CS 66: Machine Learning

Prof. Sara Mathieson

Spring 2019
Outline for March 6

- Lab 2 examples
- Ensemble methods
  - Bagging
  - Random Forests
  - Boosting

- Lab 4 due Friday
- Check-in today during lab
  - should be done with one of Logistic Regression or Naïve Bayes
Outline for March 6

• Lab 2 examples
• Ensemble methods
  – Bagging
  – Random Forests
  – Boosting
Lab 2 (heart): Henrik & Prav
Lab 2 (heart): Lyla & Sam

Tree Depth vs Accuracy of Train/Test Data

- **training**
- **testing**
Lab 2 (heart): Raymond & Kenny

Decision Tree Accuracy for Heart Dataset

Accuracy (%) vs. Maximum Depth

- **Train data**
- **Test data**
Lab 2 (heart): Mikey & Dylan
<table>
<thead>
<tr>
<th>Depth</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.825</td>
<td>0.7281553398058253</td>
</tr>
<tr>
<td>2</td>
<td>0.86</td>
<td>0.7281553398058253</td>
</tr>
<tr>
<td>3</td>
<td>0.875</td>
<td>0.7184466019417476</td>
</tr>
<tr>
<td>4</td>
<td>0.91</td>
<td>0.6893203883495146</td>
</tr>
<tr>
<td>5</td>
<td>0.915</td>
<td>0.6893203883495146</td>
</tr>
<tr>
<td>6</td>
<td>0.91</td>
<td>0.6796116504854369</td>
</tr>
<tr>
<td>7</td>
<td>0.92</td>
<td>0.6601941747572816</td>
</tr>
<tr>
<td>8</td>
<td>0.92</td>
<td>0.6601941747572816</td>
</tr>
<tr>
<td>9</td>
<td>0.92</td>
<td>0.6601941747572816</td>
</tr>
<tr>
<td>10</td>
<td>0.92</td>
<td>0.6601941747572816</td>
</tr>
<tr>
<td>11</td>
<td>0.92</td>
<td>0.6601941747572816</td>
</tr>
<tr>
<td>12</td>
<td>0.92</td>
<td>0.6601941747572816</td>
</tr>
<tr>
<td>13</td>
<td>0.92</td>
<td>0.6601941747572816</td>
</tr>
<tr>
<td>14</td>
<td>0.92</td>
<td>0.6601941747572816</td>
</tr>
<tr>
<td>15</td>
<td>0.92</td>
<td>0.6601941747572816</td>
</tr>
<tr>
<td>16</td>
<td>0.92</td>
<td>0.6601941747572816</td>
</tr>
<tr>
<td>17</td>
<td>0.92</td>
<td>0.6601941747572816</td>
</tr>
<tr>
<td>18</td>
<td>0.92</td>
<td>0.6601941747572816</td>
</tr>
<tr>
<td>19</td>
<td>0.92</td>
<td>0.6601941747572816</td>
</tr>
<tr>
<td>20</td>
<td>0.92</td>
<td>0.6601941747572816</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Depth</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7544910179640718</td>
<td>0.7</td>
</tr>
<tr>
<td>2</td>
<td>0.7544910179640718</td>
<td>0.7</td>
</tr>
<tr>
<td>3</td>
<td>0.7544910179640718</td>
<td>0.69</td>
</tr>
<tr>
<td>4</td>
<td>0.7529940119760479</td>
<td>0.73</td>
</tr>
<tr>
<td>5</td>
<td>0.7964071856287425</td>
<td>0.69</td>
</tr>
<tr>
<td>6</td>
<td>0.7934131736526946</td>
<td>0.73</td>
</tr>
<tr>
<td>7</td>
<td>0.8143712574850299</td>
<td>0.69</td>
</tr>
<tr>
<td>8</td>
<td>0.8203592814371258</td>
<td>0.64</td>
</tr>
<tr>
<td>9</td>
<td>0.8218562874251497</td>
<td>0.65</td>
</tr>
<tr>
<td>10</td>
<td>0.8353293413173652</td>
<td>0.7</td>
</tr>
<tr>
<td>11</td>
<td>0.842814371257485</td>
<td>0.7</td>
</tr>
<tr>
<td>12</td>
<td>0.8458083832335329</td>
<td>0.69</td>
</tr>
<tr>
<td>13</td>
<td>0.8502994011976048</td>
<td>0.67</td>
</tr>
<tr>
<td>14</td>
<td>0.8502994011976048</td>
<td>0.67</td>
</tr>
<tr>
<td>15</td>
<td>0.8502994011976048</td>
<td>0.67</td>
</tr>
<tr>
<td>16</td>
<td>0.8502994011976048</td>
<td>0.67</td>
</tr>
<tr>
<td>17</td>
<td>0.8502994011976048</td>
<td>0.67</td>
</tr>
<tr>
<td>18</td>
<td>0.8502994011976048</td>
<td>0.67</td>
</tr>
<tr>
<td>19</td>
<td>0.8502994011976048</td>
<td>0.67</td>
</tr>
<tr>
<td>20</td>
<td>0.8502994011976048</td>
<td>0.67</td>
</tr>
</tbody>
</table>
Lab 2 (diabetes): Haochen & Nav
Lab 2 (diabetes): Greg & Hari

diabetes data accuracy

[Graph showing accuracy over depths for diabetes training and testing data]
Lab 2 (percent train): Mikey & Dylan
Lab 2 (multi-class): Gabriel & Keton

