Introduction to Perceptron Algorithm

Yuh-Jye Lee

Data Science and Machine Intelligence
National Chiao-Tung University

March 7, 2017

Binary Classification Problem

Given a training dataset

$$S = \{(x^i, y_i) | x^i \in \mathbb{R}^n, y_i \in \{-1, 1\}, i = 1, \ldots, \ell\}$$

$$x^i \in A_+ \Leftrightarrow y_i = 1 \;\&\; x^i \in A_- \Leftrightarrow y_i = -1$$

Main Goal:

Predict the unseen class label for new data

Find a function $f : \mathbb{R}^n \to \mathbb{R}$ by learning from data

$$f(x) \geq 0 \Rightarrow x \in A_+ \text{ and } f(x) < 0 \Rightarrow x \in A_-$$

The simplest function is linear: $f(x) = w^\top x + b$

Perceptron Algorithm (Primal Form) Rosenblatt, 1956

- An on-line and mistake-driven procedure Repeat:

  for $i = 1$ *to* ℓ

  if $y_i(\langle w^k \cdot x^i \rangle + b_k) \leq 0$ *then*

  $w^{k+1} \leftarrow w^k + \eta y_i x^i$

  $b_{k+1} \leftarrow b_k + \eta y_i R^2$

  $$R = \max_{1 \leq i \leq \ell} \|x^i\|$$

  $k \leftarrow k + 1$

  end if

  until no mistakes made within the for loop return: $k, (w^k, b_k)$.

  What is k ?

$y_i(\langle w^{k+1} \cdot x^i \rangle + b_{k+1}) > y_i(\langle w^k \cdot x^i \rangle) + b_k$? $w^{k+1} \longleftarrow w^k + \eta y_i x^i$ and $b_{k+1} \longleftarrow b_k + \eta y_i R^2$

$$
\begin{aligned}
y_i(\langle w^{k+1} \cdot x^i \rangle + b_{k+1}) &= y_i(\langle (w^k + \eta y_i x^i) \cdot x^i \rangle + b_k + \eta y_i R^2) \\
&= y_i(\langle w^k \cdot x^i \rangle + b_k) + y_i(\eta y_i(\langle x^i \cdot x^i \rangle + R^2)) \\
&= y_i(\langle w^k \cdot x^i \rangle + b_k) + \eta(\langle x^i \cdot x^i \rangle + R^2)
\end{aligned}
$$

$$
\boxed{R = \max_{1 \le i \le \ell} \|x^i\|}
$$

Perceptron Algorithm Stop in Finite Steps

Theorem(Novikoff)
Let S be a non-trivial training set, and let

$$R = \max_{1 \leq i \leq \ell} \|x^i\|$$

Suppose that there exists a vector w_{opt} *such that* $\|w_{opt}\| = 1$ and

$$y_i(\langle w_{opt} \cdot x^i \rangle + b_{opt}) \textit{ for } 1 \leq i \leq \ell.$$

Then the number of mistakes made by the on-line perceptron algorithm on S is almost $(\frac{2R}{r})^2$.

Perceptron Algorithm (Dual Form)

$$w = \sum_{i=1}^{\ell} \alpha_i y_i x^i$$

Given a linearly separable training set S and $\alpha = 0$, $\alpha \in \mathbb{R}^{\ell}$, $b = 0$, $R = \max_{1 \leq i \leq \ell} \|x_i\|$.

Repeat: *for* $i = 1$ *to* ℓ

if $y_i(\sum_{j=1}^{\ell} \alpha_j y_j \langle x^j \cdot x^i \rangle + b) \leq 0$ *then*

$\alpha_i \leftarrow \alpha_i + 1$; $b \leftarrow b + y_i R^2$

end if

end for

Until no mistakes made within the for loop return: (α, b)

What We Got in the Dual Form of Perceptron Algorithm?

- The number of updates equals: $\sum_{i=1}^{\ell} \alpha_i = \|\alpha\|_1 \leq (\frac{2R}{r})^2$
- $\alpha_i > 0$ implies that the training point (x_i, y_i) has been misclassified in the training process at least once.
- $\alpha_i = 0$ implies that removing the training point (x_i, y_i) will not affect the final results.
- The training data only appear in the algorithm through the entries of the Gram matrix, $G \in \mathbb{R}^{\ell \times \ell}$ which is defined below:

$$G_{ij} = \langle x_i, x_j \rangle$$

Reference

- C. J. C Burges. "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, Vol. 2, No. 2, (1998) 121-167.
- N. Cristianini and J. Shawe-Taylor. "An Introduction to Support Vector Machines", Cambridge University Press,(2000).