Machine Learning & Data Mining
CS/CNS/EE 155

Lecture 6:
Boosting & Ensemble Selection
Today

• High Level Overview of Ensemble Methods

• Boosting
  – Ensemble Method for Reducing Bias

• Ensemble Selection
Recall: Test Error

• “True” distribution: \( P(x, y) \)
  – Unknown to us

• Train: \( h_S(x) = y \)
  – Using training data: \( S = \{(x_i, y_i)\}_{i=1}^N \)
  – Sampled from \( P(x, y) \)

• Test Error:
  \[
  L_P(h_S) = E_{(x, y) \sim P(x, y)} \left[ L(y, h_S(x)) \right]
  \]

• Overfitting: Test Error >> Training Error
### True Distribution $P(x,y)$

<table>
<thead>
<tr>
<th>Person</th>
<th>Age</th>
<th>Male?</th>
<th>Height &gt; 55”</th>
</tr>
</thead>
<tbody>
<tr>
<td>James</td>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Jessica</td>
<td>14</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Alice</td>
<td>14</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Amy</td>
<td>12</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Bob</td>
<td>10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Xavier</td>
<td>9</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Cathy</td>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Carol</td>
<td>13</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Eugene</td>
<td>13</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Rafael</td>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Dave</td>
<td>8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Peter</td>
<td>9</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Henry</td>
<td>13</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Erin</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Rose</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Iain</td>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Paulo</td>
<td>12</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Margaret</td>
<td>10</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Frank</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Jill</td>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Leon</td>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Sarah</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gena</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Patrick</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### Training Set $S$

<table>
<thead>
<tr>
<th>Person</th>
<th>Age</th>
<th>Male?</th>
<th>Height &gt; 55”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>14</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Bob</td>
<td>10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Carol</td>
<td>13</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Dave</td>
<td>8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Erin</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Frank</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Gena</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Test Error:

$$\mathcal{L}(h) = E_{(x,y) \sim P(x,y)}[L(h(x), y)]$$
Recall: Test Error

• Test Error:

\[ L_P(h) = \mathbb{E}_{(x,y) \sim P(x,y)} [L(y, h(x))] \]

• Treat \( h_S \) as random variable:

\[ h_S = \arg \min_h \sum_{(x_i, y_i) \in S} L(y_i, h(x_i)) \]

• Expected Test Error: aka test error of model class

\[ E_S \left[ L_P(h_S) \right] = E_S \left[ \mathbb{E}_{(x,y) \sim P(x,y)} [L(y, h_S(x))] \right] \]
Recall: Bias-Variance Decomposition

\[ E_S \left[ L_P(h_S) \right] = E_S \left[ E_{(x,y) \sim P(x,y)} \left[ L(y, h_S(x)) \right] \right] \]

- For squared error:

\[ E_S \left[ L_P(h_S) \right] = E_{(x,y) \sim P(x,y)} \left[ E_S \left[ (h_S(x) - H(x))^2 \right] + (H(x) - y)^2 \right] \]

**Variance Term**

**Bias Term**

\[ H(x) = E_S \left[ h_S(x) \right] \]

“Average prediction on x”
Recall: Bias-Variance Decomposition
Recall: Bias-Variance Decomposition

Some models experience high test error due to high bias. (Model class to simple to make accurate predictions.)

Some models experience high test error due to high variance. (Model class unstable due to insufficient training data.)
General Concept: Ensemble Methods

• Combine multiple learning algorithms or models
  – Previous Lecture: Bagging & Random Forests
  – Today: Boosting & Ensemble Selection

• “Meta Learning” approach
  – Does not innovate on base learning algorithm/model
  – Ex: Bagging
    • New training sets via bootstrapping
    • Combines by averaging predictions
Intuition: Why Ensemble Methods Work

• **Bias-Variance Tradeoff!**

• **Bagging reduces variance of low-bias models**
  – Low-bias models are “complex” and unstable
  – Bagging averages them together to create stability

• **Boosting reduces bias of low-variance models**
  – Low-variance models are simple with high bias
  – Boosting trains sequence of simple models
  – Sum of simple models is complex/accurate
Boosting
“The Strength of Weak Classifiers”*

Terminology: Shallow Decision Trees

• Decision Trees with only a few nodes
• Very high bias & low variance
  – Different training sets lead to very similar trees
  – Error is high (barely better than static baseline)
• Extreme case: “Decision Stumps”
  – Trees with exactly 1 split
Stability of Shallow Trees

- Tends to learn more-or-less the same model.
- $h_S(x)$ has low variance
  - Over the randomness of training set $S$
Terminology: Weak Learning

• **Error rate:** $\mathcal{E}_{h,P} = E_{P(x,y)} \left[ 1[h(x)\neq y] \right]$

• **Weak Classifier:** $\mathcal{E}_{h,P}$ slightly better than 0.5
  – Slightly better than random guessing

• **Weak Learner:** can learn a weak classifier
Terminology: Weak Learning

- **Error rate:** \( \mathcal{E}_{h,P} = E_{P(x,y)} \left[ 1[h(x) \neq y] \right] \)

- **Weak Classifier:** \( \mathcal{E}_{h,P} \) slightly better than 0.5
  - Slightly better than random guessing

Shallow Decision Trees are Weak Classifiers!

