Exercise 7

Machine Learning I

7A-1.

Version 1: From antiderivative to derivative

Derive $\sigma(a)$ manually via product rule and rearrange terms:

Auxiliary calculation

$$
\begin{aligned}
(*) \quad -1 + e^{-a}\sigma &= -1 + \frac{e^{-a}}{(1+e^{-a})} \\
&= \frac{(-1-e^{-a}) + e^{-a}}{(1+e^{-a})} \\
&= -\sigma
\end{aligned}
$$

$$
\begin{aligned}
\frac{d\sigma}{da} &= e^{-a}(1+e^{-a})^{-2} \\
&= \underbrace{(1+e^{-a})^{-1}}_{\sigma} e^{-a} \underbrace{(1+e^{-a})^{-1}}_{\sigma} \\
&= \sigma\left(1 \underbrace{-1 + e^{-a}\sigma}_{(*)}\right) \\
&= \sigma(1-\sigma)
\end{aligned}
$$

Version 2: From derivative to antiderivative

Given is a first order homogeneous ODE. As usual, we first check if $f(\sigma)$is globally Lipschitz continuous. If this condition is met, we immediately know that a solution exists and is unique on its interval of existence (Picard Lindelöf Existence/Uniqueness Theorem).

Afterwards, we try to find a maximum solution of the ODE.

Prerequisites

Given an initial value problem of the form

$$\dot{x} = f(t,x), \qquad x(t_0) = x_0$$

whereas $f: Z_{a,b} \to \mathbb{R}^N$ is continuous with domain

$$Z_{a,b} \coloneqq [t_0 - a, t_0 + a] \times \overline{U_b^N(x_0)}, a > 0, b > 0$$

Additionally, there exists an $L \geq 0$ with

$$\|f(t,x) - f(t,y)\| \leq L\|x - y\| \; for \; all \; (t,x),(t,y) \in Z_{a,b}$$

Then there is exactly one solution of the IVP on $[t_0 - \alpha, t_0 + \alpha]$, whereas

$$\alpha \coloneqq \min\left\{a, \frac{b}{M}\right\}, M \coloneqq \max\{\|f(t,x)\| : (t,x) \in Z_{a,b}\}$$

The trivial case $M = 0$ is included with $\frac{b}{M} = \infty$.

Above definition comes from page 60 of "Gewöhnliche Differenzialgleichungen" by Bernd Aulbach. I like this book very much.

The domain of our ODE is $D \coloneqq I \times J = \mathbb{R} \times (0,1)$. Because of the boundedness of x, we satisfy Lipschitz-continuity:

$$\begin{aligned} & \frac{\|x(1-x) - y(1-y)\|}{\|x - y\|} && \leq L \\ \Leftrightarrow \quad & \left\| \frac{-(x^2 - y^2)}{x - y} + \frac{x - y}{x - y} \right\| && \leq L \\ \Leftrightarrow \quad & \|-x - y + 1\| && \leq L \end{aligned}$$

is satisfied for any fixed $L \geq 1$ and arbitrary $x, y \in (0,1)$. This means a unique solution exists on D.
Note that $\frac{d\sigma}{da} = \sigma(1-\sigma) = g(a)h(\sigma)$ separable, which means there is a standard way to solve this equation.

To avoid multiplying by nonarchimedean quantities $d\sigma$, da, we do the following:

Let $h(\sigma) \neq 0$ on J_0. Let $H: J_0 \to \mathbb{R}$ be an antiderivative of $\frac{1}{h(\sigma)}$ and $H^{-1}: H(J_0) \to J_0$ its inverse.
Additionally, let $G: I \to \mathbb{R}$ be an antiderivative of $g(a)$.

It can be shown that $\lambda_\beta(a) \coloneqq H^{-1}(G(a) + \beta)$ solves the ODE on $I_0 \subseteq I$.

