Introduction to Machine Learning (67577)

Recitation 06
Classification - Half-space Classifiers

Second Semester, 2021

Contents

1 Half-Space Classifier

Similar to linear regression, one of the simplest families of classifiers is that of linear classifiers. In these, we are interested in separating a given dataset into two classes using a linear separator function, as seen in Figure 1.

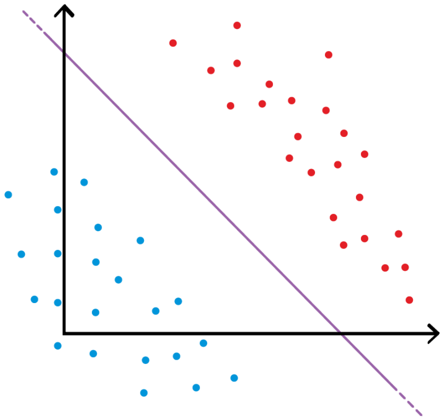

Figure 1: *Half-space Classification Illustration:*
For a domain-set $\mathcal{X} \in \mathbb{R}^2$ the two classes, coded as red and blue colors, are linearly separable

As we have seen in linear regression, the family of linear functions can be described as:

$$L_d := \left\{\mathbf{x} \mapsto \mathbf{x}^\top \mathbf{w} + b | \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\right\}$$

where the linearity refers to the functions being linear in the parameters $\mathbf{w}$. Next, consider the following definitions:

Definition 1.1 Let $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. The hyperplane defined by $(\mathbf{w}, b)$ is the set

$$\left\{\mathbf{x} | \langle \mathbf{w}, \mathbf{x} \rangle = b, \, \mathbf{x} \in \mathbb{R}^d\right\}$$

Definition 1.2 Let $(\mathbf{w}, b)$ be an hyperplane, so the half-space of $(\mathbf{w}, b)$ is defined as the set

$$\left\{\mathbf{x} | \langle \mathbf{w}, \mathbf{x} \rangle \geq b, \mathbf{x} \in \mathbb{R}^d\right\}$$

or equivalently as $\left\{\mathbf{x} | sign\left(\langle \mathbf{w}, \mathbf{x} \rangle - b\right) \geq 0, \mathbf{x} \in \mathbb{R}^d\right\}$.

Notice that the family of linear functions L_d is in-fact the family of hyperplanes. As such, we can look at the family of functions that is the composition of the *sign* function and L_d, $sign \circ L_d$. This family defines the hypothesis class of the half-space classifiers. Denote $h_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ then:

$$\mathcal{H}_{half} := \left\{h_{\mathbf{w},b}(\mathbf{x}) = sign\left(\langle \mathbf{x}, \mathbf{w} \rangle + b\right) | \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\right\} = \left\{\mathbf{x} \mapsto sign\left(\langle \mathbf{x}, \mathbf{w} \rangle + b\right)\right\} \tag{1}$$

So why are functions in the form seen in 1 are half-space classifiers? Let us assume at first that $b = 0$.

We can express the domain set as a disjoint union of the following:

$$\mathbb{R}^d = \left\{\mathbf{x} \in \mathbb{R}^d | \mathbf{x}^\top \mathbf{w} > 0\right\} \biguplus \left\{\mathbf{x} \in \mathbb{R}^d | \mathbf{x}^\top \mathbf{w} = 0\right\} \biguplus \left\{\mathbf{x} \in \mathbb{R}^d | \mathbf{x}^\top \mathbf{w} < 0\right\}$$

These sets correspond to the open half spaces on either side of the hyperplane $\mathbf{w}^\perp = \left\{\mathbf{x} \in \mathbb{R}^d | \mathbf{x}^\top \mathbf{w} = 0\right\}$ and points on the hyper-plane itself. As such. each vector $\mathbf{w} \in \mathbb{R}^d$ defines a hyper-plane $\mathbf{w}^\perp$ that divides $\mathbb{R}^d$ into two half-spaces.

