w_2, \ldots, w_N \} \), \( \forall n : w_n = 1 / N \)
- for \( m = 1 \) to \( M \), do
  - \( f_m(x) = \arg \min_f \sum_{n=1}^{N} w_n (t_n - f(x_n))^2 \)
  - \( e_m = \sum_{n=1}^{N} w_n I(t_n, f_m(x_n)) \)
  - \( \alpha_m = \log \frac{1-e_m}{e_m} \)
  - \( w_n \leftarrow w_n \exp(\alpha_m I(t_n, f_m(x_n))) \)
  - \( w_n \leftarrow \sum_{n=1}^{N} w_n \)

AdaBoost

\[ F(x) = \sum_{m=1}^{M} \alpha_m f_m(x) \]

\[ l(t, y) = -t \cdot y \]

Discrete AdaBoost

- Set \( W = \{ w_1, w_2, \ldots, w_N \}, \forall n : w_n = 1/N \)
- for \( m = 1 \) to \( M \), do
  - \( f_m(x) = \arg \min_f \sum_{n=1}^{N} w_n (t_n - f(x_n))^2 \)
  - \( e_m = \sum_{n=1}^{N} w_n l(t_n, f_m(x_n)) \)
  - \( \alpha_m = \log \frac{1-e_m}{e_m} \)
  - \( w_n \leftarrow w_n \exp(\alpha_m l(t_n, f_m(x_n))) \)
  - \( w_n \leftarrow \sum_{n=1}^{N} w_n \)


Amir Saffari
An Introduction to Ensemble and Boosting Methods
Outline

Ensemble Methods
- Introduction
- Model Averaging
- Bagging

Stagewise Additive Modeling
- Stagewise Additive Modeling
- Boosting
- Practical Example
Tracking visual objects

Tracking as binary classification


Tracking visual objects

Tracking visual objects

- Tracking as binary classification problem

- Object and background changes are robustly handled by on-line updating!