Introduction to supervised machine learning, k-fold cross-validation, nearest neighbors, and linear models

Toby Dylan Hocking

Supervised machine learning

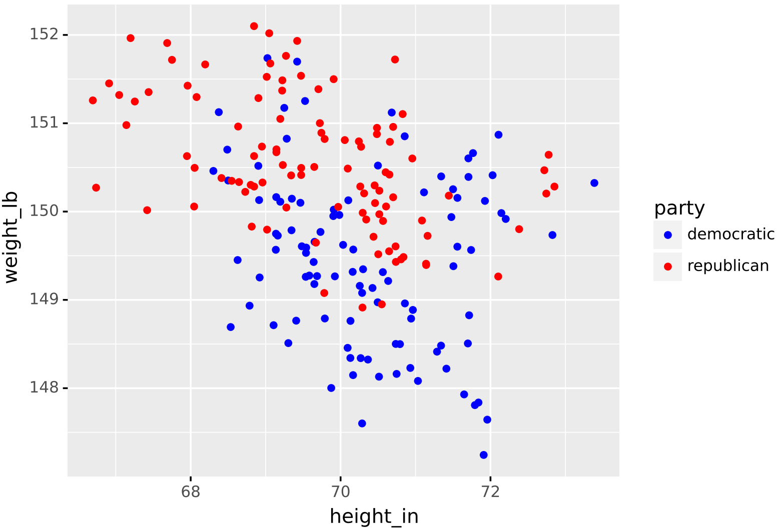
- Goal is to learn a function $f(\mathbf{x}) = y$ where $\mathbf{x}$ is an input/feature vector and y is an output/label.
- $x =$image of digit/clothing, $y \in \{0, \dots, 9\}$ (ten classes).
- $x =$vector of word counts in email, $y \in \{1, 0\}$ (spam or not).
- $x =$image of retina, $y =$risk score for heart disease.
- This week we will focus on a specific kind of supervised learning problem called binary classification, which means $y \in \{1, 0\}$.

Learning algorithm

- We want a learning algorithm LEARN which inputs a training data set and outputs a prediction function f.
- We typically represent a training data set with n observations and p features as a matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ with a corresponding label vector $\mathbf{y} \in \{0,1\}^n$.
- We will use three such data sets from Elements of Statistical Learning book by Hastie et al. (mixture slightly modified)

name	observations, n	inputs/features, p	outputs/labels
zip.test	images, 623	pixel intensities, 256	0/1 digits
spam	emails, 4601	word counts, 57	spam=1/not=0
mixture	people, 200	height/weight, 2	democratic/republican

Visualize iris data with labels



Visualize iris data without labels

- Let $X \in \mathbb{R}^{150\times 2}$ be the data matrix (input for clustering).