Automating Tidyverse functionality

Dr. Henry WJ Reeve

Teaching block 1 2021

Introduction

This document describes your ninth assignment for Statistical Computing and Empirical Methods (Unit EMATM0061) on the MSc in Data Science. Before starting the assignment it is recommended that you first watch video lectures 21, 22 and 23.

1 Basic concepts in classification

Write down your explanation of each of the following concepts. Give an example where appropriate.

1. A classification rule
2. A learning algorithm
3. Training data
4. Feature vector
5. Label
6. Test error
7. Train error
8. The train test split
9. Linear classifier

2 The train test split

Suppose you want to build a classifier to predict whether a hawk belongs to either the "Sharp-shinned" or the "Cooper's" species of hawks. The feature vector will be a four dimensional row vector containing the weight, and the lengths of the wing, the tail and the hallux. The labels will be binary - 1 if the hawk is "Sharp-shinned" and 0 if the hawk belongs to "Cooper's" species.

Begin by loading the "Hawks" data frame from the "Stat2Data" library. Now extract a subset of the data called "hawks_total" with five columns - "Weight,"Wing","Hallux","Tail" and "Species". The data frame should only include rows corresponding to hawks from either the "Sharp-shinned" or the "Cooper's" species, and not the "Red-tailed" species. Convert the Species column to a binary variable with a 1 if the hawk belongs to the sharp-shinned species and 0 if the hawk belongs to the Cooper's species. Finally remove any rows with missing values from one of the relevant columns.

Now implement a train test split for your "hawks_total" data frame. You should use 60% of your data within your training data and 40% in your test data. You should create a data frame consisting of training data called "hawks_train" and a data frame consisting of test data called "hawks_test". Display the number of rows in each data frame.

Next extract a data frame called "hawks_train_x" from your training data (from "hawks_train") containing the feature vectors and no labels. In addition extract a vector called "hawks_train_y" consisting of labels from your training data. Similarly, create data frames called "hawks_test_x" and "hawks_test_y" corresponding to the feature vectors and labels within the test set, respectively.

Now let's consider a very simple (and not very effective) classifier which entirely ignores the feature vectors. Instead the classifier simply predicts a single fixed value $y \in \{0, 1\}$. Hence, your classifier is of the form $\phi_y(x) \equiv y$ for all $x \in \mathbb{R}^4$. Begin by choosing a value $\hat{y} \in \{0, 1\}$ based on your training data - choose the value which minimises the training error.

Next compute the train and test error of $\phi_{\hat{y}}$

What does this tell you about the relative sizes of your classes?

3 Linear discriminant analysis

Describe the probabilistic model that underpins linear discriminant analysis.

Train a linear discriminant analysis model to carry out the classification task described above. That is, to predict whether a hawk belongs to either the "Sharp-shinned" or the "Cooper's" species of hawks, based on a four-dimensional feature vector containing the weight, and the lengths of the wing, the tail and the hallux.

