Finding a Higgs in a Haystack
A selection of data modeling techniques for separating signal from noise at ATLAS

Andrew Farrar
The ATLAS detector is one of the main particle detectors on the Large Hadron Collider at CERN.
The ATLAS detector is one of the main particle detectors on the Large Hadron Collider at CERN.

During an experiment, the detector records millions of collisions and decay events.
Context

- The ATLAS detector is one of the main particle detectors on the Large Hadron Collider at CERN
- During an experiment, the detector records millions of collisions and decay events
- Much of this is not interesting, we are looking for very specific signals
Data Science Steps Up

- Obviously too much data to sift through by hand
Data Science Steps Up

- Obviously too much data to sift through by hand
- Instead, build models using labeled synthetic data to apply to the real thing

Sample data:

| Feature 1 | Feature 2 | Feature 3 | Feature 4 | Feature 5 | Feature 6 | Feature 7 | Feature 8 | Feature 9 | Feature 10 | Feature 11 | Feature 12 | Feature 13 | Feature 14 | Feature 15 | Feature 16 | Feature 17 | Feature 18 | Feature 19 | Feature 20 | Feature 21 | Feature 22 | Feature 23 | Feature 24 | Feature 25 | Feature 26 | Feature 27 | Feature 28 | Feature 29 | Feature 30 | Feature 31 |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 138.47    | 51.655    | 97.827    | 27.98     | 0.91      | 124.711   | 2.666     | 3.064     | 41.928    | 197.76    | 1.582     | 1.396     | 0.2       | 32.638    | 1.017     | 0.381     | 51.626    | 2.273     | -2.414    | 16.824    | -0.277    | 258.733   | 2         | 67.435    | 2.15      | 0.444     | 46.062    | 1.24      | -2.475    | 113.497   | s         |

Some measured directly, others derived using known physical laws and equations from observed data.
Data Science Steps Up

- Obviously too much data to sift through by hand
- Instead, build models using labeled synthetic data to apply to the real thing
- Goal: identify the specific decay events of a Higgs boson (labeled “s” for signal) in the sea of other data (labeled “b” for background)

There are 30 features in addition to the class label. Sample:

138.47, 51.655, 97.827, 27.98, 0.91, 124.711, 2.666, 3.064, 41.928, 197.76, 1.582, 1.396, 0.2, 32.638, 1.017, 0.381, 51.626, 2.273, -2.414, 16.824, -0.277, 258.733, 2, 67.435, 2.15, 0.444, 46.062, 1.24, -2.475, 113.497, s

Some measured directly, others derived using known physical laws and equations from observed data.
Data Science Steps Up

- Obviously too much data to sift through by hand
- Instead, build models using labeled synthetic data to apply to the real thing
- Goal: identify the specific decay events of a Higgs boson (labeled “s” for signal) in the sea of other data (labeled “b” for background)
- There are 30 features in addition to the class label. Sample:

  138.47,51.655,97.827,27.98,0.91,124.711,2.666,
  3.064,41.928,197.76,1.582,1.396,0.2,32.638,
  1.017,0.381,51.626,2.273,-2.414,16.824,-0.277,
  258.733,2,67.435,2.15,0.444,46.062,1.24,-2.475,113.497,s

- Some measured directly, others derived using known physical laws and equations from observed data
Data Science Steps Up

- Obviously too much data to sift through by hand
- Instead, build models using labeled synthetic data to apply to the real thing
- Goal: identify the specific decay events of a Higgs boson (labeled “s” for signal) in the sea of other data (labeled “b” for background)
- There are 30 features in addition to the class label. Sample:
  
  138.47, 51.655, 97.827, 27.98, 0.91, 124.711, 2.666, 
  3.064, 41.928, 197.76, 1.582, 1.396, 0.2, 32.638, 
  1.017, 0.381, 51.626, 2.273, -2.414, 16.824, -0.277, 
  258.733, 2, 67.435, 2.15, 0.444, 46.062, 1.24, -2.475, 113.497, s
- Some measured directly, others derived using known physical laws and equations from observed data
What Models To Use?

- Clearly a binary classification problem (s vs b)

- Decided to go with:
  - Logistic Regression
  - Naive Bayes
  - Decision Tree

- Also performed two types of feature selection:
  - PCA (selected the first 14 features)
  - Information Gain (ranked features 1, 2, 3, 14, 12, 11, 5, 6, 7, 10, 13, 25, 4, 30, 23, 24, 22, 28, 8, 20, 26, 18, 27 & 29)

- Performance evaluated using 10-fold cross-validation

- All analysis performed through the WEKA tool created at the University of Waikato
What Models To Use?