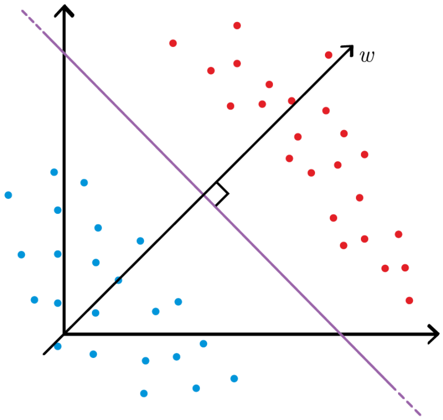

Figure 2: *Corresponding Hyperplane to* $\mathbf{w}^\perp$

The case where $b = 0$ is called the **homogeneous** case, as the hyperplane $\mathbf{w}^\perp$ is a linear subspace going through the origin. When $b \neq 0$ the hyperplane does not go through the origin and is called the non-homogeneous case. Recall that we have seen how we could transition from the non-homogeneous to the homogeneous case in the linear regression chapter.

Given a sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, we would like to find an hypothesis $h_{\mathbf{w},b} \in \mathcal{H}_{half}$ such that all data points in S that are labeled 1 are on the one side of the hyper-plane and all those labeled -1 are on the other side. To find such an hypothesis we must first make the assumption that the dataset is **linearly separable**. That is, there exists a hyper-plane such that samples of opposing labels are on opposite sides. Mathematically, we assume that

$$\exists \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R} \quad s.t \quad \forall i \in [m] \quad y_i \cdot sign\left(\langle \mathbf{x}, \mathbf{w} \rangle + b\right) = 1$$

or equivalently since the inner product will be negative for all samples with $y_i < 0$ and positive for all samples with $y_i > 0$:

$$\exists \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R} \quad s.t \quad \forall i \in [m] \quad y_i \cdot \left(\langle \mathbf{x}, \mathbf{w} \rangle + b\right) > 0$$

Note, that assuming that a given training set is linearly separable is a **realizability assumption**. Namely, the labels are generated by a function in our hypothesis class $\mathcal{H}_{half}$.

1.1 Learning Linearly Separable Data Via ERM

To train a model over the defined hypothesis class of homogenous half-spaces ($\mathbf{w} \in \mathbb{R}^d, b = 0$) observe the following: For any hypothesis $h_{\mathbf{w}} \in \mathcal{H}_{half}$, the misclassified training samples are exactly those where $y_i \cdot sign(\langle \mathbf{x}, \mathbf{w} \rangle) = -1$ or equivalently $y_i \langle \mathbf{x}, \mathbf{w} \rangle < 0$. So defining the loss of a given hypothesis over S is:

$$L_S(h_{\mathbf{w}}) := \sum_{i=1}^{m} \mathbb{1}[y_i \langle \mathbf{x}, \mathbf{w} \rangle < 0]$$

Since we assumed that S is linearly separable, we would like to find $h_{\mathbf{w}} \in \mathcal{H}_{half}$ that perfectly separates the training set. Such an hypothesis will be one that achieves $L_S(h_{\mathbf{w}}) = 0$. In other words, we are applying the ERM principle and seeking for any separating hyperplane $\mathbf{w}^{\perp}$, corresponding to an hypothesis $h_{\mathbf{w}}$ that minimizes the empirical risk $L_S(h_{\mathbf{w}})$.

1.2 The Perceptron Algorithm

Once we realize what learning principle we want to apply, we need to find a computationally efficient algorithm to find the desired hypothesis. Even though half-spaces is an ERM problem, we can solve it by using the Perceptron algorithm, suggested by Frank Rosenblatt in 1958. This is an iterative algorithm that constructs a series of vectors $\mathbf{w}^{(0)}, \mathbf{w}^{(1)}, \ldots$, each derived from the previous. At each iteration t we search for a sample i which is misclassified by $w^{(t)}$. Then, we update $\mathbf{w}^{(t)}$ by moving it in the direction of thhe misclassified sample $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_i \mathbf{x}_i$.