Weak Learners are Low Variance & High Bias!
How to “Boost” Weak Models?

\[ E_S \left[ L_P (h_S) \right] = E_{(x,y) \sim P(x,y)} \left[ E_S \left[ (h_S(x) - H(x))^2 \right] + (H(x) - y)^2 \right] \]

- Weak Models are High Bias & Low Variance
- Bagging would not work
  - Reduces variance, not bias
First Try (for Regression)

- 1 dimensional regression
- Learn Decision Stump
  - (single split, predict mean of two partitions)

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>36</td>
</tr>
</tbody>
</table>
First Try (for Regression)

- 1 dimensional regression
- Learn Decision Stump
  - (single split, predict mean of two partitions)

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>y_1</th>
<th>h_1(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td>25</td>
<td>30.5</td>
</tr>
<tr>
<td>36</td>
<td>30.5</td>
</tr>
</tbody>
</table>
First Try (for Regression)

• 1 dimensional regression
• Learn Decision Stump
  – (single split, predict mean of two partitions)

\[ y_t = y - h_{1:t-1}(x) \]

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( y_1 )</th>
<th>( h_1(x) )</th>
<th>( y_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>-6</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>-5</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>-2</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>-3</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>25</td>
<td>30.5</td>
<td>-5.5</td>
</tr>
<tr>
<td>36</td>
<td>30.5</td>
<td>5.5</td>
</tr>
</tbody>
</table>
First Try (for Regression)

- 1 dimensional regression
- Learn Decision Stump
  - (single split, predict mean of two partitions)

\[
\begin{align*}
S & = \begin{array}{c|c|c|c|c|}
0 & 0 & & \\
1 & 1 & & \\
2 & 4 & & \\
3 & 9 & & \\
4 & 16 & & \\
5 & 25 & & \\
6 & 36 & & \\
\end{array}
\end{align*}
\]

\[
\begin{array}{c|c|c|c|c}
y_1 & h_1(x) & y_2 & h_2(x) \\
\hline
0 & 6 & -6 & -5.5 \\
1 & 6 & -5 & -5.5 \\
4 & 6 & -2 & 2.2 \\
9 & 6 & -3 & 2.2 \\
16 & 6 & 10 & 2.2 \\
25 & 30.5 & -5.5 & 2.2 \\
36 & 30.5 & 5.5 & 2.2 \\
\end{array}
\]

\[y_t = y - h_{1:t-1}(x)\]
\[h_{1:t}(x) = h_1(x) + ... + h_t(x)\]

“residual”
First Try (for Regression)

• 1 dimensional regression
• Learn Decision Stump
  — (single split, predict mean of two partitions)

\[ y_t = y - h_{1:t-1}(x) \]
\[ h_{1:t}(x) = h_1(x) + \ldots + h_t(x) \]
First Try (for Regression)

- 1 dimensional regression
- Learn Decision Stump
  - (single split, predict mean of two partitions)

\[
S = \begin{array}{c|c}
  x & y \\
  \hline
 0 & 0 \\
 1 & 1 \\
 2 & 4 \\
 3 & 9 \\
 4 & 16 \\
 5 & 25 \\
 6 & 36 \\
\end{array}
\]

\[
\begin{array}{c|c|c|c|c|c|c|c}
  y_t & h_1(x) & y_2 & h_2(x) & h_{1:2}(x) & y_3 & h_3(x) & h_{1:3}(x) \\
  \hline
  0 & 6 & -6 & -5.5 & 0.5 & -0.5 & -0.55 & -0.05 \\
  1 & 6 & -5 & -5.5 & 0.5 & 0.5 & -0.55 & -0.05 \\
  4 & 6 & -2 & 2.2 & 8.2 & -4.2 & -0.55 & 7.65 \\
  9 & 6 & -3 & 2.2 & 8.2 & 0.8 & -0.55 & 7.65 \\
  16 & 6 & 10 & 2.2 & 8.2 & 7.8 & -0.55 & 7.65 \\
  25 & 30.5 & -5.5 & 2.2 & 32.7 & -7.7 & -0.55 & 32.15 \\
  36 & 30.5 & 5.5 & 2.2 & 32.7 & 3.3 & 3.3 & 36 \\
\end{array}
\]

\[
y_t = y - h_{1:t-1}(x) \\
h_{1:t}(x) = h_1(x) + \ldots + h_t(x)
\]

“residual”
First Try (for Regression)

\[ h_{1:t}(x) = h_1(x) + \ldots + h_t(x) \]

\[ y_t = y - h_{1:t-1}(x) \]
Gradient Boosting (Simple Version)

(Why is it called “gradient”?)
(Answer next slides.)