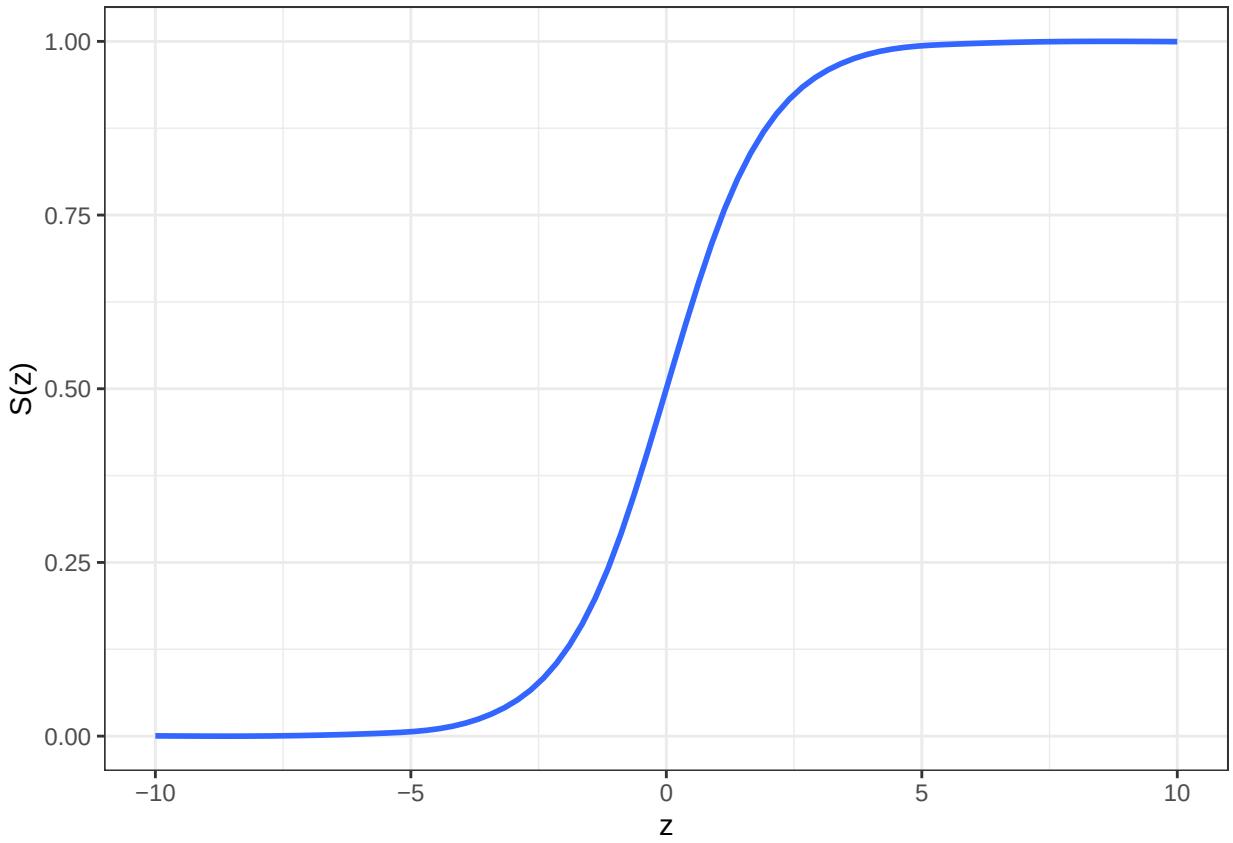
Compute and report the train error and the test error.

As a challenging optional extra implement your own linear discriminant analysis model.

4 Logistic regression

Describe the probabilistic model which underpins logistic regression.

Recall that the sigmoid function $\mathrm{S} : \mathbb{R} \to (0, 1)$ is defined by $\mathrm{S}(z) = 1/(1 + e^{-z})$. Generate the following plot which displays the sigmoid function:



Now train a logistic regression model to predict whether a hawk belongs to either the "Sharp-shinned" or the "Cooper's" species of hawks, based on a four-dimensional feature vector containing the weight, and the lengths of the wing, the tail and the hallux.

Compute and report both the training error and the test error.

As an optional extra consider the following formula for the log-likelihood of the weights $w \in \mathbb{R}^d$ and bias $w^0 \in \mathbb{R}$, given data $\mathcal{D} = ((X_1, Y_1), \cdots, (X_n, Y_n))$:

$$\log \ell(w, w^0) = \sum_{i=1}^{n} \log \mathrm{S}\left((2Y_i - 1) \cdot (wX_i^\top + w^0)\right)$$

Demonstrate the following formulas for the derivatives:

$$\frac{\partial}{\partial w} \log \ell(w, w^0) = \sum_{i=1}^{n} (2Y_i - 1)\mathrm{S}\left((1 - 2Y_i) \cdot (wX_i^\top + w^0)\right) X_i,$$
$$\frac{\partial}{\partial w^0} \log \ell(w, w^0) = \sum_{i=1}^{n} (2Y_i - 1)\mathrm{S}\left((1 - 2Y_i) \cdot (wX_i^\top + w^0)\right).$$

Explain the role the above formula has in training a logistic regression model.

You can learn more about the the glmnet approach to logistic regression here.