Algorithm 1 Batch-Perceptron

```
procedure PERCEPTRON($S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$)
    $\mathbf{w}^{(0)} \leftarrow 0$                                ▷ Initialize parameters
    for $t = 1, 2, \ldots$ do
        if $\exists i \; s.t. \; y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \leq 0$ then
            $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_i \mathbf{x}_i$
        else
            return $\mathbf{w}^{(t)}$
        end if
    end for
end procedure
```

R The Perceptron algorithm is in fact a simple case of the more general algorithm of Subgradient Descent that we saw before. Furthermore, we can modify the algorithm is such a way that rather than requiring an entire dataset S, it will each time get a single sample and update based on that one sample. We will encounter this variation in Online Learning.

1.3 Hard-SVM

Let's start with the **realizable case**. To implement our learning principle of maximal margin, we need to search, among all the hyperplane separating S, for the hyperplane with maximum margin. Namely, the hypothesis $h_{\mathbf{w},b} \in \mathcal{H}_{SVM}$ our learner will choose is the solution to the following optimization problem:

$$\begin{array}{ll} maximize & M((\mathbf{w}, b), S) \\ subject\,to & y_i \cdot (\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1 \quad i = 1, \ldots, m \end{array} \tag{2}$$

The optimization variables are $\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$. Comparing with the linear program of half-spaces (**??**) we see that the constraints are kept, which ensure the hyperplane chosen separates the training sample, but instead of a trivial objective, we seek to minimize the margin (We don't worry about "maximize" instead of "minimize" as we can just multiply the objective by -1).

1.3.1 Solving Hard-SVM

So is the Hard-SVM a convex optimization problem? Recall, that by our optimization problem 2, we are searching of a separating hyperplane that maximizes the margin from all points. As for any $c > 0$ it holds that $(\mathbf{w}, b) = (c\mathbf{w}, cb)$ we can w.l.o.g constraint ourselfs to $||\mathbf{w}|| = 1$. This way, each hyperplane has a unique vector $\mathbf{w}$ that corresponds to it.

Definition 1.3 Let $\mathbf{x} \in \mathbb{R}^d$ and $B \subseteq \mathbb{R}^d$. The distance from $\mathbf{x}$ to B is:

$$\underset{\mathbf{v} \in B}{inf} ||\mathbf{x} - \mathbf{v}||^2$$

Exercise 1.1 Let $(\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R}$ be a hyperplane where $||w|| = 1$ and $\mathbf{x} \in \mathbb{R}^d$ then the distance between $\mathbf{x}$ and the hyperplane $(\mathbf{w}, b)$ is $|\langle \mathbf{w}, \mathbf{x} \rangle + b|$. ■

Proof. To solve this we begin with defining some point in the hyperplane, calculate it's distance from $\mathbf{x}$ and then showing minimality. Let $\mathbf{v} := \mathbf{x} - (\langle \mathbf{w}, \mathbf{x} \rangle + b) \cdot \mathbf{w}$. This point $\mathbf{v}$ is indeed in the hyperplane:

$$\begin{aligned} \langle \mathbf{w}, \mathbf{v} \rangle + b &= \langle \mathbf{w}, \mathbf{x} - (\langle \mathbf{w}, \mathbf{x} \rangle + b) \cdot \mathbf{w} \rangle + b \\ &= \langle \mathbf{w}, \mathbf{x} \rangle - (\langle \mathbf{w}, \mathbf{x} \rangle + b) ||\mathbf{w}||^2 + b \\ &= \langle \mathbf{w}, \mathbf{x} \rangle - \langle \mathbf{w}, \mathbf{x} \rangle - b + b = 0 \end{aligned}$$

with a distance from x of:

$$||\mathbf{x} - \mathbf{v}|| = |\langle \mathbf{w}, \mathbf{x} \rangle + b| \cdot ||\mathbf{w}|| = |\langle \mathbf{w}, \mathbf{x} \rangle + b|$$