Calculation of G:

$$\begin{aligned} G(a) &= \int_{a_0}^{a} 1 \, da \\ &= a - a_0 \end{aligned}$$

Calculation of H:

$$\begin{aligned} H(\sigma) &= \int_{\sigma_0}^{\sigma} \frac{1}{s(1-s)} ds \\ &= \int_{\sigma_0}^{\sigma} \frac{1}{s^2(s^{-1} - 1)} ds \end{aligned}$$

Substitution: $z = s^{-1} - 1$.
Then we have:

$$\begin{aligned} \frac{dz}{ds} &= -s^{-2} \\ \Leftrightarrow \quad -s^2 dz &= ds \end{aligned}$$

Plugged into our integral we receive:

$$\begin{aligned} H(\sigma) &= -\int_{\sigma_0}^{\sigma} \frac{1}{s^2 z} s^2 dz \\ &= -\int_{\sigma_0}^{\sigma} \frac{1}{z} dz \\ &= -[\ln|z|]_{\sigma_0}^{\sigma} \\ &= -\ln|\sigma^{-1} - 1| + \ln\left|\sigma_0^{-1} - 1\right| \end{aligned}$$

Calculation of inverse:

$$\begin{aligned} & H(\sigma) && = -\ln|\sigma^{-1} - 1| + \ln\left|\sigma_0^{-1} - 1\right| \\ \Leftrightarrow \quad & e^{H(\sigma)} && = (\sigma^{-1} - 1)^{-1} \cdot (\sigma_0^{-1} - 1) \\ \Leftrightarrow \quad & \sigma^{-1} - 1 && = (\sigma_0^{-1} - 1) e^{-H(\sigma)} \\ \Leftrightarrow \quad & \frac{1}{1 + (\sigma_0^{-1} - 1) e^{-H(\sigma)}} && = \sigma \\ & && = H^{-1} \end{aligned}$$

Let $\beta = 0$. Plugged into our solution:

$$\begin{aligned} \lambda_\beta(a) &:= H^{-1}(G(a) + \beta) \\ &= \frac{1}{1 + (\sigma_0^{-1} - 1) e^{-H(G(a)+\beta)}} \\ &= \frac{1}{1 + (\sigma_0^{-1} - 1) e^{-a+a_0}} \end{aligned}$$

The equation given in the task is thereby a specific solution with initial values ($\sigma_0 = \frac{1}{2}, \backslash a_0 = 0$):

