Exercise 3 – Classification I

Introduction to Machine Learning

Exercise 1: Logistic vs softmax regression

This exercise is only for lecture group A

Learning goals

Solve "show equivalence"-type questions

Binary logistic regression is a special case of multiclass logistic, or softmax, regression. The softmax function is the multiclass analogue to the logistic function, transforming scores $\theta^\top \mathbf{x}$ to values in the range [0, 1] that sum to one. The softmax function is defined as:

$$\pi_k(\mathbf{x}|\theta) = \frac{\exp(\theta_k^\top \mathbf{x})}{\sum_{j=1}^{g} \exp(\theta_j^\top \mathbf{x})}, k \in \{1, ..., g\}$$

Show that logistic and softmax regression are equivalent for $g = 2$.

Solution

As we would expect, the two formulations are equivalent (up to reparameterization). In order to see this, consider the softmax function components for both classes:

$$\pi_1(\mathbf{x}|\theta) = \frac{\exp(\theta_1^\top \mathbf{x})}{\exp(\theta_1^\top \mathbf{x}) + \exp(\theta_2^\top \mathbf{x})}$$

$$\pi_2(\mathbf{x}|\theta) = \frac{\exp(\theta_2^\top \mathbf{x})}{\exp(\theta_1^\top \mathbf{x}) + \exp(\theta_2^\top \mathbf{x})}$$

Since we know that $\pi_1(\mathbf{x}|\theta) + \pi_2(\mathbf{x}|\theta) = 1$, it is sufficient to compute one of the two scoring functions. Let's pick $\pi_1(\mathbf{x}|\theta)$ and relate it to the logistic function:

$$\pi_1(\mathbf{x}|\theta) = \frac{1}{1 + \exp(\theta_2^\top \mathbf{x} - \theta_1^\top \mathbf{x})} = \frac{1}{1 + \exp(-\theta^\top \mathbf{x})}$$

where $\theta := \theta_1 - \theta_2$. Thus, we obtain the binary-case logistic function, reflecting that we only need one scoring function (and thus one set of parameters θ rather than two θ_1, θ_2).

Exercise 2: Hyperplanes

Learning goals

1. Understand that hyperplanes bisect the space with a linear boundary
2. Get a feeling for coefficients in hyperplane equations

Linear classifiers like logistic regression learn a decision boundary that takes the form of a (linear) hyperplane. Hyperplanes are defined by equations $\theta^\top \mathbf{x} = b$ with coefficients θ and a scalar $b \in \mathbb{R}$.

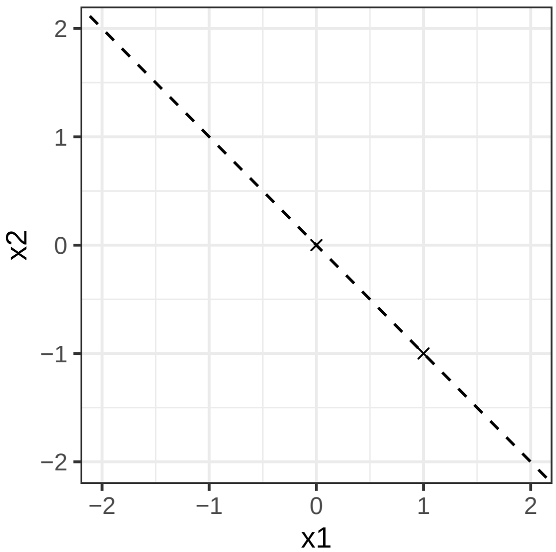
In order to see that such expressions actually describe hyperplanes, consider $\theta^\top \mathbf{x} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$. Sketch the hyperplanes given by the following coefficients and explain the difference between the parameterizations:

- $\theta_0 = 0, \theta_1 = \theta_2 = 1$
- $\theta_0 = 1, \theta_1 = \theta_2 = 1$
- $\theta_0 = 0, \theta_1 = 1, \theta_2 = 2$

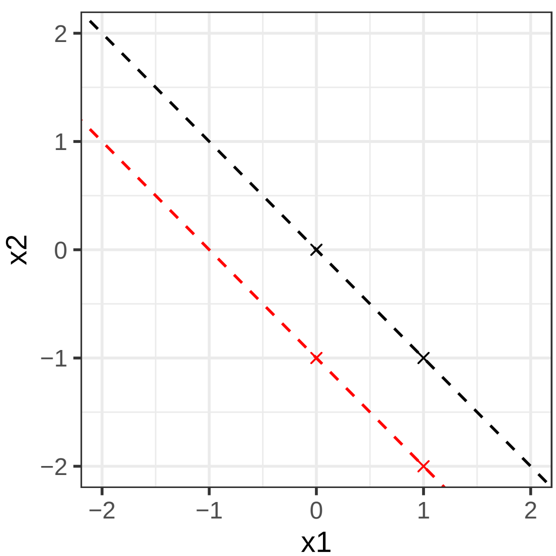
Solution

A hyperplane in 2D is just a line. We know that two points are sufficient to describe a line, so all we need to do is pick two points fulfilling the hyperplane equation.