For Regression Only

\[ S = \{(x_i, y_i)\}_{i=1}^N \]

\[ h(x) = h_1(x) + h_2(x) + ... + h_n(x) \]

\[ S_1 = \{(x_i, y_i)\}_{i=1}^N \]
\[ S_2 = \{(x_i, y_i - h_1(x_i))\}_{i=1}^N \]
\[ S_n = \{(x_i, y_i - h_{1:n-1}(x_i))\}_{i=1}^N \]
Axis Aligned Gradient Descent

(For Linear Model)

• Linear Model: \( h(x) = w^T x \)

• Squared Loss: \( L(y, y') = (y - y')^2 \)

• Similar to Gradient Descent
  – But only allow axis-aligned update directions
  – Updates are of the form:

  \[
  w = w - \eta g_d e_d \\
  g = \sum_i \nabla_w L(y_i, w^T x_i)
  \]

  Unit vector along d-th dimension

  \[ e_d = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \]

  Projection of gradient along d-th dimension

  Update along axis with greatest projection

Training Set

\[ S = \{(x_i, y_i)\}_{i=1}^N \]
Axis Aligned Gradient Descent

Update along axis with largest projection

This concept will become useful in ~5 slides
Function Space & Ensemble Methods

• Linear model = one coefficient per feature
  – Linear over the input feature space

• Ensemble methods = one coefficient per model
  – Linear over a function space
  – E.g., \( h = h_1 + h_2 + \ldots + h_n \)

  “Function Space”
  (Span of all shallow trees)
  (Potentially infinite)
  (Most coefficients are 0)

Coefficient=1 for models used
Coefficient=0 for other models
Properties of Function Space

• Generalization of a Vector Space

• Closed under Addition
  – Sum of two functions is a function

• Closed under Scalar Multiplication
  – Multiplying a function with a scalar is a function

• Gradient descent: adding a scaled function to an existing function
Function Space of Models

• Every “axis” in the space is a weak model
  – Potentially infinite axes/dimensions

• Complex models are linear combinations of weak models
  – $h = \eta_1 h_1 + \eta_2 h_2 + \ldots + \eta_n h_n$
  – Equivalent to a point in function space
    • Defined by coefficients $\eta$
Recall: Axis Aligned Gradient Descent

Project to closest axis & update (smallest squared dist)

Imagine each axis is a weak model.

Every point is a linear combination of weak models
**Functional Gradient Descent**

(Gradient Descent in Function Space)

(Derivation for Squared Loss)

- Init $h(x) = 0$
- Loop $n=1,2,3,4,...$

$$h = h - \arg\max_{h_n} \left( \text{project}_{h_n} \left( \sum_i \nabla_h L(y_i, h(x_i)) \right) \right)$$

$$= h + \arg\min_{h_n} \sum_i (y_i - h(x_i) - h_n(x_i))^2$$

Equivalent to finding the $h_n$ that minimizes residual loss

$$S = \{(x_i, y_i)\}_{i=1}^N$$
Reduction to Vector Space

- Function space = axis-aligned unit vectors
  - Weak model = axis-aligned unit vector:
    \[ e_d = \begin{bmatrix} \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix} \]

- Linear model \( w \) has same functional form:
  - \( w = \eta_1 e_1 + \eta_2 e_2 + \ldots + \eta_D e_D \)
  - Point in space of \( D \) “axis-aligned functions”

- Axis-Aligned Gradient Descent = Functional Gradient Descent on space of axis-aligned unit vector weak models.
Gradient Boosting (Full Version)
(Instance of Functional Gradient Descent) (For Regression Only)

\[ S = \{(x_i, y_i)\}_i^{N} \]

\[ h_{1:n}(x) = h_1(x) + \eta_2 h_2(x) + \ldots + \eta_n h_n(x) \]

\[ S_1 = \{(x_i, y_i)\}_i^{N} \rightarrow S_2 = \{(x_i, y_i - h_1(x_i))\}_i^{N} \rightarrow S_n = \{(x_i, y_i - h_{1:n-1}(x_i))\}_i^{N} \]
Recap: Basic Boosting

• Ensemble of many weak classifiers.
  – \( h(x) = \eta_1 h_1(x) + \eta_2 h_2(x) + \ldots + \eta_n h_n(x) \)

• **Goal:** reduce bias using low-variance models

• **Derivation:** via Gradient Descent in Function Space
  – Space of weak classifiers

• We’ve only seen the regression so far...
AdaBoost

Adaptive Boosting for Classification

Boosting for Classification

• Gradient Boosting was designed for regression

• Can we design one for classification?

