TD 3

Exercice 1 – Perceptron

Q 1.1 Recall the least squares cost function for a binary learning problem. Provide a few examples to show that correctly classified samples contribute to the cost function.

Q 1.2 What is the cost function used by the perceptron algorithm ?

Q 1.3 Assuming a function f of infinite complexity (able to model any decision boundary), draw by hand the optimal decision boundary according to the costs defined previously for the two toy problems in figure 1. Are these boundaries *interesting* ? What problems arise ?

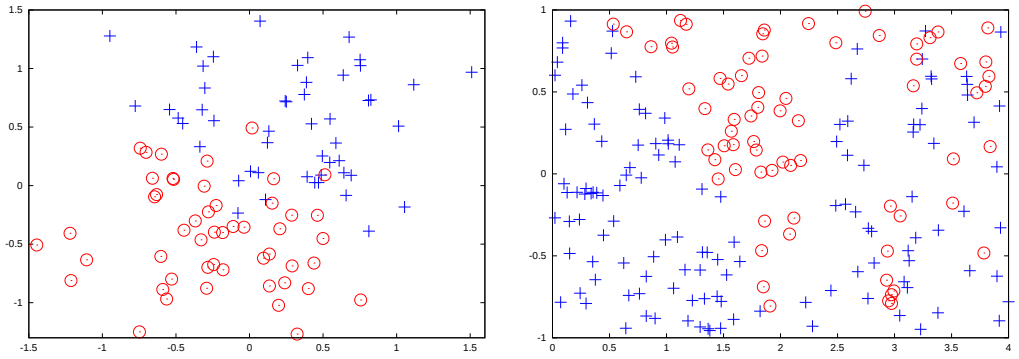

FIGURE 1 – Non-linearly separable Gaussians

Q 1.4 Let $\mathbf{w} = (2, 1)$ be the weight vector of a linear separator. Draw this separator in the plane. Indicate the quantities $< \mathbf{w}, \mathbf{x} >$ with respect to a well-classified and a misclassified example. What happens to the scalar product in the case of a misclassified example with the update $\mathbf{w} \leftarrow \mathbf{w} + y\mathbf{x}$?

Q 1.5 How do the following classifiers compare to the one from the previous question : $w^1 = (1, 0.5)$, $w^2 = (200, 100)$, $w^3 = (-2, -1)$?

Q 1.6 Show that the perceptron algorithm corresponds to gradient descent. Is the solution unique ?

Q 1.7 What problem can arise for certain values of w ? How can it be addressed ?

Q 1.8 What is the difference between stochastic gradient descent and batch gradient descent ? And mini-batch ?

Q 1.9 Provide a perceptron that implements the logical `AND` operation between the binary inputs x_1 and x_2 (positive if both are 1, negative otherwise), and another one for the logical `OR` operation.

Exercice 2 – Convergence of the Perceptron

In this exercise, we assume a data set $\{\mathbf{x}^i, y^i)\}_{i=0}^N$ that is linearly separable. We will study the convergence of the perceptron algorithm as iterations progress. For this, we will only consider the "useful" iterations, that is, those where an update of the weight vector $\mathbf{w}$ is made. We also assume that $\mathbf{w}^0 = 0$. Additionally, we denote $R = \max_{i=1}^N \|\mathbf{x}^i\|$.

Q 2.1 Let $\gamma > 0$ and $\mathbf{w}^*$ be such that $y_i \frac{<\mathbf{w}^* \cdot \mathbf{x}^i>}{\|\mathbf{w}^*\|} \geq \gamma$ for all examples in the data set. What does the existence of γ and $\mathbf{w}^*$ mean geometrically ? If $\mathbf{w}^*$ exists, is it unique ?

Q 2.2 Assuming the algorithm has not converged at time t, give an upper bound for $\|\mathbf{w}^t\|^2$ in terms of t and R using an induction rule on t.

Q 2.3 Assuming the algorithm has not converged at time t, give a lower bound for $< \mathbf{w}^t, \mathbf{w}^* >$.

Q 2.4 Using the Cauchy-Schwarz inequality $| < \mathbf{w}, \mathbf{w}^* > | \leq \|\mathbf{w}\|\|\mathbf{w}^*\|$, prove Novikoff's theorem : the number of iterations t of the algorithm is bounded by $\frac{R^2}{\gamma^2}$.

Exercice 3 – Expressiveness of Linear Separators

We are working in the space of linear separators : $f_{\mathbf{w}}(\mathbf{x}) = \sum_{j=1}^{d} x_j w_j$.

Q 3.1 What is the dimension of the vector $\mathbf{w}$? Recall the matrix form of $f_{\mathbf{w}}(\mathbf{x})$. Approximate the optimal decision boundaries using a basic linear model on the figure 1.

Q 3.2 We will increase the expressiveness of the model by extending the initial representation space in the 2D case : $\mathbf{x} = [x_1, x_2]$. Let the following transformation ϕ be defined : $\phi(\mathbf{x}) = [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2]$, and consider the linear model $f_{\mathbf{w}}(\mathbf{x}) = \sum_{j=1}^{d} \phi_j(\mathbf{x}) w_j$.

- What is the dimension of the vector $\mathbf{w}$ in this case ?
- What does the projection ϕ correspond to ?
- Redraw the optimal decision boundaries on the figure using this new representation.
- Can we recover the linear decision boundaries from the previous question in this new space ? If so, give the associated coefficients w_j.

Q 3.3 Are the decision boundaries more *interesting* when using the first or the second data representation ? Can you roughly compare the magnitude of the cost function (e.g., least squares) in the linear and quadratic cases ? What can we infer from this ? On what element do you base your measurement of the model's quality ?

Q 3.4 To increase the expressiveness of our separator class, we turn to Gaussian representations. The input space is discretized using a grid of N^2 points $\mathbf{p}^{i,j}$, and then we measure the Gaussian similarity of the point $\mathbf{x}$ with respect to each point on the grid :

$$s(\mathbf{x}, \mathbf{p}^{i,j}) = K e^{-\frac{||\mathbf{x}-\mathbf{p}^{\mathbf{i},\mathbf{j}}||^2}{\sigma}}.$$

The new representation of the example is the vector containing the similarity of the example to each grid point :

$$\phi(\mathbf{x}) = (s(\mathbf{x}, \mathbf{p}^{1,1}), s(\mathbf{x}, \mathbf{p}^{1,2}), \ldots)$$

- What is the dimension of the vector $\mathbf{w}$?
- Provide the literal expression for the decision function.
- What role does the parameter σ play ?

Q 3.5 Pragmatic introduction to kernels

- What happens in dimension 3 if we want to keep the spatial resolution of the grid ?
- To address this issue, we propose using the learning set instead of the grid : the points serving as the support for the projection will be those from the learning set. Express the literal form of the decision function in this new framework. What is the new dimension of the parameter $\mathbf{w}$?
- What happens when σ approaches 0 ? What happens when σ approaches infinity ? Do we need all the dimensions of $\mathbf{w}$, or can we recover the same decision boundary by limiting the number of learning data ? What does this correspond to for $||\mathbf{w}||$?