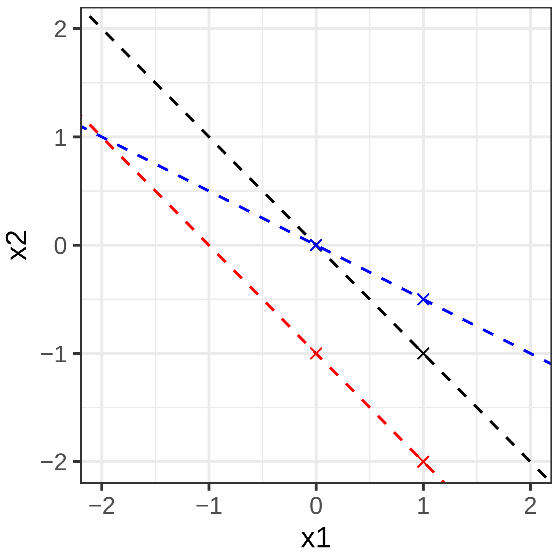
- $\theta_0 = 0, \theta_1 = \theta_2 = 1 \rightsquigarrow$ e.g., (0, 0) and (1, -1). Sketch it:



- $\theta_0 = 1, \theta_1 = \theta_2 = 1 \rightsquigarrow$ e.g., (0, -1) and (1, -2). The change in θ_0 promotes a horizontal shift:



- $\theta_0 = 0, \theta_1 = 1, \theta_2 = 2 \rightsquigarrow$ e.g., (0, 0) and (1, -0.5). The change in θ_2 pivots the line around the intercept:



We see that a hyperplane is defined by the points that lie directly on it and thus fulfill the hyperplane equation.

Exercise 3: Decision Boundaries & Thresholds in Logisitc Regression

Learning goals

1) Understand that logistic regression finds a linear decision boundary
2) Get a feeling for how parameterization changes predicted probabilities

In logistic regression (binary case), we estimate the probability $p(y = 1|\mathbf{x}, \theta) = \pi(\mathbf{x}|\theta)$. In order to decide about the class of an observation, we set $\hat{y} = 1$ iff $\pi(\mathbf{x}|\theta) \geq \alpha$ for some $\alpha \in (0, 1)$.

Show that the decision boundary of the logistic classifier is a (linear) hyperplane.

Hint

Derive the value of $\theta^\top \mathbf{x}$ (depending on α) starting from which you predict $\hat{y} = 1$ rather than $\hat{y} = 0$.

Solution

We evaluate

$$
\begin{aligned}
\pi(\mathbf{x}) = \frac{1}{1+\exp(-\theta^\top \mathbf{x})} &= \alpha \\
\Leftrightarrow\; 1+\exp(-\theta^\top \mathbf{x}) &= \frac{1}{\alpha} \\
\Leftrightarrow\; \exp(-\theta^\top \mathbf{x}) &= \frac{1}{\alpha} - 1 \\
\Leftrightarrow\; -\theta^\top \mathbf{x} &= \log\left(\frac{1}{\alpha} - 1\right) \\
\Leftrightarrow\; \theta^\top \mathbf{x} &= -\log\left(\frac{1}{\alpha} - 1\right).
\end{aligned}
$$

$\theta^\top \mathbf{x} = -\log\left(\frac{1}{\alpha} - 1\right)$ is the equation of the linear hyperplane comprised of all linear combinations $\theta^\top \mathbf{x}$ that are equal to $-\log\left(\frac{1}{\alpha} - 1\right)$. The equation therefore describes the decision rule for setting $\hat{y}$ equal to 1 by taking all points that lie on or above this hyperplane.

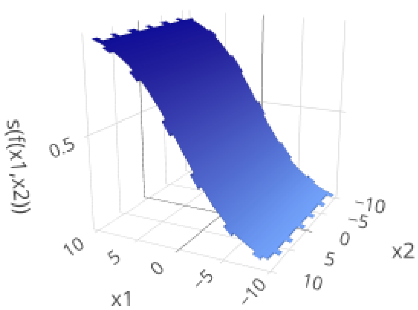
Below you see the logistic function for a binary classification problem with two input features for different values $\theta^\top = (\theta_1, \theta_2)^\top$ (plots 1-3) as well as α (plot 4). What can you deduce for the values of θ_1, θ_2, and α? What are the implications for classification in the different scenarios?