• AdaBoost
  – Adaptive Boosting
AdaBoost = Functional Gradient Descent

• AdaBoost is also instance of functional gradient descent:
  \[ h(x) = \text{sign}( a_1 h_1(x) + a_2 h_2(x) + \ldots + a_n h_n(x) ) \]

• E.g., weak models \( h_i(x) \) are classification trees
  – Always predict -1 or +1
  – (Gradient Boosting used regression trees)
Combining Multiple Classifiers

Aggregate Scoring Function:

\[ f(x) = 0.1h_1(x) + 1.5h_2(x) + 0.4h_3(x) + 1.1h_4(x) \]

Aggregate Classifier:

\[ h(x) = \text{sign}(f(x)) \]

<table>
<thead>
<tr>
<th>Data Point</th>
<th>( h_1(x) )</th>
<th>( h_2(x) )</th>
<th>( h_3(x) )</th>
<th>( h_4(x) )</th>
<th>( f(x) )</th>
<th>( h(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>0.1 + 1.5 + 0.4 - 1.1 = 0.9</td>
<td>+1</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>+1</td>
<td>0.1 + 1.5 + 0.4 + 1.1 = 3.1</td>
<td>+1</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
<td>-0.1 + 1.5 − 0.3 − 1.1 = -0.1</td>
<td>-1</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td>-1</td>
<td>-0.1 − 1.5 + 0.3 − 1.1 = -2.4</td>
<td>-1</td>
</tr>
</tbody>
</table>
Also Creates New Training Sets

• Gradients in Function Space
  – Weak model that outputs residual of loss function
    • Squared loss target: \( y-h(x) \)
  – Algorithmically equivalent to training weak model on modified training set
    • Gradient Boosting = train on \((x_i, y_i-h(x_i))\)

• What about AdaBoost?
  – Classification problem.
Reweighting Training Data

• Define weighting $D$ over $S$:
  – Sums to 1: $\sum_i D(i) = 1$

• Examples:

<table>
<thead>
<tr>
<th>Data Point</th>
<th>D(i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(x_1, y_1)$</td>
<td>$\frac{1}{3}$</td>
</tr>
<tr>
<td>$(x_2, y_2)$</td>
<td>$\frac{1}{3}$</td>
</tr>
<tr>
<td>$(x_3, y_3)$</td>
<td>$\frac{1}{3}$</td>
</tr>
</tbody>
</table>

• Weighted loss function:

$$L_D(h) = \sum_i D(i)L(y_i, h(x_i))$$
Training Decision Trees with Weighted Training Data

• Slight modification of splitting criterion.

• Example: Bernoulli Variance:

\[ L(S') = |S'| p_{S'} (1 - p_{S'}) = \frac{\# pos \times \# neg}{|S'|} \]

• Estimate fraction of positives as:

\[ p_{S'} = \frac{\sum_{(x_i, y_i) \in S'} D(i) 1_{[y_i=1]}}{|S'|} \quad |S'| \equiv \sum_{(x_i, y_i) \in S'} D(i) \]
**AdaBoost Outline**

\[ h(x) = \text{sign}(a_1 h_1(x) + a_2 h_2(x) + \ldots + a_n h_n(x)) \]

\[ S = \{(x_i, y_i)\}_{i=1}^{N} \]

\[ y_i \in \{-1, +1\} \]

\((S, D_1 = \text{Uniform}) \rightarrow h_1(x) \)

\((S, D_2) \rightarrow h_2(x) \)

\((S, D_n) \rightarrow h_n(x) \)

\[ D_t \rightarrow \text{weighting on data points} \]

\[ a_t \rightarrow \text{weight of linear combination} \]


Stop when validation performance plateaus (will discuss later)
Intuition

Aggregate Scoring Function:
\[ f(x) = 0.1 \cdot h_1(x) + 1.5 \cdot h_2(x) + 0.4 \cdot h_3(x) + 1.1 \cdot h_4(x) \]

Aggregate Classifier:
\[ h(x) = \text{sign}(f(x)) \]

### Aggregate Scoring Function

**Data Point** | **Label** | **f(x)** | **h(x)**
--- | --- | --- | ---
\( x_1 \) | \( y_1=+1 \) | 0.9 | +1
\( x_2 \) | \( y_2=+1 \) | 3.1 | +1
\( x_3 \) | \( y_3=+1 \) | -0.1 | -1
\( x_4 \) | \( y_4=-1 \) | -2.4 | -1

---

Safely Far from Decision Boundary

Violates Decision Boundary

Somewhat close to Decision Boundary
Intuition

Thought Experiment:
When we train new $h_5(x)$ to add to $f(x)$...
... what happens when $h_5$ mispredicts on everything?

