Cognitive Algorithms - Exercise Sheet 1

Intro & Perceptron

Department of Machine Learning - TU Berlin

Disclaimer

For each exercise, but particularly for exercises involving calculations:

Show your work or you will not receive (full) credit!

Furthermore: Exercises marked with an asterisk * are not required and don't contribute towards the credit you receive, but you are welcome to do them because they are fun.

Introduction

Binary linear classification is defined for both NCC (Nearest Centroid Classifier), also known as Prototype Classifier, and Perceptron as:

$$\mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x} - \beta \quad \Rightarrow \begin{cases} \text{class 1} & \text{if } \mathbf{w}^\top \mathbf{x} - \beta < 0 \\ \text{class 2} & \text{if } \mathbf{w}^\top \mathbf{x} - \beta \geq 0 \end{cases} \tag{1}$$

Both NCC and Perceptron are described by (1), but $\mathbf{w}$ and β are determined differently.
For NCC: $\mathbf{w}$ and β are determined as shown in (2) and (3). For Perceptron: $\mathbf{w}$ and β are determined via gradient descent.

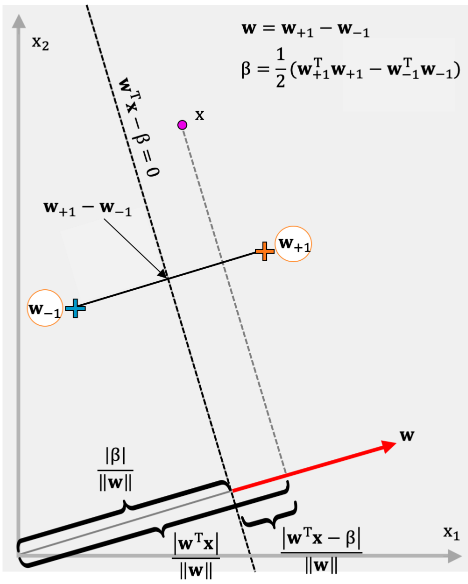


Figure 1: NCC Algorithm including weight vector $\mathbf{w}$ and bias β. For the sake of classification it is sufficient to define: $\mathbf{w}_{NCC} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. Source: Prof. Benjamin Blankertz, TU Berlin

Task 1 - Prototype classifier [5 points]

The goal of this task is to compute and visualize a prototype classifier for a very simple two-class classification problem in the 2D-space. Consider the following data points:

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

Vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ belong to class -1, while $\mathbf{x}_3$ and $\mathbf{x}_4$ belong to class $+1$. Note that bold characters denote vectors. Remember the following formulas:

$$\mathbf{w} = \mathbf{w}_{+1} - \mathbf{w}_{-1} \tag{2}$$

$$\beta = \frac{1}{2}(\mathbf{w}_{+1}^\top \mathbf{w}_{+1} - \mathbf{w}_{-1}^\top \mathbf{w}_{-1}) \tag{3}$$

1. [**1 point**] Compute class means $\mathbf{w}_{-1}$ and $\mathbf{w}_{+1}$.
2. [**1 point**] Compute the classification boundary $\mathbf{w}^\top \mathbf{x} - \beta = 0$ of the prototype classifier.
3. [**1 point**] For each point, compute the assigned class label $\text{sign}(\mathbf{w}^\top \mathbf{x} - \beta)$. Are all points correctly classified?
4. [**2 points**] Sketch the data points, their class means $\mathbf{w}_{-1}$ and $\mathbf{w}_{+1}$, the normal vector $\mathbf{w}$, and the classification boundary in the x_1-x_2 space where x_1 denotes the horizontal axis.

Task 2 - Linear classification boundary [5 points]

Consider a linear classification boundary $\mathbf{w}^\top \mathbf{x} - \beta = 0$ in the 2D-space, i.e. two features per data point. Draw a sketch in 2D to visualize the classification boundary and answer the following questions:

1. [**1 point**] Suppose $\beta = 0$ and $\|\mathbf{w}\| = 1$. How large is the distance of a point $\mathbf{z}$ to the classification boundary?
2. [**1 point**] How large is the distance between a point $\mathbf{z}$ and a classification boundary $\mathbf{w}$ if $\|\mathbf{w}\| = 1$ but $\beta \neq 0$?
3. [**1 point**] How large is the distance between a point $\mathbf{z}$ and a classification boundary $\mathbf{w}$ for arbitrary β and $\mathbf{w}$?
4. [**1 point**] How large is the distance between a classification boundary $\mathbf{w}$ and the origin for arbitrary β and $\mathbf{w}$?
5. [**1 point**] Is β in the general case the intercept of the classification boundary $\mathbf{w}^\top \mathbf{x} - \beta = 0$ with the x_2-axis? If yes, explain why. If not, give a counter-example.

Task 3 - Convergence of the perceptron [*]

Suppose we have N points $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^d$ with class labels $y_1, \ldots, y_N \in \{-1, +1\}$, and that the data set is linearly separable. Prove that the perceptron learning algorithm converges to a separating hyperplane in a finite number of steps for a learning rate $\eta = 1$.