Wine Region Prediction Accuracy vs Depth

Accuracy

Maximum Tree Depth

Train
Test
Outline for March 6

• Lab 2 examples
• Ensemble methods
  – Bagging
  – Random Forests
  – Boosting
Ensemble Notation

$T$: # of models/classifiers (index $t$)

$\tilde{x}$: test example

$y$: test label (binary)

$X_t$: bootstrap training dataset $t$

$h^{(t)}(x)$: hypothesis about $x$ from model $t$

$r$: prob of error for each model

$R$: # votes for wrong class
Bagging (Bootstrap Aggregation)

Training data

$\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_n$

Why $n$ examples?

Prob didn't choose a data point:

$$\left(\frac{n-1}{n}\right)^n = \left(1 - \frac{1}{n}\right)^n \approx e^{-\frac{1}{n}}$$

$$\lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n = e^x$$

$T = 3$
Algorithm Idea

Train
for $t = 1, 2, \ldots, T$:
- create bootstrap sample $X_t$
- train on $X_t$
to get model

Test:
for $\tilde{x}$ in test:
$h(\tilde{x}) = \arg\max_{y \in \{0, 1, \beta\}}$

\[
R = \sum_{t=1}^{T} \mathbb{1}(h^{(t)}(x) = \hat{y})
\]
wrong class

\[
P(R = k) = \binom{T}{k} r^k (1-r)^{T-k}
\]
"T choose k" $k$ times

right $(T-k)$ times

\[
\mathbb{P}(\text{overfit}) = \mathbb{P}(R = k) = \frac{T!}{k!(T-k)!}
\]

\[
\sum_{t=1}^{T} \mathbb{1}(h^{(t)}(x) = y)
\]

\[
\begin{array}{c|ccc}
\hline
& h^0 & h^1 & h^2 \\
\hline
0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
\hline
\end{array}
\]

\[
\mathbb{P}(err) = \binom{3}{2} = 3
\]
\[ R = \sum_{t=1}^{T} 1(h_t(x) = y) \] 

Wrong class.

\[ P(R = k) = \binom{T}{k} r^k (1-r)^{T-k} \] 

\( T \) choose \( k \) \( \) times wrong \( \) times \( (T-k) \) times 

Right:

\[ P(\text{overall wrong}) = P(R > \frac{T}{2}) \]

\[ = \sum_{k=\frac{T+1}{2}}^{T} \binom{T}{k} r^k (1-r)^{T-k} \]

As \( T \to \infty \)

\[ P(\text{err}) \to 0 \]

If \( r < 0.5 \)
Example

$T = 3$

$\bar{x}_{test} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$

$\bar{y}_{test} = 1$

<table>
<thead>
<tr>
<th>$h^{(1)}$</th>
<th>$h^{(2)}$</th>
<th>$h^{(3)}$</th>
<th>$P_{\text{error}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$(\frac{1}{4})^3$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>$(\frac{1}{4})^2 (\frac{3}{4})$</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>$(\frac{1}{4})^2 (\frac{3}{4})$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>$(\frac{1}{4})^2 (\frac{3}{4})$</td>
</tr>
</tbody>
</table>

$P_{\text{error}} = (\frac{1}{4})^3 + (\frac{1}{4})^2 (\frac{3}{4}) \cdot 3$

$= \frac{10}{64} \approx 0.16$

compare to $r$ and it's better!

Decision stump

```
weather
  /\  \\
 sun /  \ rain
  \   / wind
    \ no
```

depth 1
1 feature
Outline for March 6

• Lab 2 examples
• Ensemble methods
  – Bagging
  – Random Forests
  – Boosting
Random Forests

Idea: choose a different subset of features for every classifier $t$.

Goal: decorrelate trees' models

In practice: choose $\sqrt{p}$ features

* without replacement for each model
* every model independent data points and independent features
Heart data: bagging vs. random forest

**FIGURE 8.8.** Bagging and random forest results for the Heart data. The test error (black and orange) is shown as a function of $B$, the number of bootstrapped training sets used. Random forests were applied with $m = \sqrt{p}$. The dashed line indicates the test error resulting from a single classification tree. The green and blue traces show the OOB error, which in this case is considerably lower.
Heart data: most important features

**FIGURE 8.9.** A variable importance plot for the Heart data. Variable importance is computed using the mean decrease in Gini index, and expressed relative to the maximum.