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Label</th>
<th>$f(x)$</th>
<th>$h(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$y_1=+1$</td>
<td>0.9</td>
<td>+1</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$y_2=+1$</td>
<td>3.1</td>
<td>+1</td>
</tr>
<tr>
<td>$x_3$</td>
<td>$y_3=+1$</td>
<td>-0.1</td>
<td>-1</td>
</tr>
<tr>
<td>$x_4$</td>
<td>$y_4=-1$</td>
<td>-2.4</td>
<td>-1</td>
</tr>
</tbody>
</table>

Somewhat close to Decision Boundary
Violates Decision Boundary
Safely Far from Decision Boundary
**Intuition**

**Aggregate Scoring Function:**
\[ f_{1:5}(x) = f_{1:4}(x) + 0.5h_5(x) \]

**Aggregate Classifier:**
\[ h_{1:5}(x) = \text{sign}(f_{1:5}(x)) \]

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Label</th>
<th>( f_{1:4}(x) )</th>
<th>( h_{1:4}(x) )</th>
<th>Worst case ( h_5(x) )</th>
<th>Worst case ( f_{1:5}(x) )</th>
<th>Impact of ( h_5(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( y_1 = +1 )</td>
<td>0.9</td>
<td>+1</td>
<td>-1</td>
<td>0.4</td>
<td>Kind of Bad</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( y_2 = +1 )</td>
<td>3.1</td>
<td>+1</td>
<td>-1</td>
<td>2.6</td>
<td>Irrelevant</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>( y_3 = +1 )</td>
<td>-0.1</td>
<td>-1</td>
<td>-1</td>
<td>-0.6</td>
<td>Very Bad</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>( y_4 = -1 )</td>
<td>-2.4</td>
<td>-1</td>
<td>+1</td>
<td>-1.9</td>
<td>Irrelevant</td>
</tr>
</tbody>
</table>

Suppose \( a_5 = 0.5 \)

\( h_5(x) \) that mispredicts on everything
Intuition

$h_5(x)$ should definitely classify $(x_3, y_3)$ correctly!
$h_5(x)$ should probably classify $(x_1, y_1)$ correctly.
Don’t care about $(x_2, y_2) \& (x_4, y_4)$
Implies a weighting over training examples

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Label</th>
<th>$f_{1:4}(x)$</th>
<th>$h_{1:4}(x)$</th>
<th>Worst case $h_5(x)$</th>
<th>Worst case $f_{1:5}(x)$</th>
<th>Impact of $h_5(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$y_1=+1$</td>
<td>0.9</td>
<td>+1</td>
<td>-1</td>
<td>0.4</td>
<td>Kind of Bad</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$y_2=+1$</td>
<td>3.1</td>
<td>+1</td>
<td>-1</td>
<td>2.6</td>
<td>Irrelevant</td>
</tr>
<tr>
<td>$x_3$</td>
<td>$y_3=+1$</td>
<td>-0.1</td>
<td>-1</td>
<td>-1</td>
<td>-0.6</td>
<td>Very Bad</td>
</tr>
<tr>
<td>$x_4$</td>
<td>$y_4=-1$</td>
<td>-2.4</td>
<td>-1</td>
<td>+1</td>
<td>-1.9</td>
<td>Irrelevant</td>
</tr>
</tbody>
</table>

$h_5(x)$ that mispredicts on everything
Intuition

Aggregate Scoring Function:

\[ f_{1:4}(x) = 0.1*h_1(x) + 1.5*h_2(x) + 0.4*h_3(x) + 1.1*h_4(x) \]

Aggregate Classifier:

\[ h_{1:4}(x) = \text{sign}(f_{1:4}(x)) \]

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Label</th>
<th>( f_{1:4}(x) )</th>
<th>( h_{1:4}(x) )</th>
<th>Desired ( D_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( y_1=+1 )</td>
<td>0.9</td>
<td>+1</td>
<td>Medium</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( y_2=+1 )</td>
<td>3.1</td>
<td>+1</td>
<td>Low</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>( y_3=+1 )</td>
<td>-0.1</td>
<td>-1</td>
<td>High</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>( y_4=-1 )</td>
<td>-2.4</td>
<td>-1</td>
<td>Low</td>
</tr>
</tbody>
</table>
AdaBoost

- Init \( D_1(x) = 1/N \)
- Loop \( t = 1 \ldots n \):
  - Train classifier \( h_t(x) \) using \( (S, D_t) \)
  - Compute error on \( (S, D_t) \):
    \[ \varepsilon_t \equiv L_{D_t}(h_t) = \sum_i D_t(i)L(y_i, h_t(x_i)) \]
  - Define step size \( a_t \):
    \[ a_t = \frac{1}{2} \log \left\{ \frac{1 - \varepsilon_t}{\varepsilon_t} \right\} \]
  - Update Weighting:
    \[ D_{t+1}(i) = \frac{D_t(i) \exp\{ -a_t y_i h_t(x_i) \}}{Z_t} \]
- Return: \( h(x) = \text{sign}(a_1 h_1(x) + \ldots + a_n h_n(x)) \)

E.g., best decision stump

\[ S = \{ (x_i, y_i) \}_{i=1}^N \]
\[ y_i \in \{-1, +1\} \]
Example

\[ \varepsilon_t = L_{D_t}(h_t) = \sum_i D_t(i) L(y_i, h_t(x_i)) \]

\[ a_t = \frac{1}{2} \log \left( \frac{1 - \varepsilon_t}{\varepsilon_t} \right) \]

\[ D_{t+1}(i) = \frac{D_t(i) \exp \{-a_t y_i h_t(x_i)\}}{Z_t} \]

\[ y_i h_t(x_i) = -1 \text{ or } +1 \]

Normalization Factor
s.t. \( D_{t+1} \) sums to 1.