$$\lambda(a; \sigma_0, a_0) = \frac{1}{1 + e^{-a}}$$

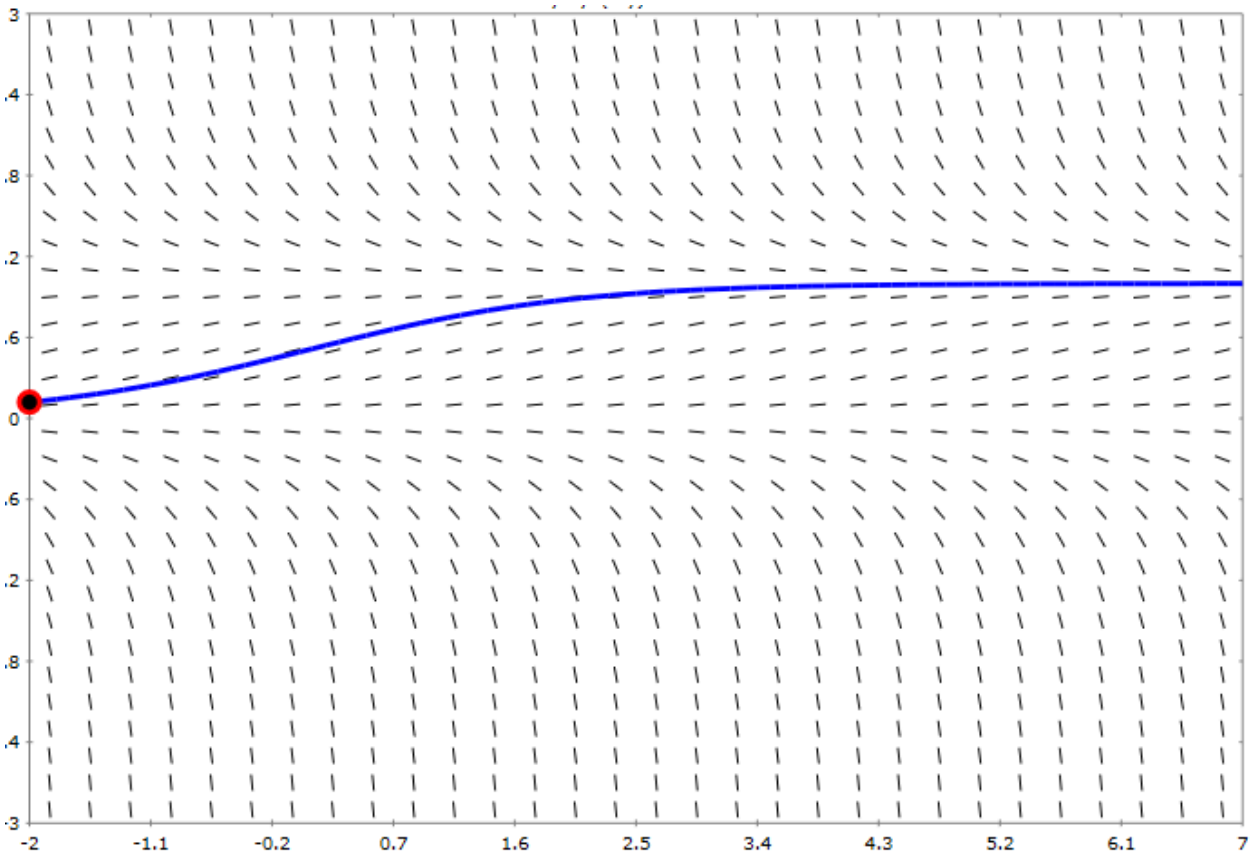

Figure 1 Slope field of $\frac{d\sigma}{da}$. It can be seen that the solution $\sigma = 1$ is attractive and $\sigma = 0$ is divergent. Because of Picard Lindelöf, two solutions cannot overlap, which guarantees that our specific solution (blue line) with the initial values $\sigma_0 = \frac{1}{2}, \backslash a_0 = 0$ stays in the corridor (0,1) (exclusive). You can generate your own slope field plot on http://slopefield.nathangrigg.net/.

7A-2.

Auxiliary calculation

It can be easily shown that by calculating $\frac{\partial}{\partial w_i}$ and collecting terms,
we can generalize as follows:

$$(*) \quad \nabla \ln \sigma(\mathbf{w}^T \boldsymbol{\phi}) = \frac{\sigma(1-\sigma)}{\sigma}\boldsymbol{\phi} = (1-\sigma)\boldsymbol{\phi}$$

$$(**) \quad \nabla \ln\big(1-\sigma(\mathbf{w}^T \boldsymbol{\phi})\big) = \frac{-\sigma(1-\sigma)}{1-\sigma}\boldsymbol{\phi} = -\sigma\boldsymbol{\phi}$$

Please note: For some people the gradient is a row vector. If this interpretation is chosen, the result $E(\mathbf{w})$ is transposed of what follows below.

$$
\begin{aligned}
\nabla E(\mathbf{w}) &= -\sum_{n=1}^{N}\left\{t_n \underbrace{\nabla \ln y_n}_{(*)} + (1-t_n)\underbrace{\nabla \ln(1-y_n)}_{(**)}\right\} \\
&= -\sum_{n=1}^{N}\{t_n(1-\sigma_n)\boldsymbol{\phi}_n - (1-t_n)\sigma_n\boldsymbol{\phi}_n\} \\
&= -\sum_{n=1}^{N}\{t_n-\sigma_n\}\boldsymbol{\phi}_n \\
&= \sum_{n=1}^{N}\{y_n-t_n\}\boldsymbol{\phi}_n
\end{aligned}
$$

Another way is to doubly apply the chain rule (credits to Dario Hett for this idea):

$$
\begin{aligned}
\nabla E(\mathbf{w}) &= -\sum_{n=1}^{N}\left[\frac{\partial\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}}{\partial y_n}\frac{\partial y_n}{\partial a}\frac{\partial a}{\partial \mathbf{w}}\right] \\
&= -\sum_{n=1}^{N}\left[\left\{\frac{t_n}{y_n}-\frac{1-t_n}{1-y_n}\right\}y_n(1-y_n)\boldsymbol{\phi}_n\right] \\
&= \sum_{n=1}^{N}\{y_n-t_n\}\boldsymbol{\phi}_n
\end{aligned}
$$

This has the advantage that we can just use scalar derivative rules for the logarithm.

7A-3.

The reason is the same as for linear regression: Introducing nonlinear basis functions allows us to create nonlinear decision boundaries. For example, $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x_n}) = -1 + x_1^2 + x_2^2$ clusters our binary dataset in a circular manner.