Lastly, let us conclude that such $\mathbf{v}$ is the closest point in the hyperplane to $\mathbf{x}$. Let $\mathbf{u}$ be some point in the hyperplane, then:

$$\begin{aligned} ||\mathbf{x} - \mathbf{u}||^2 &= ||\mathbf{x} - \mathbf{v} + \mathbf{v} - \mathbf{u}||^2 \\ &= ||\mathbf{x} - \mathbf{v}||^2 + ||\mathbf{v} - \mathbf{u}||^2 + 2 \langle \mathbf{x} - \mathbf{v}, \mathbf{v} - \mathbf{u} \rangle \\ &\geq ||\mathbf{x} - \mathbf{v}||^2 + 2 \langle \mathbf{x} - \mathbf{v}, \mathbf{v} - \mathbf{u} \rangle \\ &= ||\mathbf{x} - \mathbf{v}||^2 + 2 \langle (\langle \mathbf{w}, \mathbf{x} \rangle + b) \mathbf{w}, \mathbf{v} - \mathbf{u} \rangle \\ &= ||\mathbf{x} - \mathbf{v}||^2 + 2 (\langle \mathbf{w}, \mathbf{x} \rangle + b) \langle \mathbf{w}, \mathbf{v} - \mathbf{u} \rangle \\ &= ||\mathbf{x} - \mathbf{v}||^2 + 2 (\langle \mathbf{w}, \mathbf{x} \rangle + b) (\langle \mathbf{w}, \mathbf{v} \rangle - \langle \mathbf{w}, \mathbf{u} \rangle) \\ &= ||\mathbf{x} - \mathbf{v}||^2 \end{aligned}$$

■

So, as the margin **??** between a given hyperplane $(\mathbf{w}, b)$ and a set of points S is the minimal distance between the hyperplane and any point in the set, we derive that our optimization problem is infact of the form:

$$\begin{aligned} &\underset{||\mathbf{w}||=1, b}{argmax} \quad \underset{i \in [m]}{min} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| \\ &subject\,to \quad y_i \cdot (\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1 \quad i = 1, \ldots, m \end{aligned} \tag{3}$$

While the constraints enforce $\mathbf{w}$ to define a separating hyperplane, the objective will make us choose a separating hyperplane with the maximal margin. To solve this problem numerically, we will need to perform a slight manipulation and write it in a different form.

Claim 1.1 Let $(\mathbf{v}^*, c^*)$ be an optimal solution of:

$$\begin{aligned} &\underset{(\mathbf{w},b)}{argmin} && ||\mathbf{w}||^2 \\ &subject\,to && y_i \cdot (\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1 \quad i = 1, \ldots, m \end{aligned} \tag{4}$$

Then, $\mathbf{w}^* := \gamma \mathbf{v}^*, b^* := \gamma c^*$ for $\gamma = ||\mathbf{v}^*||^{-1}$ is an optimal solution for 1.3.1.

Proof. Let us begin with simplifying (1.3.1). Consider a *feasible* solution $\mathbf{w}$ to the problem (i.e. that satisfies all constraints). It holds that $|\langle \mathbf{w}, \mathbf{x}_i \rangle + b| = y_i \left(\mathbf{x}_i^\top \mathbf{w} + b \right)$. Hence, we can rewrite (1.3.1) as:

$$\begin{aligned} &\underset{||\mathbf{w}||=1,b}{argmax} && \min_{1 \leq i \leq m} y_i \left(\mathbf{x}_i^\top \mathbf{w} + b \right) \\ &\text{subject to} && y_i \cdot \left(\mathbf{x}_i^\top \mathbf{w} + b \right) \geq 1 \quad i = 1, \ldots, m \end{aligned}$$