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Label</th>
<th>( D_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( y_1=+1 )</td>
<td>0.01</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( y_2=+1 )</td>
<td>0.01</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>( y_3=+1 )</td>
<td>0.01</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>( y_4=-1 )</td>
<td>0.01</td>
</tr>
</tbody>
</table>

\[ \vdots \]
Example

$$\varepsilon_t = L_{D_t}(h_t) = \sum_i D_t(i)L(y_i, h_t(x_i))$$

$$a_t = \frac{1}{2} \log \left\{ \frac{1 - \varepsilon_t}{\varepsilon_t} \right\}$$

$$\varepsilon_1 = 0.4$$
$$a_1 = 0.2$$

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Label</th>
<th>$D_1$</th>
<th>$h_1(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$y_1=+1$</td>
<td>0.01</td>
<td>+1</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$y_2=+1$</td>
<td>0.01</td>
<td>-1</td>
</tr>
<tr>
<td>$x_3$</td>
<td>$y_3=+1$</td>
<td>0.01</td>
<td>-1</td>
</tr>
<tr>
<td>$x_4$</td>
<td>$y_4=-1$</td>
<td>0.01</td>
<td>-1</td>
</tr>
</tbody>
</table>

$$y_i h_t(x_i) = -1 \text{ or } +1$$

$$D_{t+1}(i) = \frac{D_t(i) \exp \{-a_t y_i h_t(x_i)\}}{Z_t}$$

Normalization Factor s.t. $D_{t+1}$ sums to 1.
Example

$$\varepsilon_t = L_{D_t}(h_t) = \sum_i D_t(i) L(y_i, h_t(x_i))$$

$$a_t = \frac{1}{2} \log \left\{ \frac{1 - \varepsilon_t}{\varepsilon_t} \right\}$$

$$\varepsilon_1 = 0.4$$
$$a_1 = 0.2$$

$$y_i h_t(x_i) = -1 \text{ or } +1$$

$$D_{t+1}(i) = \frac{D_t(i) \exp \{-a_t y_i h_t(x_i)\}}{Z_t}$$

Normalization Factor s.t. $D_{t+1}$ sums to 1.

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Label</th>
<th>$D_1$</th>
<th>$h_1(x)$</th>
<th>$D_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$y_1=+1$</td>
<td>0.01</td>
<td>+1</td>
<td>0.008</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$y_2=+1$</td>
<td>0.01</td>
<td>-1</td>
<td>0.012</td>
</tr>
<tr>
<td>$x_3$</td>
<td>$y_3=+1$</td>
<td>0.01</td>
<td>-1</td>
<td>0.012</td>
</tr>
<tr>
<td>$x_4$</td>
<td>$y_4=-1$</td>
<td>0.01</td>
<td>-1</td>
<td>0.008</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Example

\[ \epsilon_t = L_{D_t}(h_t) = \sum_i D_t(i) L(y_i, h_t(x_i)) \]

\[ a_t = \frac{1}{2} \log \left( \frac{1 - \epsilon_t}{\epsilon_t} \right) \]

\[ y_i h_t(x_i) = -1 \text{ or } +1 \]

\[ D_{t+1}(i) = \frac{D_t(i) \exp\{-a_t y_i h_t(x_i)\}}{Z_t} \]

Normalization Factor
s.t. \( D_{t+1} \) sums to 1.

\[ \epsilon_1 = 0.4 \quad a_1 = 0.2 \]
\[ \epsilon_2 = 0.45 \quad a_2 = 0.1 \]

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Label</th>
<th>( D_1 )</th>
<th>( h_1(x) )</th>
<th>( D_2 )</th>
<th>( h_2(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( y_1 = +1 )</td>
<td>0.01</td>
<td>+1</td>
<td>0.008</td>
<td>+1</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( y_2 = +1 )</td>
<td>0.01</td>
<td>-1</td>
<td>0.012</td>
<td>+1</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>( y_3 = +1 )</td>
<td>0.01</td>
<td>-1</td>
<td>0.012</td>
<td>-1</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>( y_4 = -1 )</td>
<td>0.01</td>
<td>-1</td>
<td>0.008</td>
<td>+1</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
</tbody>
</table>
\[ \varepsilon_t \equiv L_{D_t}(h_t) = \sum_i D_t(i)L(y_i, h_t(x_i)) \]
\[ a_t = \frac{1}{2} \log \left( \frac{1 - \varepsilon_t}{\varepsilon_t} \right) \]
\[ y_i h_t(x_i) = -1 \text{ or } +1 \]
\[ D_{t+1}(i) = \frac{D_t(i) \exp \{-a_t y_i h_t(x_i)\}}{Z_t} \]