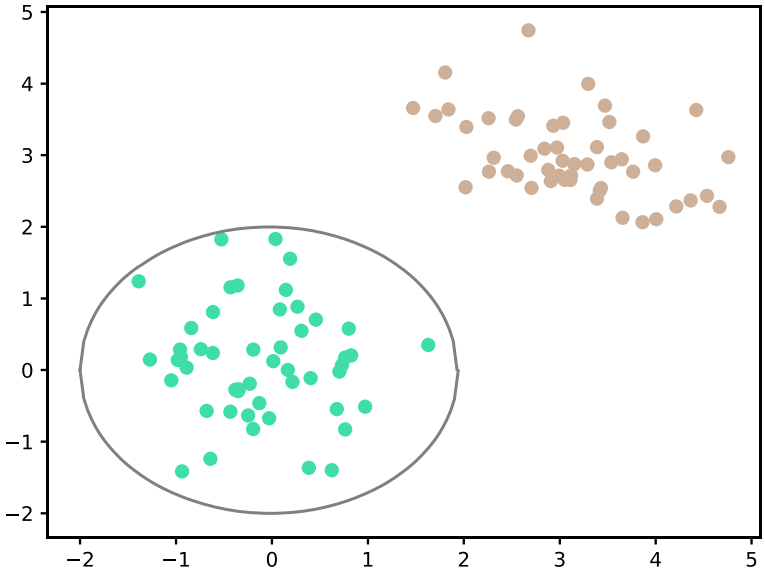

Figure 2 Nonlinear basis functions allow us to create nonlinear boundaries.
The plotted boundary is not perfect but possible using $\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n}) = w_0 + w_1x_1^2 + w_2x_2^2$.

7A-4.

As the data is linearly separable, there exists a partition such that:

$$(*) \quad \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x_n}) > 0 \; if \; t_n = 1,$$

$$\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x_n}) < 0 \; if \; t_n = 0$$

The error function $E(\mathbf{w})$ is minimal, if $y_n = t_n \; \forall n$. We can now take the limit of the original assignment $(*)$:

$$y_n = \lim_{\|\mathbf{w}\| \to \infty} \frac{1}{1 + e^{-\mathbf{w}^T \boldsymbol{\phi}}} = \begin{cases} 1, & if \; t_n = 1 \\ 0, & if \; t_n = 0 \end{cases}$$

The last equality holds, because

$$\begin{aligned} \mathbf{w}^T \boldsymbol{\phi} &= \langle \mathbf{w}, \boldsymbol{\phi} \rangle \\ &= \|\mathbf{w}\| \cdot \|\boldsymbol{\phi}\| \cdot \cos \alpha \end{aligned}$$

which entails:

$$\lim_{\|\mathbf{w}\| \to \infty} \mathbf{w}^T \boldsymbol{\phi} = \begin{cases} +\infty, & if \; \alpha \in [0, 0.5\pi] \cup [1.5\pi, 2\pi) \\ -\infty, & if \; \alpha \in (0.5\pi, 1.5\pi] \end{cases}$$

7A-5.