Now, it is clear that the constraints are redundant. If $\mathbf{w}$ is infeasible then $min_i y_i \left(\mathbf{x}_i^\top \mathbf{w} + b \right) < 0$, achieving a lower objective than any feasible solution. Therefore, we can re-write the problem as:

$$\underset{||\mathbf{w}||=1,b}{argmax} \quad \min_{1 \leq i \leq m} y_i \left(\mathbf{x}_i^\top \mathbf{w} + b \right) \tag{5}$$

Next, let us examine the optimization in claim. Notice that an optimal solution $\mathbf{v}^*, c^*$ will always satisfy:

$$\min_{1 \leq i \leq m} y_i \left(\langle \mathbf{v}^*, \mathbf{x}_i \rangle + c^* \right) = 1$$

since otherwise, we can divide $\mathbf{v}^*$ by a positive number and get a feasible solution with a lower objective. As scalar multiplication does not change the hyperplane, denote $\gamma = ||\mathbf{v}^*||^{-1}$ and let $\mathbf{w}^* := \gamma \mathbf{v}^*, b^* := \gamma c^*$. Since $(\mathbf{v}^*, c^*)$ is a is the optimal solution of (4) it must satisfy all constraints and is therefore a feasible solution for (5) with an objective:

$$\min_i y_i \left(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^* \right) = \min_i y_i \left(\langle \gamma \mathbf{v}^*, \mathbf{x}_i \rangle + \gamma c^* \right) = \gamma$$

Assume towards contradiction that there is a solution $(\mathbf{w}_2, b_2) \neq (\mathbf{w}^*, b^*)$ achieving a higher objective:

$$\min_i y_i \left(\langle \mathbf{w}_2, \mathbf{x}_i \rangle + b_2 \right) = \delta > \gamma$$

But then $(\mathbf{w}_2/\delta, b_2/\delta)$ is a feasible solution for (4) which $||\mathbf{w}_2/\delta|| = \delta^{-1} < \gamma^{-1}$ contradicting optimality of $(\mathbf{v}^*, c^*)$. ■

This means in fact that maximizing the margin is equivalent to minimizing the size of the hyperplane. The optimization problem written in 4 is a quadratic program for which there exist efficient solves. By using them to solve problem 4 we can obtain an optimal solution for the Hard-SVM optimization problem.

R But so how is it that minimizing $||\mathbf{w}||^2$ is equivalent to maximizing the margin? Let us denote the width of the total margin (i.e. the sum of margin from both sides) by l, and let x_+ and x_-

be the positive- and negative support vectors . To calculate the value of l we will project the vector $x_+ - x_-$ onto the normalized normal $\mathbf{w}$:

$$\begin{aligned} l &= \left\langle x_+ - x_-, \frac{\mathbf{w}}{||\mathbf{w}||} \right\rangle \\ &= (\langle x_+, \mathbf{w}\rangle - \langle x_-, \mathbf{w}\rangle)\,||\mathbf{w}|| \\ &= (1 - b - (-1 - b))\,/\,||\mathbf{w}|| \\ &= 2/\,||\mathbf{w}|| \end{aligned}$$

where support vectors satisfy $y_i\,(\langle \mathbf{w}, \mathbf{x}_i\rangle + b)$ and that for positive samples $y_i = 1$ and negative samples $y_i = -1$. This shows how minimizing $||\mathbf{w}||$ maximizes l.

1.4 Soft-SVM

The basic assumption of Hard-SVM is that the training sample is linearly separable. If that is not the case then the optimization problem has no solutions as for any candidate $(\mathbf{w}, b)$ at least one of the constraints $y_i \cdot (\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1$ cannot be satisfied.

