

---

# Explanations for Multinomial Classifiers

Tips and Tricks for Practitioners

---

**Pramit Choudhary\***  
Los Angeles, CA  
pramit.choudhary@h2o.ai

**Navdeep Gill\***  
Mountain View, CA  
navdeep.gill@h2o.ai

**Patrick Hall†**  
Washington, DC  
patrick.hall@h2o.ai

## Abstract

### 1 Introduction

### 2 Notation

To facilitate technical descriptions of explanatory techniques, notation for input and output spaces, datasets, and models is defined.

#### 2.1 Spaces

- Input features come from the set  $\mathcal{X}$  contained in a  $P$ -dimensional input space,  $\mathcal{X} \subset \mathbb{R}^P$ .
- Known labels corresponding to instances of  $\mathcal{X}$  come from the set  $\mathcal{Y}$  contained in a  $C$ -dimensional input space,  $\mathcal{Y} \subset \mathbb{R}^C$ .
- Learned output responses come from the set  $\hat{\mathcal{Y}}$ . For classification models the set  $\hat{\mathcal{Y}}$  typically contains a column vector for each unique class in  $\mathcal{Y}$ . In this text, the space  $\hat{\mathcal{Y}}$  is said to be contained in a  $C'$ -dimensional output space,  $\hat{\mathcal{Y}} \subset \mathbb{R}^{C'}$ .

#### 2.2 Datasets

- The input dataset  $\mathbf{X}$  is composed of observed instances of the set  $\mathcal{X}$  with a corresponding dataset of labels  $\mathbf{Y}$ , observed instances of the set  $\mathcal{Y}$ .
- Each  $i$ -th observation of  $\mathbf{X}$  is denoted as  $\mathbf{x}^{(i)} = [x_0^{(i)}, x_1^{(i)}, \dots, x_{P-1}^{(i)}]$ , with corresponding  $i$ -th labels in  $\mathbf{Y}$ ,  $\mathbf{y}^{(i)} = [y_0^{(i)}, y_1^{(i)}, \dots, y_{C-1}^{(i)}]$ , and corresponding predictions in  $\hat{\mathbf{Y}}$ ,  $\hat{\mathbf{y}}^{(i)} = [\hat{y}_0^{(i)}, \hat{y}_1^{(i)}, \dots, \hat{y}_{C'-1}^{(i)}]$ .
- $\mathbf{X}$  and  $\mathbf{Y}$  consist of  $N$  tuples of observations:  $[(\mathbf{x}^{(0)}, \mathbf{y}^{(0)}), (\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \dots, (\mathbf{x}^{(N-1)}, \mathbf{y}^{(N-1)})]$ .
- Each  $j$ -th input column vector of  $\mathbf{X}$  is denoted as  $X_j = [x_j^{(0)}, x_j^{(1)}, \dots, x_j^{(N-1)}]^T$ .

---

\*H2O.ai

†H2O.ai and George Washington University

## 2.3 Models

- A type of machine learning model  $g$ , selected from a hypothesis set  $\mathcal{H}$ , is trained to represent an unknown signal-generating function  $f$  observed as  $\mathbf{X}$  with labels  $\mathbf{Y}$  using a training algorithm  $\mathcal{A}$ :
- $g$  generates learned output responses on the input dataset  $g(\mathbf{X}) = \hat{\mathbf{Y}}$ , and on the general input space  $g(\mathcal{X}) = \hat{\mathcal{Y}}$ .
- The model to be explained is denoted as  $g$ .

## 3 Global Analysis

### 3.1 Decision Tree Surrogate

### 3.2 Decision Boundary Plots

### 3.3 Shapley Global Feature Importance

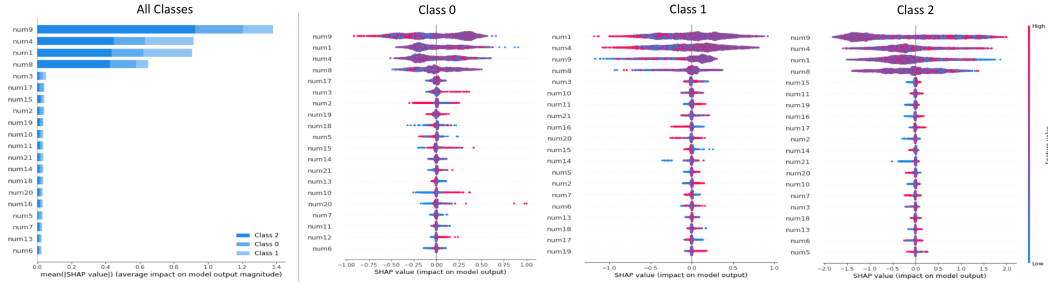


Figure 1:

