Solution 1: Logistic vs softmax regression

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:

$$\pi_1(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}_1^\top \mathbf{x})}{\exp(\boldsymbol{\theta}_1^\top \mathbf{x}) + \exp(\boldsymbol{\theta}_2^\top \mathbf{x})}$$
$$\pi_2(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}_2^\top \mathbf{x})}{\exp(\boldsymbol{\theta}_1^\top \mathbf{x}) + \exp(\boldsymbol{\theta}_2^\top \mathbf{x})}.$$

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

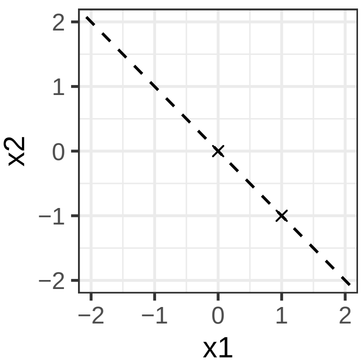
$$\begin{aligned}
\pi_1(\mathbf{x} \mid \boldsymbol{\theta}) &= \frac{1}{\frac{\exp(\boldsymbol{\theta}_1^\top \mathbf{x}) + \exp(\boldsymbol{\theta}_2^\top \mathbf{x})}{\exp(\boldsymbol{\theta}_1^\top \mathbf{x})}} \\
&= \frac{1}{1 + \exp(\boldsymbol{\theta}_2^\top \mathbf{x} - \boldsymbol{\theta}_1^\top \mathbf{x})} \\
&= \frac{1}{1 + \exp(-\boldsymbol{\theta}^\top \mathbf{x})} \\
&= \pi(\mathbf{x} \mid \boldsymbol{\theta}),
\end{aligned}$$

i.e., we obtain the binary-case logistic function if we set $\boldsymbol{\theta} := \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$, reflecting that we only need one scoring function (and thus one set of parameters $\boldsymbol{\theta}$ rather than two $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$).

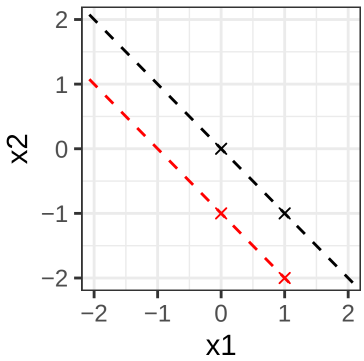
Solution 2: Hyperplanes

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.

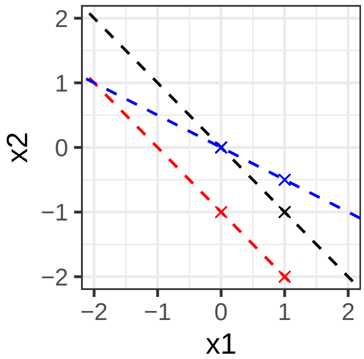
- $\theta_0 = 0, \theta_1 = \theta_2 = 1 \rightsquigarrow$ e.g., (0, 0) and (1, -1).
  Sketch it:



- $\theta_0 = 1, \theta_1 = \theta_2 = 1 \rightsquigarrow$ 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.

Solution 3: Decision Boundaries & Thresholds in Logistic Regression

a) We evaluate

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

$\boldsymbol{\theta}^\top \mathbf{x} = -\log\left(\frac{1}{\alpha} - 1\right)$ is the equation of the linear hyperplane comprised of all linear combinations $\boldsymbol{\theta}^\top \mathbf{x}$ 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.

b) 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.
- 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%.

c) We make use of our results from a):

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

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

d) 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.