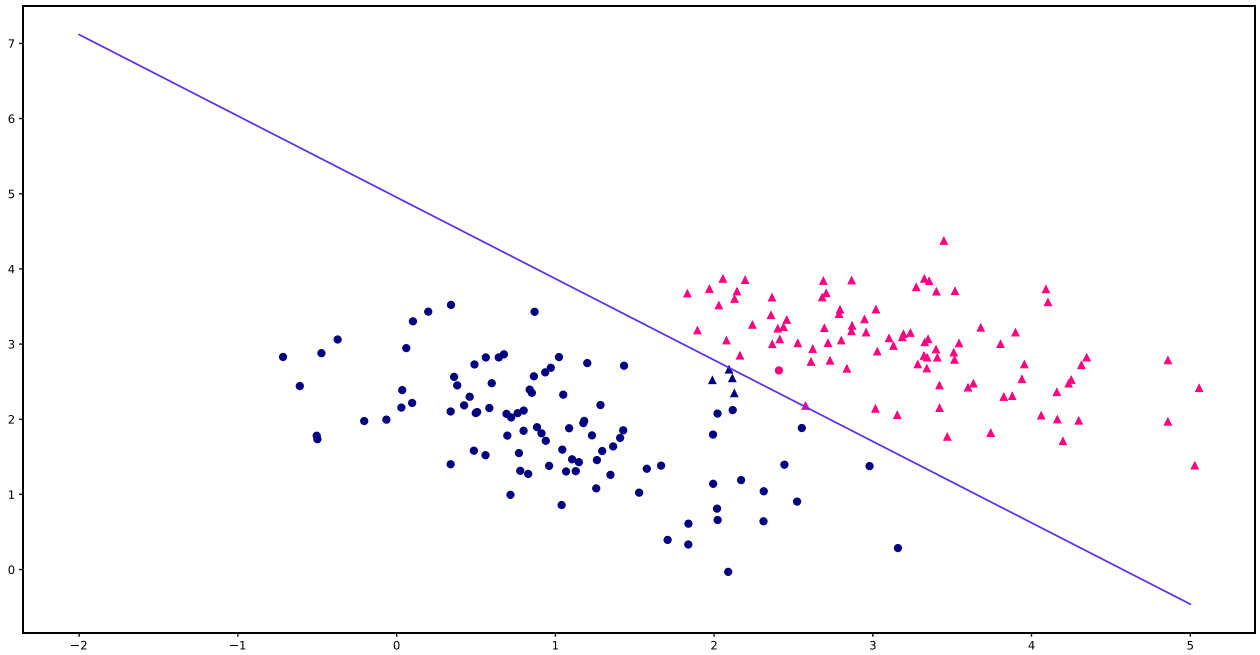

Figure 3 Two different classes (circle and triangle) in juxtaposition. The color indicates predicted class labels (blue=circle, pink=triangle) As there is some crossover between the classes, the decision boundary does not perfectly separate the data. This means some triangles have blue color and one circle pink color. As this is a vector graphic, you can zoom in for a clearer view without sacrificing image quality.

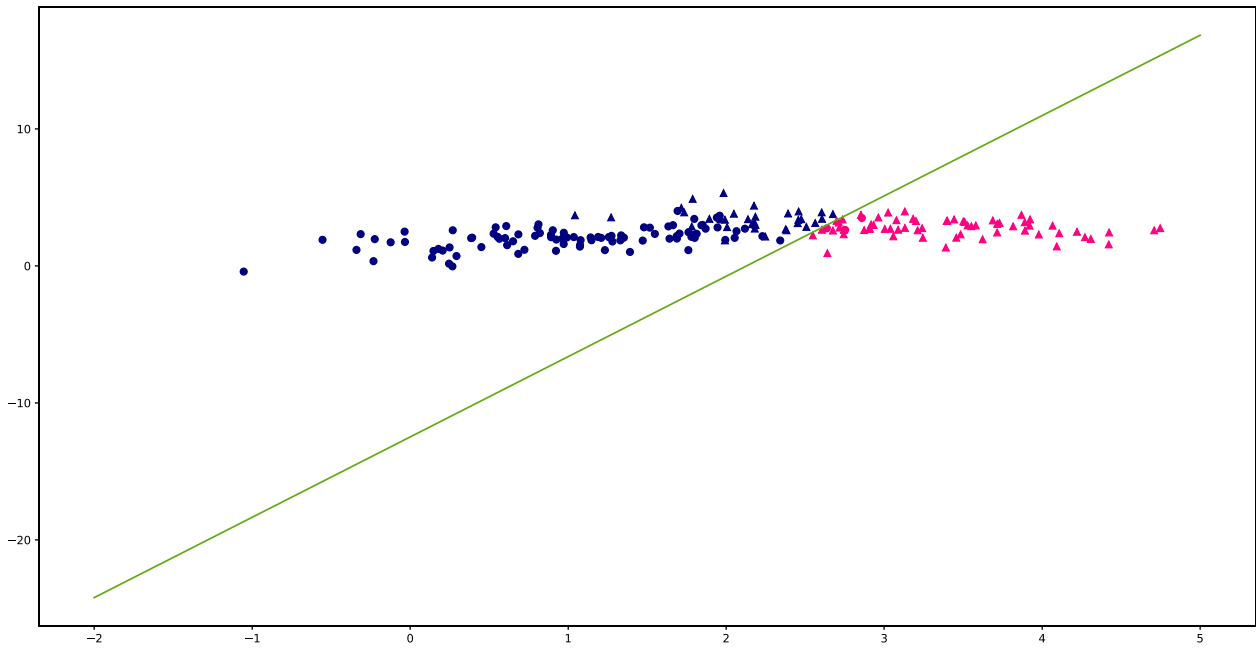

Figure 4 In this example, there is more interplay between the classes. We see many misclassified triangles. As this is a vector graphic, you can zoom in for a clearer view without sacrificing image quality.

We call α the learn rate. If you decrease alpha, the step size becomes slower. This increases computational time but is not detrimental to the overall error in our estimate for $\mathbf{w}$. If you set α too large, however, it is possible that the gradient algorithm does diverge. Depending on the topology of $E(\mathbf{w})$, a correct step size (i.e. a step size that is not too large) is sometimes impossible to find. For more, google “Stiff problems ODE solvers”.