- Clearly a binary classification problem (s vs b)
- Decided to go with:
  - Logistic Regression
  - Naive Bayes
  - Decision Tree

- Also performed two types of feature selection:
  - PCA (selected the first 14 features)
  - Information Gain (ranked features 1, 2, 3, 14, 12, 11, 5, 6, 7, 10, 13, 25, 4, 30, 23, 24, 22, 28, 8, 20, 26, 18, 27 & 29)

- Performance evaluated using 10-fold cross-validation
- All analysis performed through the WEKA tool created at the University of Waikato
What Models To Use?

- Clearly a binary classification problem (s vs b)
- Decided to go with:
  - Logistic Regression
  - Naive Bayes
  - Decision Tree
- Also performed two types of feature selection:
  - PCA (selected the first 14 features)
  - Information Gain (ranked features 1, 2, 3, 14, 12, 11, 5, 6, 7, 10, 13, 25, 4, 30, 23, 24, 22, 28, 8, 20, 26, 18, 27 & 29)
What Models To Use?

- Clearly a binary classification problem (s vs b)
- Decided to go with:
  - Logistic Regression
  - Naive Bayes
  - Decision Tree
- Also performed two types of feature selection:
  - PCA (selected the first 14 features)
  - Information Gain (ranked features 1, 2, 3, 14, 12, 11, 5, 6, 7, 10, 13, 25, 4, 30, 23, 24, 22, 28, 8, 20, 26, 18, 27 & 29)
- Performance evaluated using 10-fold cross-validation
What Models To Use?

- Clearly a binary classification problem (s vs b)
- Decided to go with:
  - Logistic Regression
  - Naive Bayes
  - Decision Tree
- Also performed two types of feature selection:
  - PCA (selected the first 14 features)
  - Information Gain (ranked features 1, 2, 3, 14, 12, 11, 5, 6, 7, 10, 13, 25, 4, 30, 23, 24, 22, 28, 8, 20, 26, 18, 27 & 29)
- Performance evaluated using 10-fold cross-validation
- All analysis performed through the WEKA tool created at the University of Waikato
Results!

Evaluating based on percent incorrectly classified and classification matrix:

\[
\begin{bmatrix}
\text{found signal} & \text{missed signal} \\
\text{false signal} & \text{true background}
\end{bmatrix}
\]

- Without feature selection:
  - Classification matrix:
    \[
    \begin{bmatrix}
    45653 & 40014 \\
    22341 & 141992
    \end{bmatrix},
    \% \text{incorrect} = 24.94\%
    \]

- With PCA:
  - Classification matrix:
    \[
    \begin{bmatrix}
    42300 & 43367 \\
    20212 & 144121
    \end{bmatrix},
    \% \text{incorrect} = 25.43\%
    \]

- With Info Gain:
  - Classification matrix:
    \[
    \begin{bmatrix}
    42300 & 43367 \\
    20212 & 144121
    \end{bmatrix},
    \% \text{incorrect} = 24.93\%
    \]
Results!

Evaluating based on percent incorrectly classified and classification matrix:

\[
\begin{bmatrix}
\text{found signal} & \text{missed signal} \\
\text{false signal} & \text{true background}
\end{bmatrix}
\]

▶ Logistic Regression
Results!

Evaluating based on percent incorrectly classified and classification matrix:

\[
\begin{bmatrix}
\text{found signal} & \text{missed signal} \\
\text{false signal} & \text{true background}
\end{bmatrix}
\]

- **Logistic Regression**
  - Without feature selection:
    
    Classification matrix \[
    \begin{bmatrix}
    45653 & 40014 \\
    22341 & 141992
    \end{bmatrix}, \text{ % incorrect } = 24.94\]
Results!

Evaluating based on percent incorrectly classified and classification matrix:

\[
\begin{bmatrix}
\text{found signal} & \text{missed signal} \\
\text{false signal} & \text{true background}
\end{bmatrix}
\]

- **Logistic Regression**
  - Without feature selection:
    
    Classification matrix
    \[
    \begin{bmatrix}
    45653 & 40014 \\
    22341 & 141992
    \end{bmatrix}, \quad \% \text{incorrect} = 24.94
    \]
  
  - With PCA:
    
    Classification matrix
    \[
    \begin{bmatrix}
    42300 & 43367 \\
    20212 & 144121
    \end{bmatrix}, \quad \% \text{incorrect} = 25.43
    \]
Results!

