Exercise 1: Logistic vs Softmax Regression [only for lecture group A]

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 $\boldsymbol{\theta}^\top \mathbf{x}$ to values $\in [0, 1]$ that sum to one:

$$\pi_k(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}_k^\top \mathbf{x})}{\sum_{j=1}^{g} \exp(\boldsymbol{\theta}_j^\top \mathbf{x})}, \quad k \in \{1, \ldots, g\}$$

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

Exercise 2: Hyperplanes [only for lecture group B]

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

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

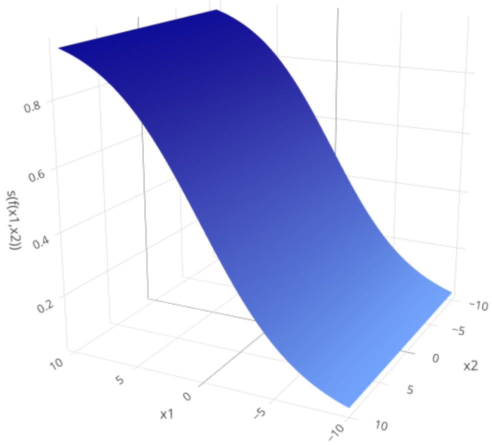
- $\theta_0 = 0, \theta_1 = \theta_2 = 1$
- $\theta_0 = 1, \theta_1 = \theta_2 = 1$
- $\theta_0 = 0, \theta_1 = 1, \theta_2 = 2$

Exercise 3: Decision Boundaries & Thresholds in Logistic Regression

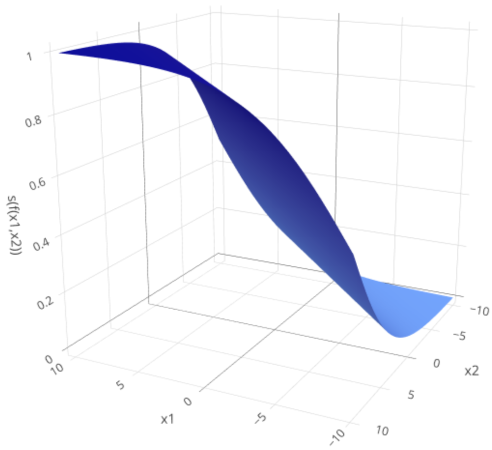
In logistic regression (binary case), we estimate the probability $\mathbb{P}(y = 1 \mid \mathbf{x}, \boldsymbol{\theta}) = \pi(\mathbf{x} \mid \boldsymbol{\theta})$. In order to decide about the class of an observation, we set $\hat{y} = 1$ iff $\pi(\mathbf{x} \mid \boldsymbol{\theta}) \geq \alpha$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

a) Show that the decision boundary of the logistic classifier is a (linear) hyperplane.
Hint: derive the value of $\boldsymbol{\theta}^\top \mathbf{x}$ (depending on α) starting from which you predict $\hat{y} = 1$ rather than $\hat{y} = 0$.

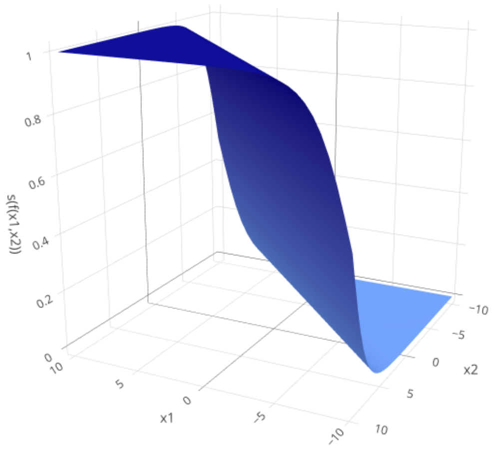
b) Below you see the logistic function for a binary classification problem with two input features for different values $\boldsymbol{\theta} = (\theta_1, \theta_2)$ (plots 1-3) as well as α (plot 4). What can you deduce for the values of θ_1, θ_2 and α? What are the implications for classification in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



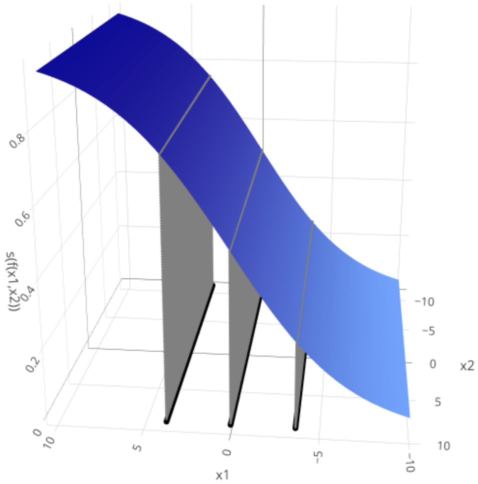
Plot (1)



Plot (2)



Plot (3)



Plot (4)

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

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