But what if the training sample is almost linearly separable? That is, what if most of the samples are linearly separable with only a few violating the constraints by “not too much“? Recall that if $y_i \cdot (\langle \mathbf{w}, \mathbf{x}_i\rangle + b) < 0$ then sample $\mathbf{x}_i$ is on the “wrong side“ of the hyperplane. This means that:

$$\exists \xi_i > 0 \quad s.t. \quad y_i \cdot (\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i$$

Therefore, sample $\mathbf{x}_i$ is on the "wrong" side of the **margin** by an amount proportional to ξ_i (Figure **??**). To allow training samples to violate the constraints “a little“, we modify the optimization problem to:

$$\begin{array}{ll} minimize & ||\mathbf{w}||^2 \\ subject\,to & \begin{cases} y_i \cdot (\langle \mathbf{x}_i, \mathbf{w}\rangle + b) \geq 1 - \xi_i & i = 1, \ldots, m \\ \xi_i \geq 0 \quad \wedge \quad \frac{1}{m}\sum_{i=1}^{m} \xi_i \leq C \end{cases} \end{array} \tag{6}$$

where $C > 0$ is a constant we specify. The variables $\xi_1, \ldots, \xi_m$ are new auxiliary variables we introduce (sometimes known as slack variables). Notice that the larger we choose C to be, the more violations of margin we allow. On the one hand, we want to allow “noisy“ samples to violate the margin, so the hyperplane will ignore them. On the other hand, if we allow too many violations, we lose touch with the training sample and its structure. This is exactly the bias-variance trade-off: the larger C, the more freedom the learner has to “chase after the training sample“.

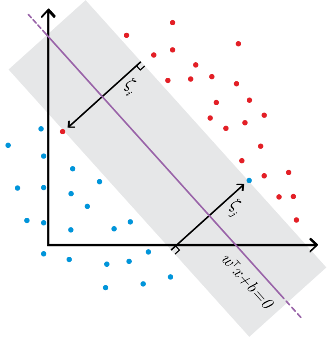

Figure 3: *Slack variables of data-points that are on the "wrong" side of the hyper-plane.*

Instead of specifying C directly, we often prefer working with a slightly different optimization problem, where instead of constraining the value of $\frac{1}{m}\sum \mathbf{x}_i$ we jointly minimize the norm of $\mathbf{w}$ (related to the margin) and the average of ξ_i (corresponding margin violations).

$$\begin{array}{ll} \underset{\mathbf{w},\{\xi_i\}}{minimize} & \lambda\left\|\mathbf{w}\right\|^2 + \frac{1}{m}\sum_{i=1}^{m}\xi_i \\ subject\,to & y_i\cdot(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) \geq 1-\xi_i,\ \xi_i \geq 0 \quad i=1,\ldots,m \end{array} \tag{7}$$

To simplify the above optimization problem let use define the **hinge** loss function:

$$\ell^{hinge}(a) = \max\{0, 1-a\},\ a \in \mathbb{R} \tag{8}$$

Claim 1.2 Given a training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$ and hyperplane $(\mathbf{w}, b)$, the Soft-SVM optimization problem (7) is equivalent to

$$\min_{\mathbf{w},b}\left(\lambda\left\|\mathbf{w}\right\|^2 + L_S^{hinge}((\mathbf{w},b))\right)$$

where $L_S^{hinge}((\mathbf{w},b)) := \frac{1}{m}\sum \ell^{hinge}(y_i\langle \mathbf{x}_i, \mathbf{w}\rangle)$

Proof. Given a specific hyperplane $(\mathbf{w}, b)$ consider the minimization over $\xi_1,\ldots,\xi_m$. Since we defined the auxiliary variables to be nonnegative, the optimal assignment of ξ_i is

$$\xi_i := \begin{cases} 0 & y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) \\ 1 - y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b) & otherwise \end{cases}$$

Thus $\xi_i = \ell^{hinge}(y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b))$ ■

The hyper-parameter λ controls the trade-off between the two norm of $\mathbf{w}$ and the violations of margin. The larger λ, the less sensitive the solution will be to the term $\frac{1}{m}\sum_{i=1}^{m}\xi_i$, and will allow more violations. The smaller λ, the more sensitive and will allow less violations. If we choose to work with this optimization problem to choose h, the constant λ also moves us along different members of a family of learners, each with a different bias-variance tradeoff. λ is known as a **regularization parameter**. This topic is covered in