Evaluating based on percent incorrectly classified and classification matrix:

\[
\begin{bmatrix}
\text{found signal} & \text{missed signal} \\
\text{false signal} & \text{true background}
\end{bmatrix}
\]

- **Logistic Regression**
  - Without feature selection:
    
    Classification matrix \[
    \begin{bmatrix}
    45653 & 40014 \\
    22341 & 141992
    \end{bmatrix},
    \%
    \text{incorrect} = 24.94
    \]
  
  - With PCA:
    
    Classification matrix \[
    \begin{bmatrix}
    42300 & 43367 \\
    20212 & 144121
    \end{bmatrix},
    \%
    \text{incorrect} = 25.43
    \]
  
  - With Info Gain:
    
    Classification matrix \[
    \begin{bmatrix}
    42300 & 43367 \\
    20212 & 144121
    \end{bmatrix},
    \%
    \text{incorrect} = 24.93
Results! (cont.)

- Naive Bayes

  - Without feature selection:
    - Classification matrix:
      \[
      \begin{bmatrix}
      46857 & 38810 \\
      42130 & 122203 \\
      \end{bmatrix},
      \%	ext{ incorrect} = 32.38
      \]

  - With PCA:
    - Classification matrix:
      \[
      \begin{bmatrix}
      53515 & 32152 \\
      43709 & 120624 \\
      \end{bmatrix},
      \%	ext{ incorrect} = 30.34
      \]

  - With Info Gain:
    - Classification matrix:
      \[
      \begin{bmatrix}
      47097 & 38570 \\
      42403 & 121930 \\
      \end{bmatrix},
      \%	ext{ incorrect} = 32.39
      \]
Results! (cont.)

- Naive Bayes
  - Without feature selection:

    Classification matrix
    \[
    \begin{bmatrix}
    46857 & 38810 \\
    42130 & 122203
    \end{bmatrix}, \quad \% \text{incorrect} = 32.38
    \]
Results! (cont.)

- Naive Bayes
  - Without feature selection:
    
    $$\text{Classification matrix} = \begin{bmatrix} 46857 & 38810 \\ 42130 & 122203 \end{bmatrix}, \quad \% \text{incorrect} = 32.38$$

  - With PCA:
    
    $$\text{Classification matrix} = \begin{bmatrix} 53515 & 32152 \\ 43709 & 120624 \end{bmatrix}, \quad \% \text{incorrect} = 30.34$$
Results! (cont.)

- Naive Bayes
  - Without feature selection:
    
    Classification matrix \[
    \begin{bmatrix}
    46857 & 38810 \\
    42130 & 122203 \\
    \end{bmatrix},
    \%
    \text{incorrect} = 32.38
    \]

  - With PCA:
    
    Classification matrix \[
    \begin{bmatrix}
    53515 & 32152 \\
    43709 & 120624 \\
    \end{bmatrix},
    \%
    \text{incorrect} = 30.34
    \]

  - With Info Gain:
    
    Classification matrix \[
    \begin{bmatrix}
    47097 & 38570 \\
    42403 & 121930 \\
    \end{bmatrix},
    \%
    \text{incorrect} = 32.39
    \]
Results! (cont. II)

- Decision Tree

Without feature selection:

Classification matrix

\[
\begin{bmatrix}
53345 & 32322 \\
16938 & 147395
\end{bmatrix},
\%	ext{incorrect} = 19.70
\]

With PCA:

Classification matrix

\[
\begin{bmatrix}
52605 & 33062 \\
16158 & 148175
\end{bmatrix},
\%	ext{incorrect} = 19.69
\]

With Info Gain:

Classification matrix

\[
\begin{bmatrix}
52605 & 33062 \\
16158 & 148175
\end{bmatrix},
\%	ext{incorrect} = 19.69
\]
Results! (cont. II)

- Decision Tree
  - Without feature selection:
    
    Classification matrix \[
    \begin{bmatrix}
    53345 & 32322 \\
    16938 & 147395 \\
    \end{bmatrix}, \%
    \text{ incorrect} = 19.70
    \]
Results! (cont. II)

- **Decision Tree**
  - Without feature selection:
    
    Classification matrix $\begin{bmatrix} 53345 & 32322 \\ 16938 & 147395 \end{bmatrix}$, % incorrect = 19.70

  - With PCA:
    
    Classification matrix $\begin{bmatrix} 52605 & 33062 \\ 16158 & 148175 \end{bmatrix}$, % incorrect = 19.69
Results! (cont. II)

- **Decision Tree**
  - Without feature selection:
    
    Classification matrix
    \[
    \begin{bmatrix}
    53345 & 32322 \\
    16938 & 147395
    \end{bmatrix}, \text{ % incorrect } = 19.70
    \]

  - With PCA:
    
    Classification matrix
    \[
    \begin{bmatrix}
    52605 & 33062 \\
    16158 & 148175
    \end{bmatrix}, \text{ % incorrect } = 19.69
    \]

  - With Info Gain:
    
    Classification matrix
    \[
    \begin{bmatrix}
    52605 & 33062 \\
    16158 & 148175
    \end{bmatrix}, \text{ % incorrect } = 19.69
    \]