Lecture 8

1 The Perceptron Algorithm

The perceptron algorithm solves the classical problem of online learning of halfspaces. There are m examples $(x_i, l(x_i))$ where $x_i \in \mathbb{R}^n$ are feature vectors and $l(x_i) = \pm 1$ are labels. The examples are correctly classified by a halfspace, that is $l(x_i) = sign(w \cdot x + b)$ for some w, b. An online algorithm is given x_i in some order, asked to predict $l(x_i)$ and then the correct label is revealed. The goal is to minimize the number of classification mistakes.

Wlog we can assume that the separating halfspace passes through the origin and is of form $w.x$ as a $sign(\sum w_i x_i + b)$ can be simulated by adding an extra feature (coordinate) that is always equal to 1. We can further assume that $|x_i| = 1$ as scaling x_i does not change $sign(w.x)$. Let w^* be the unit vector in the direction w.

The angular margin γ for a set of normalized feature vectors is the minimum distance of the x_i from the halfspace $w^*.x = 0$.

$$\gamma = \min_{i \in [m]} |x_i.w^*| \tag{1}$$

A large γ indicates that the classifier is robust, that is perturbing the examples does not change the label, see Figure 1 for an illustration.

1.1 Algorithm:

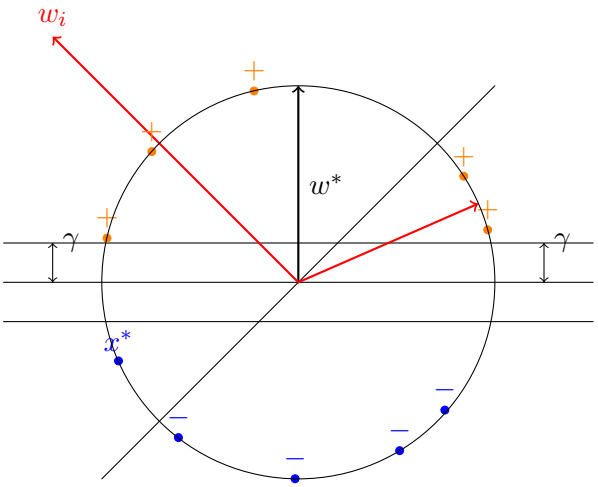
The perceptron algorithm starts with an initial guess $w_1 = 0$ for the halfspace, and does the following on receiving example x_i:

1. Predict $sign(w_i \cdot x)$ as the label for example x_i.
2. If incorrect, update $w_{i+1} = w_i + l(x_i)x_i$ else $w_{i+1} = w_i$.

Claim 1
The perceptron algorithm makes at most $1/\gamma^2$ mistakes if the points x_i are separated with angular margin γ.

Proof: The proof relies on two geometric observations illustrated in Figure 1. If the algorithm makes a mistake on x^*, the unit vector $l(x^*)x^*$ added to w_i has projection at least γ on w^*. This follows as we updating using labels $l(x_i x_i$ and the points are separated by margin γ. If the algorithm makes M mistakes. The length of

$$|w_m| \geq |w_m.w^*| \geq \gamma M \tag{2}$$



Positive + and negative − examples are separated by a halfspace $w^*.x$ with angular margin γ. On making mistake x^*, the perceptron algorithm updates weights by adding $l(x^*)x^*$ to the current weight w_i. The analysis of perceptron relies on the following observations:

1. The projection of $l(x^*)x^*$ onto w^* is at least γ.
2. The unit vector $l(x^*)x^*$ makes an obtuse angle with w_i.

Figure 1: The analysis of the Percepton algorithm in pictures.

Secondly note that the unit vector $l(x^*)x^*$ makes an obtuse angle with w_i for every update, using the law of cosines $|w_i|^2 + 1 \geq |w_{i+1}|^2$. The length of vector w_m is at most $\sqrt{M}$, combining the bounds we have,

$$\sqrt{M} \geq |w_m| \geq \gamma.M \tag{3}$$

It follows that the the perceptron algorithm makes at most $1/\gamma^2$ mistakes. □

1.2 Hinge loss

The notion of the hinge loss TD_γ is introduced to handle the case where there is no separating hyperplane. The hinge loss TD_γ is the minimum total distance through which points x_i must be moved in order to make them separable by an angular margin γ.

The distance TD_γ is parallel to w^* as it is the minimum distance moved, the bound on the projection of w_m onto w^* changes to,

$$|w_m.w^*| \geq \gamma M - TD_\gamma$$

The $l(x^*)x^*$ continues to make an obtuse angle with w_i for all cases, so we have the modified bound $\sqrt{M} \geq \gamma M - TD_\gamma$. Squaring and dropping the positive term TD_γ^2 on the right hand side,

$$M \geq \gamma^2 M^2 - 2\gamma M TD_\gamma \Rightarrow \frac{1}{\gamma^2} + \frac{2TD_\gamma}{\gamma} \geq M \tag{4}$$

The number of mistakes made by the perceptron algorithm can therefore be bounded in terms of the hinge loss.

1.3 Large margin separators:

Consider the variant of the perceptron algorithm that carries out updates when the current hypothesis fails to separate x_i with margin more than $\gamma/2$. For example in Figure 1, the −

point located close to the boundary of the current hypothesis will be treated as a mistake by this algorithm.

Update for the modified perceptron continue to increase the value of $w_i.w^*$ by at least γ as points x_i are separated by an angular margin γ. The observation that $l(x^*)x^*$ makes an obtuse angle with w_i does not hold any more, instead we have that moving x_* by distance $\gamma/2$ along w_i produces a vector making an obtuse angle with w_i.

The obtuse angle condition can be written as $|w_{i+1}|^2 \leq |w_i|^2 + 1 \Rightarrow |w_{i+1}| \leq |w_i| + \frac{1}{2|w_i|}$. Moving x^* by distance $\gamma/2$ produces a vector making an obtuse angle with w_i, so we can apply the triangle inequality to obtain $|w_{i+1}| < |w_i| + \frac{1}{2|w_i|} + \frac{\gamma}{2}$. If $|w_i| \geq 2/\gamma$ we have $|w_{i+1}| \leq |w_i| + \frac{3\gamma}{4}$, summing over all mistakes that occur after $|w_i| \geq \frac{2}{\gamma}$,

$$M\gamma \leq |w_m| \leq \frac{2}{\gamma} + \frac{3\gamma M}{4} \Rightarrow M \leq \frac{8}{\gamma^2} \tag{5}$$

1.4 Kernelization:

There are good algorithms for classifying data separated by halfspaces. If the data is not separated by a halfspace, the kernel trick described in the homework may be used to reduce to the problem to learning halfspaces in an implicit high dimensonal space.