1. We denote a hyperplane by $\mathbf{w}^\top \mathbf{x} = 0$. Show that there exists a $\mathbf{w}_{\text{sep}}$ such that:

$$\mathbf{w}_{\text{sep}}^\top \mathbf{x}_i y_i \geq \|\mathbf{x}_i\|^2, \forall i \in \{1, \ldots N\}.$$

2. Given a current $\mathbf{w}_{\text{old}} \in \mathbb{R}$, the perceptron algorithm identifies a point $\mathbf{x}_m$ that is misclassified, and produces the update rule $\mathbf{w}_{\text{new}} = \mathbf{w}_{\text{old}} + \eta \mathbf{x}_m y_m$. Using this update rule, show that

$$\|\mathbf{w}_{\text{new}} - \mathbf{w}_{\text{sep}}\|^2 \leq \|\mathbf{w}_{\text{old}} - \mathbf{w}_{\text{sep}}\|^2 - \|\mathbf{x}_m\|^2 \tag{4}$$

This implies that the perceptron algorithm converges to a separating hyperplane in a finite number of steps.

1. Because the data set is linearly seperable, there exists a w_{sep}, such that for all x_i, $w_{sep}^T x_i \geq \xi$ for some $\xi \geq 0$. Let x_j be the data point with the largest squared norm $\|x_j\|^2$.

$$\begin{aligned} & w_{sep}^T x_i y_i \geq \xi \\ & \Leftrightarrow \frac{1}{\xi} w_{sep}^T x_i y_i \geq 1 \\ & \Leftrightarrow [w'_{sep}]^T x_i y_i \geq 1 \\ & \Leftrightarrow [w''_{sep}]^T x_i y_i \geq \|x_j\|^2 \geq \|x_i\|^2 \end{aligned}$$

2. Now let $\mathbf{w}^{\text{sep}} = w''_{sep}$.

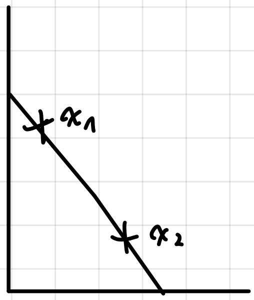
$$\begin{aligned} \|\mathbf{w}^{\text{new}} - \mathbf{w}^{\text{sep}}\|^2 &= \|\mathbf{w}^{\text{old}} + \underbrace{\eta}_{=1} \mathbf{x}_m y_m - \mathbf{w}^{\text{sep}}\|^2 \\ &= (\mathbf{w}^{\text{old}} - \mathbf{w}^{\text{sep}} + \mathbf{x}_m y_m)^\top (\mathbf{w}^{\text{old}} - \mathbf{w}^{\text{sep}} + \mathbf{x}_m y_m) \\ &= (\mathbf{w}^{\text{old}} - \mathbf{w}^{\text{sep}})^\top (\mathbf{w}^{\text{old}} - \mathbf{w}^{\text{sep}}) + 2(\mathbf{w}^{\text{old}} - \mathbf{w}^{\text{sep}})^\top \mathbf{x}_m y_m + (\mathbf{x}_m y_m)^\top \mathbf{x}_m y_m \\ &= \|\mathbf{w}^{\text{old}} - \mathbf{w}^{\text{sep}}\|^2 + 2 \underbrace{(\mathbf{w}^{\text{old}})^\top \mathbf{x}_m y_m}_{\leq 0 \text{ since } \mathbf{x}_m \text{ misclassified}} - 2(\mathbf{w}^{\text{sep}})^\top \mathbf{x}_m y_m + \underbrace{y_m^2}_{=1} \|\mathbf{x}_m\|^2 \\ &\leq \|\mathbf{w}^{\text{old}} - \mathbf{w}^{\text{sep}}\|^2 - 2(\mathbf{w}^{\text{sep}})^\top \mathbf{x}_m y_m + \|\mathbf{x}_m\|^2 \\ &\overset{3.1}{\leq} \|\mathbf{w}^{\text{old}} - \mathbf{w}^{\text{sep}}\|^2 - 2\|\mathbf{x}_m\|^2 + \|\mathbf{x}_m\|^2 \\ &= \|\mathbf{w}^{\text{old}} - \mathbf{w}^{\text{sep}}\|^2 - \|\mathbf{x}_m\|^2. \end{aligned}$$

NCC and Perceptron

	NCC	Perceptron
Problem	Classification	Classification, (Regression)
Model	$y = \text{sign}(\mathbf{w}^T\mathbf{x})$	$y = f(\mathbf{w}^T\mathbf{x})$
Error	distance to $\mathbf{w}_{+1}, \mathbf{w}_{-1}$	$-\sum_{m \in M} \mathbf{w}^T \mathbf{x}_m y_m$
Optimization	closed form	SGD
Result	always the same	can differ
Application	Cancer Prediction[1]	NLP[2]

test

Desision Boundary and w



$$0 = w^T \bar{x} + \beta$$

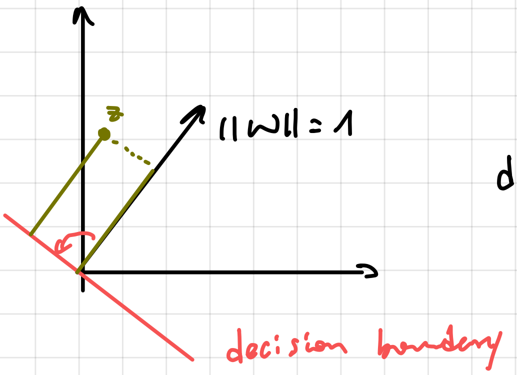
$$0 = w^T x_1 + \beta - w^T x_2 - \beta$$

$$= w^T (x_1 - x_2) = 0$$

task 2

1. $\beta = 0$ and $||w|| = 1$

$$||w||^2 = \left(\sqrt{w^T \cdot w}\right)^2 = w^T \cdot w$$



$$d = |w^T z|$$

2. $\beta \neq 0$ and $||w|| = 1$