Normalization Factor
s.t. \( D_{t+1} \) sums to 1.

\[ \varepsilon_1 = 0.4 \]
\[ a_1 = 0.2 \]
\[ \varepsilon_2 = 0.45 \]
\[ a_2 = 0.1 \]

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Label</th>
<th>D1</th>
<th>h1(x)</th>
<th>D2</th>
<th>h2(x)</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( y_1 = +1 )</td>
<td>0.01</td>
<td>+1</td>
<td>0.008</td>
<td>+1</td>
<td>0.007</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( y_2 = +1 )</td>
<td>0.01</td>
<td>-1</td>
<td>0.012</td>
<td>+1</td>
<td>0.011</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>( y_3 = +1 )</td>
<td>0.01</td>
<td>-1</td>
<td>0.012</td>
<td>-1</td>
<td>0.013</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>( y_4 = -1 )</td>
<td>0.01</td>
<td>-1</td>
<td>0.008</td>
<td>+1</td>
<td>0.009</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
</tbody>
</table>
Example

\[ \varepsilon_t = L_{D_t}(h_t) = \sum_i D_t(i) L(y_i, h_t(x_i)) \]

\[ a_t = \frac{1}{2} \log \left( \frac{1 - \varepsilon_t}{\varepsilon_t} \right) \]

What happens if \( \varepsilon = 0.5 \)?

\[ y_i h_t(x_i) = -1 \text{ or } +1 \]

Normalization Factor s.t. \( D_{t+1} \) sums to 1.

<table>
<thead>
<tr>
<th>Data Point</th>
<th>Label</th>
<th>( D_1 )</th>
<th>( h_1(x) )</th>
<th>( D_2 )</th>
<th>( h_2(x) )</th>
<th>( D_3 )</th>
<th>( h_3(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( y_1 = +1 )</td>
<td>0.01</td>
<td>+1</td>
<td>0.008</td>
<td>+1</td>
<td>0.007</td>
<td>-1</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( y_2 = +1 )</td>
<td>0.01</td>
<td>-1</td>
<td>0.012</td>
<td>+1</td>
<td>0.011</td>
<td>+1</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>( y_3 = +1 )</td>
<td>0.01</td>
<td>-1</td>
<td>0.012</td>
<td>-1</td>
<td>0.013</td>
<td>+1</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>( y_4 = -1 )</td>
<td>0.01</td>
<td>-1</td>
<td>0.008</td>
<td>+1</td>
<td>0.009</td>
<td>-1</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
</tbody>
</table>

\[ \varepsilon_1 = 0.4 \quad a_1 = 0.2 \]
\[ \varepsilon_2 = 0.45 \quad a_2 = 0.1 \]
\[ \varepsilon_3 = 0.35 \quad a_3 = 0.31 \]
Exponential Loss

\[ L(y, f(x)) = \exp\{-yf(x)\} \]

Upper Bounds
0/1 Loss!

Can prove that
AdaBoost minimizes
Exp Loss
(Homework Question)
Decomposing Exp Loss

\[ L(y, f(x)) = \exp\{-yf(x)\} \]

\[ = \exp\left\{-y \left( \sum_{t=1}^{n} a_t h_t(x) \right) \right\} \]

\[ = \prod_{t=1}^{n} \exp\left\{-ya_t h_t(x)\right\} \]

\textbf{Distribution Update Rule!}

Intuition

\[
L(y, f(x)) = \exp \left\{ -y \sum_{t=1}^{n} a_t h_t(x) \right\} = \prod_{t=1}^{n} \exp \{ -ya_t h_t(x) \}
\]

• Exp Loss operates in exponent space

• Additive update to \( f(x) = \) multiplicative update to Exp Loss of \( f(x) \)

• Reweighting Scheme in AdaBoost can be derived via residual Exp Loss
AdaBoost = Minimizing Exp Loss

- Init $D_1(x) = 1/N$
- Loop $t = 1...n$:
  - Train classifier $h_t(x)$ using $(S,D_t)$
  - Compute error on $(S,D_t)$: $\varepsilon_t \equiv L_{D_t}(h_t) = \sum_i D_t(i) L(y_i, h_t(x_i))$
  - Define step size $a_t$: $a_t = \frac{1}{2} \log \left\{ \frac{1 - \varepsilon_t}{\varepsilon_t} \right\}$
  - Update Weighting: $D_{t+1}(i) = \frac{D_t(i) \exp\{-a_t y_i h_t(x_i)\}}{Z_t}$
- Return: $h(x) = \text{sign}(a_1 h_1(x) + ... + a_n h_n(x))$

$S = \{(x_i, y_i)\}_{i=1}^N$
$y_i \in \{-1, +1\}$

Data points reweighted according to Exp Loss!