Solution

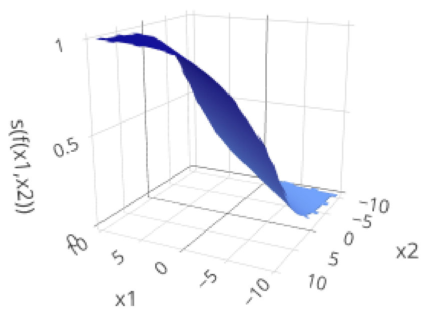
We observe

- in plot (1): the logistic function runs parallel to the x_2 axis, so it is the same for every value of x_2. In other words, x_2 does not contribute anything to the class discrimination and its associated parameter θ_2 is equal to 0.
- in plot (2): both dimensions affect the logistic function – to equal degree in this case, meaning x_1 and x_2 are equally important. If θ_1 were larger than θ_2 or vice versa the hypersurface would be more tilted towards the respective axis. Furthermore, due to θ_1 and θ_2 being positive, $\pi(\mathbf{x})$ increases with higher values for x_1 and x_2.
- in plot (3): this is the same situation as in plot (2) but the logistic function is steeper, which is due to θ_1, θ_2 having larger absolute values. We therefore get a sharper separation between classes (fewer predicted probability values close to 0.5, so we are overall more confident in our decision). As in plot (2), the increasing probability of $\hat{y} = 1$ for higher values of x_1 and x_2 indicates positive values for θ_1 and θ_2.

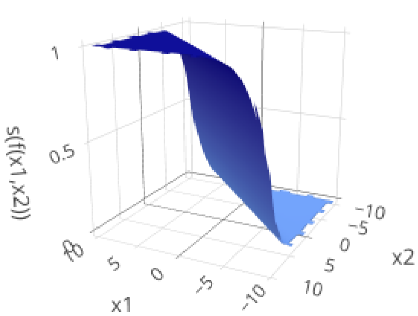
Plot (1)



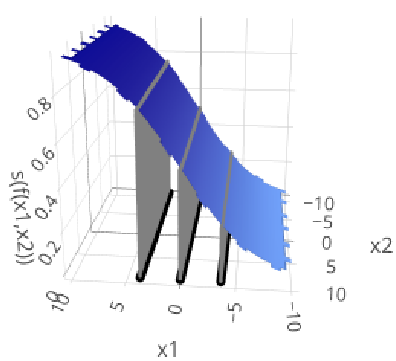
Plot (2)



Plot (3)



Plot (4)



- in plot (4): this is the same situation as in plot (1). The different values for α represent different thresholds: a high value (leftmost line) means we only assign class 1 if the estimated class-1 probability is large. Conversely, a low value (rightmost line) signifies we are ready to predict class 1 at a low threshold – in effect, this is the same as the previous scenario, only the class labels are flipped. The mid line corresponds to the common case $\alpha = 0.5$ where we assign class 1 as soon as the predicted probability is more than 50%.

Derive the equation for the decision boundary hyperplane if we choose $\alpha = 0.5$.

Solution

We make use of our results from a):

$$\begin{aligned}
\hat{y} = 1 &\Leftrightarrow \theta^\top \mathbf{x} \geq -\log\left(\frac{1}{\alpha} - 1\right) \\
&\Leftrightarrow \theta^\top \mathbf{x} \geq -\log\left(\frac{1}{0.5} - 1\right) \\
&\Leftrightarrow \theta^\top \mathbf{x} \geq -\log 1 \\
&\Leftrightarrow \theta^\top \mathbf{x} \geq 0.
\end{aligned}$$

The 0.5 threshold therefore leads to the coordinate hyperplane and divides the input space into the positive "1" halfspace where $\theta^\top \mathbf{x} \geq 0$ and the "0" halfspace where $\theta^\top \mathbf{x} < 0$.

Explain when it might be sensible to set α to 0.5.

Solution

When the threshold $\alpha = 0.5$ is chosen, the losses of misclassified observations, i.e., $L(\hat{y} = 0 \mid y = 1)$ and $L(\hat{y} = 1 \mid y = 0)$, are treated equally, which is often the intuitive thing to do. It means $\alpha = 0.5$ is a sensible threshold if we do not wish to avoid one type of misclassification more than the other. If, however, we need to be cautious to only predict class 1 if we are very confident (for example, when the decision triggers a costly therapy), it would make sense to set the threshold considerably higher.