CR12: Statistical Learning & Applications

Lecture 1. Perceptron

Lecturer: Zaid Harchaoui

1 History

- Rosenblatt (1957) creats the Perceptron, a linear recognition machine
- 60's Theoretical advances in A.I.
- 70's Minsky's book describes all the Perceptrons failures. This book killed the whole discipline (I.A.)
- 80's Support Vector Machines expansion (The return of perceptron)

2 The Perceptron

2.1 Training Set

A training set is defined as:

$$\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$$

Where for all $i \in \{1, ..., n\}$, $x_i \in \mathbf{R}^d$ and $y_i \in \{-1, 1\}$. x_i are the Perceptrons's inputs and y_i are the answers. Perceptrons should learn to use the training set inputs (predictors) to predict the answers (boolean classification). To predict answers the perceptron defines a hyperplane that separates inputs in $\mathbf{R}^d$ according to their belonging classes.

2.2 The margin

2.2.1 Definition

Given a training set, $\mathcal{D}_n$ the margin γ is defined as:

$$\gamma(\mathcal{D}_n) = max(min(yW^T x))$$

Where W is a vector orthogonal to the separating hyperplane ($||W||_2 = 1$).

2.3 Interpretation

> The distance between x and the hyperplane is defined by $|W^T x|$ > For correctly classfied data points:

$$yW^T x = |W^T x|$$
$$sign(W^T x) = y$$

2.3.1 Linearly separable

$\mathcal{D}_n$ is linearly separable if $\gamma(\mathcal{D}_n) > 0$.

2.3.2 remark

Note that when x is expanded, $\gamma(\mathcal{D}_n)$ is also expanded, to avoid expansion we took for all $i \in \{1, ..., n\}$, $||x_i||_2 = 1$.

2.4 Algorithm: Perceptron

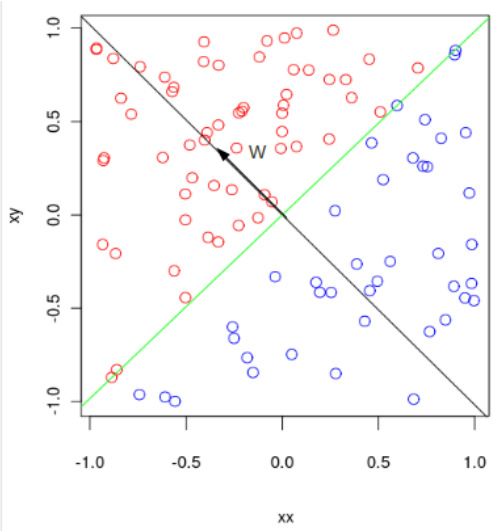

Figure 1: *The data set was created randomly (uniform distribution in [-1,1]X[-1,1]). Lets call xx and xy a point coordinate, all dots with xy>xx where painted in red (class 1) and the other points where painted in blue (class -1). Perceptron separates correctly the dataset, W represents the vector normal to the separator hyperplane.*

2.4.1 The Algorithm

Algorithm Perceptron
Input: Dn

Init: $W_1 = Y_1 * X_1$
While $\exists (x, y) \in \mathcal{D}_n$

if $yW^T x < 0$//Misclassified

Increment: $W_{t+1} = W_t + yx$
Ouput: W

2.4.2 Theorem

Assume that $\mathcal{D}_n$ is linearly separable ($\gamma(\mathcal{D}_n) > 0$) then the perceptron runs for at most:

$$(\frac{1}{\gamma(\mathcal{D}_n)})^2$$

2.4.3 Proof

Lets define $W^* = argmax(min(yW^T x))$, with $||W||_2 = 1$ and $(x, y) \in \mathcal{D}_n$
Lets call W_t the t-th iteration and x_t the t-th example.

At iteration t:
$W^{*T}(W_{t+1} - Wt) = yW^{*T}x_t$
$yW^{*T}x_t \geq \gamma(\mathcal{D}_n)$ (by definition)
by telescoping:
$W^{*T}W_{t+1} \geq \gamma(\mathcal{D}_n) \cdot t$
Using Cauchy-Shwarz :
$W^{*T}W_{t+1} \leq ||W^*||||W_{t+1}|| = ||W_{t+1}||$
$||W_{t+1}||^2 - ||W_t||^2 = 2y_t W^T x_t + ||x||^2$
Then: $t \cdot \gamma(\mathcal{D}_n) \leq W^{*T}W_{t+1} \leq ||W_{t+1}|| \leq \sqrt{t}$

2.4.4 Corollary

For any $\mathcal{D}_n$ where $\gamma(\mathcal{D}_n) > 0$, there exist $(\frac{1}{\gamma(\mathcal{D}_n))^2}$ examples such that a hyperplane can make zero training error. (By construction of the training set: we only choose the usefull data points (missclassified points) to train the perceptron).