Normalization Factor s.t. $D_{t+1}$ sums to 1.

Story So Far: AdaBoost

• AdaBoost iteratively finds weak classifier to minimize residual Exp Loss
  – Trains weak classifier on reweighted data (S,D_t).

• Homework: Rigorously prove it!
  1. Formally prove Exp Loss ≥ 0/1 Loss
  2. Relate Exp Loss to Z_t:

\[
D_{t+1}(i) = \frac{D_t(i) \exp\{-a_t y_i h_t(x_i)\}}{Z_t}
\]

\[
a_t = \frac{1}{2} \log \left\{ \frac{1-\epsilon_t}{\epsilon_t} \right\}
\]

3. Justify choice of a_t:
   • Gives largest decrease in Z_t

Recap: AdaBoost

- **Gradient Descent in Function Space**
  - Space of weak classifiers

- **Final model = linear combination of weak classifiers**
  - \( h(x) = \text{sign}(a_1 h_1(x) + \ldots + a_n h_n(x)) \)
  - I.e., a point in Function Space

- **Iteratively creates new training sets via reweighting**
  - Trains weak classifier on reweighted training set
  - Derived via minimizing residual Exp Loss
Ensemble Selection
Recall: Bias-Variance Decomposition

\[
E_S \left[ L_P(h_S) \right] = E_S \left[ E_{(x,y)\sim P(x,y)} \left[ L(y, h_S(x)) \right] \right]
\]

- For squared error:

\[
E_S \left[ L_P(h_S) \right] = E_{(x,y)\sim P(x,y)} \left[ E_S \left[ (h_S(x) - H(x))^2 \right] + (H(x) - y)^2 \right]
\]

\[
H(x) = E_S \left[ h_S(x) \right]
\]

“Average prediction on x”
Ensemble Methods

• Combine base models to improve performance

• **Bagging**: averages high variance, low bias models
  – Reduces variance
  – Indirectly deals with bias via low bias base models

• **Boosting**: carefully combines simple models
  – Reduces bias
  – Indirectly deals with variance via low variance base models

• **Can we get best of both worlds?**
Insight: Use Validation Set

• Evaluate error on validation set V:

\[ L_V(h_S) = E_{(x,y) \sim V} \left[ L(y, h_S(x)) \right] \]

• Proxy for test error:

\[ E_V \left[ L_V(h_S) \right] = L_P(h_S) \]

- Expected Validation Error
- Test Error
Ensemble Selection

**Maintain ensemble model as combination of H:**

\[ h(x) = h_1(x) + h_2(x) + \ldots + h_n(x) + h_{n+1}(x) \]

**Add model from H that maximizes performance on V’**

H = \{2000 models trained using S’\}

Repeat

“Ensemble Selection from Libraries of Models”
Caruana, Niculescu-Mizil, Crew & Ksikes, ICML 2004
Reduces Both Bias & Variance

• Expected Test Error = Bias + Variance

• **Bagging:** reduce variance of low-bias models

• **Boosting:** reduce bias of low-variance models

• **Ensemble Selection:** who cares!
  – Use validation error to approximate test error
  – Directly minimize validation error
  – Don’t worry about the bias/variance decomposition
What’s the Catch?

• Relies heavily on validation set
  – Bagging & Boosting: uses training set to select next model
  – Ensemble Selection: uses validation set to select next model

• Requires validation set be sufficiently large

• In practice: implies smaller training sets
  – Training & validation = partitioning of finite data

• Often works very well in practice
Ensemble Selection often outperforms a more homogenous sets of models. Reduces overfitting by building model using validation set.

Ensemble Selection won KDD Cup 2009
http://www.niculescu-mizil.org/papers/KDDCup09.pdf

“Ensemble Selection from Libraries of Models”
Caruana, Niculescu-Mizil, Crew & Ksikes, ICML 2004
References & Further Reading


“Ensemble Methods in Machine Learning” Thomas Dietterich, Multiple Classifier Systems, 2000

“Ensemble Selection from Libraries of Models” Caruana, Niculescu-Mizil, Crew & Ksikes, ICML 2004

“Getting the Most Out of Ensemble Selection” Caruana, Munson, & Niculescu-Mizil, ICDM 2006


“Random Forests – Random Features” Leo Breiman, Tech Report #567, UC Berkeley, 1999,

“Structured Random Forests for Fast Edge Detection” Dollár & Zitnick, ICCV 2013

“ABC-Boost: Adaptive Base Class Boost for Multi-class Classification” Ping Li, ICML 2009


“Winning the KDD Cup Orange Challenge with Ensemble Selection”, Niculescu-Mizil et al., KDD 2009

Next Lectures

• Deep Learning

• Recitation Next Thursday
  – Keras Tutorial