

Exercise 1:

In logistic regression, we estimate the probability $\pi(\mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x})$. To decide if \hat{y} is 0 or 1, we follow:

$$\hat{y} = 1 \Leftrightarrow \hat{\pi}(\mathbf{x}) \geq a$$

- What happens if you are choosing $a = 0.5$? More precisely, from which value of $\theta^T x$ do you predict $\hat{y} = 1$ rather than $\hat{y} = 0$?
- Explain (using words) why $a = 0.5$ is a sensible threshold.

Exercise 2:

- What is the relationship between softmax $\pi_k(x) = \frac{\exp(\theta_k^T x)}{\sum_{j=1}^g \exp(\theta_j^T x)}$ and the logistic function $\pi(\mathbf{x}) = \frac{1}{1 + \exp(\theta^T x)}$ for $g = 2$ (binary classification)?
- The likelihood function of a multinomially distributed target variable with g target classes is given by

$$\mathcal{L}_i = \mathbb{P}(Y^{(i)} = y^{(i)} \mid x^{(i)}, \theta_1, \dots, \theta_g) = \prod_{j=1}^g \pi_j(x^{(i)})^{\mathbb{1}_{\{y^{(i)}=j\}}}$$

where the posterior class probabilities $\pi_1(x), \dots, \pi_g(x)$ are modeled with softmax regression. Derive the likelihood function of n such independent target variables. How can you transform this likelihood function into an empirical risk function?

Hints:

- By following the maximum likelihood principle, we should look for parameters $\theta_1, \dots, \theta_g$, which maximize the likelihood function.
- The expressions $\prod \mathcal{L}_i$ and $\log \prod \mathcal{L}_i$ (if this expression is defined) are maximized by the same parameters.
- The empirical risk is a *sum* of loss function values, not a *product*.
- Minimizing a scalar function multiplied with -1 is equivalent to maximizing the original function.

State the associated loss function.

- Explain how the predictions of softmax regression (multiclass classification) looks like (probabilities and classes) and define the parameter space.