2.5 Mistakes made by the Perceptron

2.5.1 Theorem

The number of mistakes ($\#\mathcal{M}$) is at most:
$\mathcal{M} \leq inf(\sum_{t \in \mathcal{M}} l(x_t, y_t, W) + \frac{1}{\gamma}\sqrt{\sum_{t \in \mathcal{M}} ||x_t||^2}))$

2.5.2 Rise of the linear hinge-loss

Lets define the following terms:
$l(x_t, y_t, w) = max(0, 1 - \frac{1}{\gamma} y W^T x)$
$D_t = \sum_{i=1}^{t} l_\gamma(x_i, y_i, W)$
If $D_n = 0$ then $\mathcal{D}_n$ is linearly separable with margin γ
$D_T = D_{t \in \mathcal{M}} l_\gamma(x_t, y_t, W)$

2.5.3 Proof

Let $u \in \mathbf{R}^d$:

$$u^T W_{t+1} - u^T W_t = y_t u^T x_t \geq u^T W_t + \gamma$$
for all u (with $||u||_2 = 1$), and for all γ:
$$u^T W_{t+1} - u^T W_t = y_t u^T x_t \geq u^T W_t + \gamma - \gamma D_t$$
$$u^T W_{t+1} \geq \gamma M - \gamma \textstyle\sum_{t \in \mathcal{M}} l_\gamma(x_t, y_t, W)$$
Cauchy-Shwarz:
$$u^T W_{t+1} \leq ||u|| ||W_{t+1}|| = ||W_{t+1}||$$
$$||W_{t+1}||^2 - ||W_t||^2 = 2 y W^T x_t + ||x_t||^2 \leq ||x_t||^2 \text{ because } sign(y_t W^T x_t) \leq 0$$
$$||W_{t+1}||^2 \leq \textstyle\sum_{t \in \mathcal{M}} ||x_t||^2$$
$$\sqrt{\textstyle\sum_{t \in \mathcal{M}} ||x_t||^2} \geq \gamma M - \gamma(\textstyle\sum_{t \in \mathcal{M}} l_\gamma(x_t, y_t, W))$$

2.6 Pocket Perceptron

2.6.1 The Pocket Perceptron algorithm

Input :	D_n, γ
Find :	$w^* = \mathbf{argmax} \, \#\{i, y_i w^t x_i \geq \gamma\}$.
Algorithm :	I=0
	Run Perceptron on D_n
	After each iteration :
	$I^* = \#\{i, y_i w^t x_i \geq \gamma\}$
	If $I^* > I : I = I^*, w^* = w$
Output :	w^*

2.6.2 Theorem

Let $D_n = \{(x_i, y_i)\}, (x_i, y_i)_{i=1...n} \overset{iid}{\sim} \mathbb{P}_{xy}, \gamma > 0$ and $\bar{\text{w}}$ the result of the Pocket Perceptron over D_n and γ. Then with probability at least $(1 - \delta), \delta > 0$

$$\mathbb{P}_{xy}(y \neq \text{sign}(w^T x)) \leq$$
$$M\gamma(D_n) + \sqrt{M\gamma(D_n) + \frac{4 \log(n/\delta)}{\gamma^2 n}} + \frac{8 \log(n/\delta)}{\gamma^2 n}$$
$$M\gamma(D_n) = n^{-1} \# I(\bar{\text{w}})$$

3 Lifting

Let $X = \mathbb{R}^d, \widetilde{X} = \mathbb{R}^p, p > d$. Then some training set not separable in X can be separable in $\widetilde{X}$. Transition $X \to \widetilde{X}$ can be set as

$$\widetilde{x_i} = [ax_i; \sqrt{1-a^2}l_i]$$

where $a > 0, l_i$ - unary vector in $\mathbb{R}^{p-d}$.

Then $D_n \to \widetilde{D_n} = \{\widetilde{x_i}; y_i\}_{i=1,2\dots n}$ where D_n is linearly separable.

3.1 Algorithm : Lifted perceptron

Input :	$D_n, a > 0$
Algorithm :	For each $i = 1 \dots n$
	$\widetilde{x_i} = [ax_i; \sqrt{1-a^2}l_i]$
	Run Perceptron on $\widetilde{D_n} = \{(\widetilde{x_i}; y_i)\}_{i=1,2\dots n}$
Output :	w

3.1.1 Lemma

Let $w^* \in \mathbb{R}^d, ||w^*|| = 1$, with margin $\gamma > 0$. For each $i \in \{1 \dots n\}$ define $l_i = \max(0, \gamma - yw^{*T}x_i)$. Then $\widetilde{D_n}$ is separable with margin

$$\gamma' = \frac{a\gamma}{\sqrt{1 + \frac{a^2}{1-a^2}\sum_i l_i^2}}$$

.

3.1.2 Proof

Denote $b = \sqrt{1-a^2}, L = \sum_i l_i^2$.

Set $w = \alpha \left[w^*; \frac{a}{b}(y_1 l_1, \dots, y_n l_n)\right]$ with $\alpha = \left(1 + \frac{a^2}{b^2}L\right)^{-1/2}$.

$$\sqrt{1 + \frac{a^2}{b^2}L} y_i w^{*T} x_i = a y_i w^{*T} x_i + a l_i \geq a y_i w^{*T} x_i + a\left(\gamma - y w^{*T} x_i\right) = a\gamma.$$

Thus

$$y_i w^{*T} x_i \geq \frac{a\gamma}{\sqrt{1 + \frac{a^2}{1-a^2}\sum_i l_i^2}}.$$

So $\widetilde{D_n}$ is linearly separable.