Exercise 1: Logistic Regression Basics

a) What is the relationship between softmax

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

and the logistic function

$$\pi(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^T \mathbf{x})}$$

for $g = 2$ (binary classification)?

b) The likelihood function for a multinomially distributed target variable with g target classes is given by[1]

$$\mathcal{L}_i(\boldsymbol{\theta}) = \mathbb{P}(y^{(i)} | \mathbf{x}^{(i)}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_g) = \prod_{j=1}^{g} \pi_j \left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)^{\mathbb{I}(y^{(i)} = j)}$$

where the posterior class probabilities $\pi_1\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right), \pi_2\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right), \ldots, \pi_g\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)$ are modeled with softmax regression. Derive the likelihood function for n independent observations.

c) We have already addressed the connection that holds between maximum likelihood estimation and empirical risk minimization. Transform the joint likelihood function into an empirical risk function.

Hints:

- By following the maximum likelihood principle, we should look for parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_g$ that maximize the likelihood function.
- The expressions $\prod \mathcal{L}_i$ and $\log \prod \mathcal{L}_i$, if defined, are maximized by the same parameters.
- Minimizing a scalar function multiplied with -1 is equivalent to maximizing the original function.

State the associated risk function.

d) Write down the discriminant functions of multiclass logistic regression resulting from this minimization objective. How do we arrive at the final prediction?

e) State the parameter space Θ and corresponding hypothesis space $\mathcal{H}$ for the multiclass case.

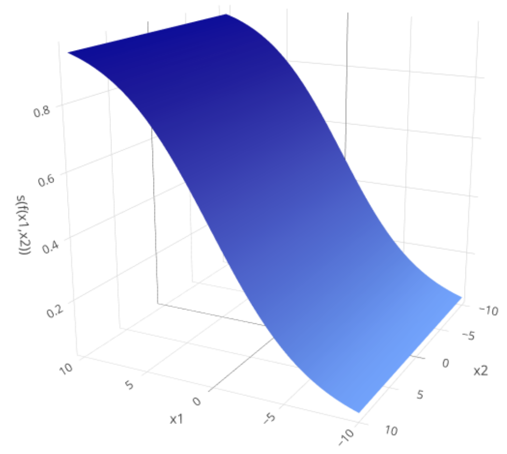
Exercise 2: Decision Boundaries & Thresholds in Logistic Regression

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 $\hat{\pi}(\mathbf{x} \mid \hat{\boldsymbol{\theta}}) \geq \alpha$ for some $\alpha \in (0, 1)$.

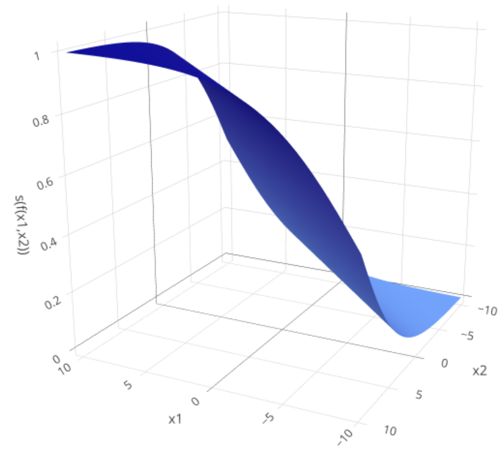
a) Show that the decision boundary of the logistic classifier is a (linear!) hyperplane.
Hint: derive the value of $\hat{\boldsymbol{\theta}}^T \mathbf{x}$ (depending on α) starting from which you predict $\hat{y} = 1$ rather than $\hat{y} = 0$.

[1] While this might look somewhat complicated, it is actually just a very concise way to express the multinomial likelihood: for each observation, all factors but the one corresponding to the true class j' will be 1 (due to the 0 exponent), so the result is simply $\pi_{j'}\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)$.

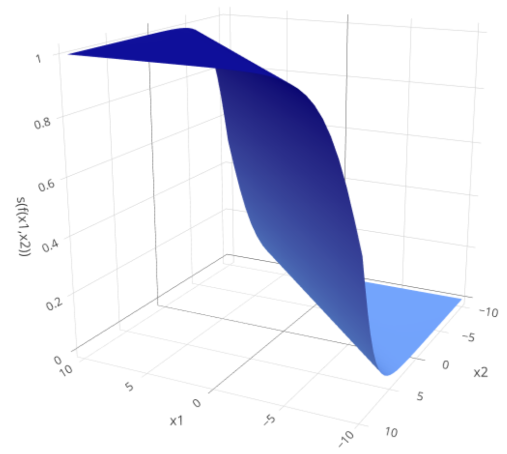
b) Below you see the logistic function for a binary classification problem with two input features for different values $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2)$ (plots 1-3) as well as α (plot 4). What can you deduce for the values of $\hat{\theta}_1$, $\hat{\theta}_2$ and α? What are the implications for classification in the different scenarios?



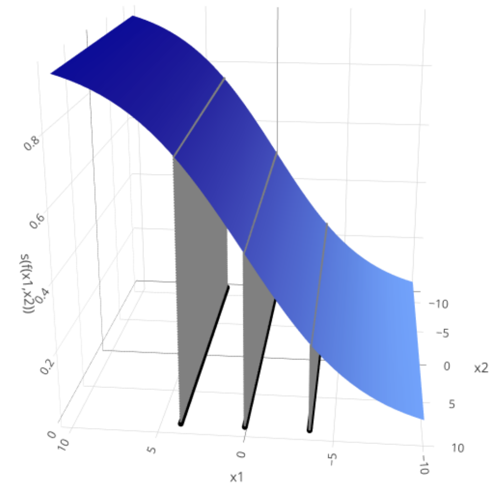
Plot (1)



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.