COMMITTEE MACHINES

CS6140

Predrag Radivojac
Khoury College of Computer Sciences
Northeastern University

Fall, 2019
Preliminaries

Given: a set of observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$

Objective: learn the posterior $p(y|x, \mathcal{D})$

Optimal Bayes Model:

$$p(y|x, \mathcal{D}) = \sum_{f \in \mathcal{F}} p(y|x, \mathcal{D}, f)p(f|x, \mathcal{D})$$

Finite $\mathcal{F}$

Idea:

Don’t just train a single model, train multiple models.

Average the outputs in some way.
Two Ways of Averaging

Static structures:
  pre-trained models, applied independently of \( x \)
  examples: bagging, boosting, random forests

Dynamic structures:
  pre-trained models, applied depending on \( x \)
  examples: mixture of experts
**STATIC STRUCTURE**

\[ D \]

\[ D_1 \rightarrow f_1(x) \]

\[ D_2 \rightarrow f_2(x) \]

\[ \vdots \]

\[ D_M \rightarrow f_M(x) \]

\[ x \]

\[ f(x) \]

One way to average:

\[ f(x) = \frac{1}{M} \sum_{i=1}^{M} f_i(x) \]
**WHY SHOULD IT WORK?**

**Given:** \( \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}, \ x_i \in \mathcal{X}, y_i \in \mathcal{Y} \)

**Idea:**

Let \( f_A(x) = \mathbb{E}[f(x, D)] \) be an “averaged” model. Then, for a fixed \( x \)

\[
\mathbb{E}[(y - f(x, D))^2] = y^2 - 2y\mathbb{E}[f(x, D)] + \mathbb{E}[f^2(x, D)]
\]

\[
\mathbb{E}[(y - f(x, D))^2] \geq (y - f_A(x))^2
\]

Jensen’s inequality

\[
\mathbb{E}[Z^2] \geq \mathbb{E}^2[Z]
\]
Bagging: Bootstrap aggregating

Approach:

Create $B$ bootstrap samples $\mathcal{D}_b$, where $b = 1, 2, \ldots, B$

Train a model $f_b(x)$ for each $\mathcal{D}_b$

Let final decision be $f(x) = \frac{1}{B} \sum_{b=1}^{B} f_b(x)$

Averaging:

Average soft predictions then threshold the model.

Use majority vote (original idea by Leo Breiman).
**Boosting**

**Idea:** Average weak models to create strong models.

Models trained on different distribution.

**Two approaches:**

- Boosting by filtering.
- Boosting by subsampling or reweighting.


Boosting by Filtering

Idea: Construct a committee of 3 experts.

1) Expert $f_1$ is trained on $n_1$ examples picked randomly

2) Expert $f_2$ is trained as follows

   2.1. Flip a coin

       Heads $\rightarrow$ pass examples through $f_1$ until one is misclassified

       include that example in training set for $f_2$

       Tails $\rightarrow$ pass examples through $f_1$ until one is correctly classified

       include that example in training set for $f_2$

   2.2. Continue until $n_1$ examples are collected, then train $f_2$
**Boosting by Filtering**

**Idea:** Construct a committee of 3 experts.

3) Expert $f_3$ is trained as follows

3.1. Select $n_1$ examples by keeping those where $f_1$ and $f_2$ disagree

Data distributions are different for training $f_1$, $f_2$, and $f_3$

The final prediction obtained by majority voting.

It can be proved that if $f_1$, $f_2$, $f_3$ all have error rate of $\epsilon < \frac{1}{2}$

then the committee machine has error $g(\epsilon) = 3\epsilon^2 - 2\epsilon^3$

**Idea:** Repeat the process recursively.
Algorithm 1 AdaBoost algorithm. Typically, $T = 100$.

Input:
\[ \mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \], where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$.
Weak learning algorithm $a$ that maps $\mathcal{X}$ to $\mathcal{Y}$
Positive integer $T$

Initialization:
Initialize sampling distribution $p^{(1)}(i) = \frac{1}{n}$ for $\forall i \in \{1, 2, \ldots, n\}$

Loop:
for $t = 1$ to $T$
\begin{itemize}
  \item Sample data set $\mathcal{D}^{(t)}$ from $\mathcal{D}$ according to $p^{(t)}(i)$
  \item Learn model $f_t(x)$ from $\mathcal{D}^{(t)}$
  \item Calculate error $\epsilon_t$ on training data $\mathcal{D}$ as $\epsilon_t = \sum_{i: f_t(x_i) \neq y_i} p^{(t)}(i)$
  \item Set $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$
  \item Set $w_t = \ln \frac{1}{\beta_t}$
  \item Set $p^{(t+1)}(i) = \frac{p^{(t)}(i)}{Z} \cdot \begin{cases} 
    \beta_t & \text{if } f_t(x_i) = y_i \\
    1 & \text{otherwise}
  \end{cases}$, where $Z$ is a normalizer
\end{itemize}
end

Output:
\[ f(x) = \arg \max_{y \in \mathcal{Y}} \left( \sum_{t=1}^{T} w_t \cdot I(f_t(x) = y) \right) \]
**AdaBoost**

**Theorem.** Assume $\epsilon_t < \frac{1}{2}$ and let $\gamma_t = \frac{1}{2} - \epsilon_t$. Then the following bound holds

$$\frac{1}{n} |\{i : f(x_i) \neq y_i\}| \leq \prod_{t=1}^{T} \sqrt{1 - 4\gamma_t^2} \leq e^{-2 \sum_{t=1}^{T} \gamma_t^2}$$

**In practice:**

Training error $\to 0$, but the test error continues to decrease

AdaBoost theory shows relationship to SVMs

Minimizing $\epsilon_t$ is equivalent to minimizing $E = \sum_{i=1}^{n} e^{-y_i f(x_i)}$
RANDOM FORESTS

Idea. Construct an ensemble of trees, 100 to 1000.

In practice:

Randomize dataset

Bootstrap the dataset

Randomize features upon which we split

At each node, keep e.g. \( F = \log_2 d + 1 \) features, then split