Exercise 1: Logistic Regression Basics

a) What is the relationship between softmax

$$\pi_k(\mathbf{x}) = \frac{\exp(\boldsymbol{\theta}_k^T \mathbf{x})}{\sum\limits_{j=1}^{g} \exp(\boldsymbol{\theta}_j^T \mathbf{x})}, \quad k = 1, 2$$

and the logistic function

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

for $g = 2$ (binary classification)?

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

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

where the posterior class probabilities $\pi_1(\mathbf{x}^{(i)}), \pi_2(\mathbf{x}^{(i)}), \ldots, \pi_g(\mathbf{x}^{(i)})$ are modeled with softmax regression. Derive the likelihood function for n such independent target variables.

c) We have already addressed the connection that holds between maximum likelihood estimation and empirical risik minimization. How can you 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 loss 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 hypothesis space $\mathcal{H}$ and corresponding parameter space Θ for the multiclass case.

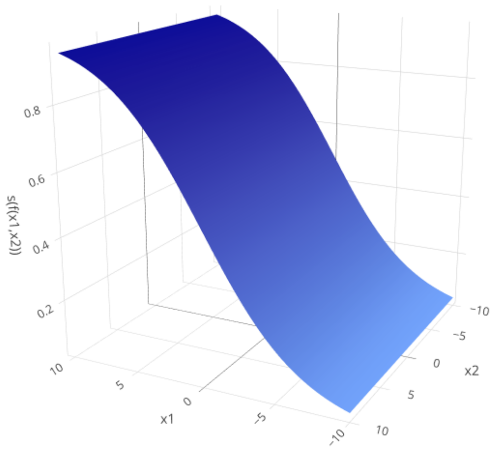
Exercise 2: Decision Boundaries & Thresholds in Logistic Regression

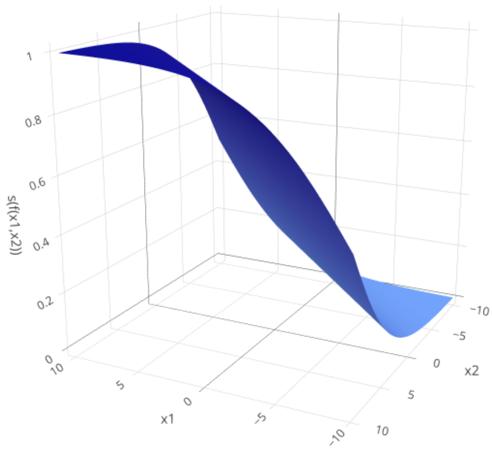
In logistic regression (binary case), we estimate the probability $\pi(\mathbf{x})$. In order to decide about the class of an observation, we set $\hat{y} = 1$ iff $\hat{\pi}(\mathbf{x}) \geq \alpha$ for some $\alpha \in \mathbb{R}$.

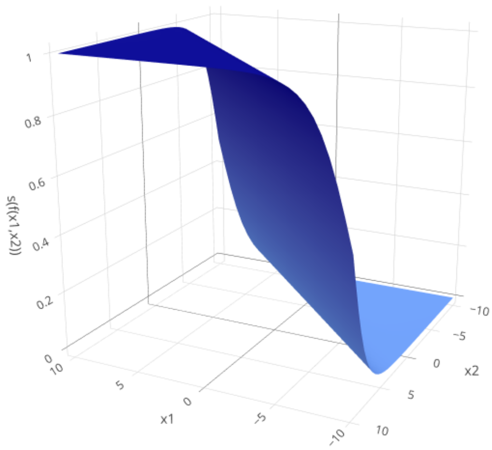
a) Show that the decision boundary of the logistic classifier is a (linear!) hyperplane.
Hint: derive the value of $\boldsymbol{\theta}^T \mathbf{x}$ (depending on α) starting from which you predict $y = 1$ rather than $y = 0$.

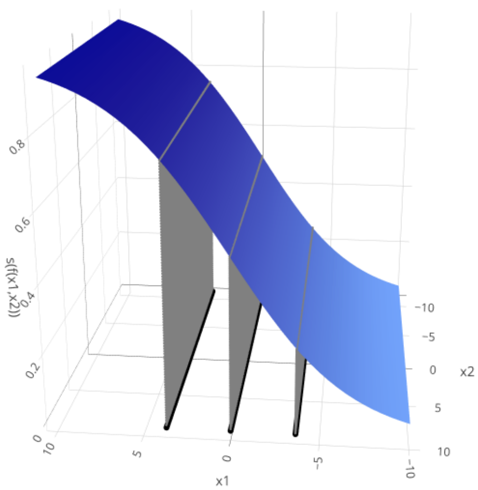
b) Below you see the logistic function for a binary classification problem with two input features for different values $\boldsymbol{\theta} = (\theta_1, \theta_2)$ (plots 1-3) as well as α (plot 4). What can you deduce for the values of θ_1, θ_2 and α? What are the implications for classification in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

[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 $\pi_j(\mathbf{x}^{(i)})$.









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.

Exercise 3: Decision Boundaries for `mlr3` Learners

We will now visualize how well different learners classify the notoriously hard `mlbench::mlbench.spirals` data set. Generate 1000 points from `spirals` (using the default standard deviation) and consider the classifiers already introduced in the lecture:

- logistic regression,
- LDA,
- QDA, and
- Naive Bayes.

Plot their decision boundaries for different settings of relevant hyperparameters (too early to touch upon hyperparameters?). Can you spot differences in the learners' separation ability? To refresh your knowledge about `mlr3` you can take a look at `https://mlr3book.mlr-org.com/basics.html`.