### 3.4 Partial Dependence and ICE

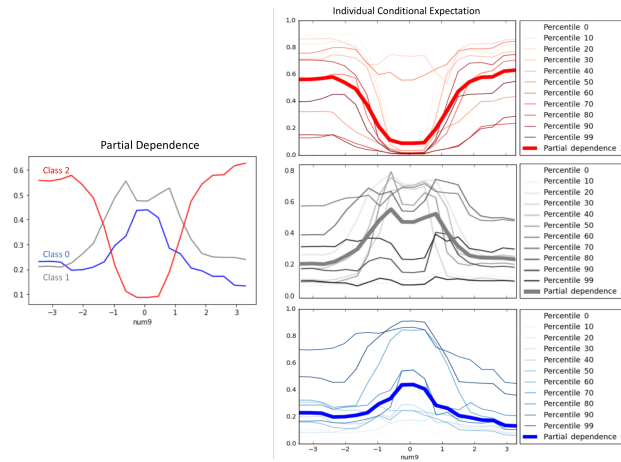


Figure 2:

### 3.5 Shapley Local Feature Importance

## 4 Supplementary Materials

## 5 Conclusion

## 6 NIPS Style examples

**Paragraphs** There is also a `\paragraph` command available, which sets the heading in bold, flush left, and inline with the text, with the heading followed by 1 em of space.

### 6.1 Citations, figures, tables, references

These instructions apply to everyone.

### 6.2 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

<http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf>

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

Hasselmo, et al. (1995) investigated...

If you wish to load the `natbib` package with options, you may add the following before loading the `nips_2018` package:

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{nips_2018}
```

As submission is double blind, refer to your own published work in the third person. That is, use “In the previous work of Jones et al. [4],” not “In our previous work [4].” If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form “A. Anonymous.”

**New preprint option for 2018** If you wish to post a preprint of your work online, e.g., on arXiv, using the NIPS style, please use the `preprint` option. This will create a nonanonymized version of your work with the text “Preprint. Work in progress.” in the footer. This version may be distributed as you see fit. Please **do not** use the `final` option, which should **only** be used for papers accepted to NIPS.

At submission time, please omit the `final` and `preprint` options. This will anonymize your submission and add line numbers to aid review. Please *do not* refer to these line numbers in your paper as they will be removed during generation of camera-ready copies.

The file `nips_2018.tex` may be used as a “shell” for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.



Figure 3: Sample figure caption.

Table 1: Sample table title

Part		
Name	Description	Size ( $\mu\text{m}$ )
Dendrite	Input terminal	$\sim 100$
Axon	Output terminal	$\sim 10$
Soma	Cell body	up to $10^6$

The formatting instructions contained in these style files are summarized in Sections ??, ??, and 6.1 below.

### 6.3 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number<sup>3</sup> in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.<sup>4</sup>

### 6.4 Figures

### 6.5 Tables

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules*. We strongly suggest the use of the booktabs package, which allows for typesetting high-quality, professional tables:

<https://www.ctan.org/pkg/booktabs>

This package was used to typeset Table 1.

The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

```
\usepackage{amsfonts}
```

followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for  $\mathbb{R}$ ,  $\mathbb{N}$  or  $\mathbb{C}$ . You can also use the following workaround for reals, natural and complex:

```
\newcommand{\RR}{\mathbb{R}} %real numbers
\newcommand{\Nat}{\mathbb{N}} %natural numbers
\newcommand{\CC}{\mathbb{C}} %complex numbers
```

---

<sup>3</sup>Sample of the first footnote.

<sup>4</sup>As in this example.

Note that `amsmath` is automatically loaded by the `amssymb` package.

### **Acknowledgments**

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.