Lecture 8

1 The Perceptron Algorithm

In this lecture we study the classical problem of online learning of halfspaces. The online learning algorithm is given a sequence of m labeled examples $(x_i, l(x_i))$ where $x_i \in \mathbb{R}^n, l(x_i) = \pm 1$ and there exists halfspace w through the origin such that $l(x_i) = sign(w \cdot x)$. The objective of the online algorithm is to minimize the number of classification mistakes. The assumption that the separating halfspace is of the form $w.x$ can be made wlog as a non zero threshold $l(x_i) = sign(\sum w_i x_i + w)$ can be simulated by padding the inputs with an extra coordinate that is always equal to 1.

Let w^* be the unit vector in the direction w, wlog we assume that the x_i are unit vectors as scaling x_i does not change $sign(w.x) = l(x_i)$. The mistake bound for the perceptron algorithm is $1/\gamma^2$ where γ is the angular margin with which hyperplane $w^*.x$ separates the points x_i. An angular margin of γ means that a point x_i must be rotated about the origin by an angle at least $2\arccos(\gamma)$ to change its label.

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

1.1 Perceptron algorithm

1. Initialize $w_1 = 0$.
2. Predict $sign(w_i \cdot x)$ for example x_i.
3. 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 for points x_i that are separated with angular margin γ.

Proof: The proof relies on upper and lower bounds on the potential function $\phi(i) = w_i.w^*$. Initially $\phi(0) = 0$ and if $\phi(i) = |w_i|$, then the classifier predicts all points x_i correctly.

The potential function increases by at least γ each time the algorithm makes a mistake,

$$w_{i+1}.w^* = (w_i + l(x_i)x_i).w^* \geq w_i.w^* + \gamma \tag{2}$$

The classifier is correct if $\phi(i) = |w_i|$ and the potential function ϕ increases by at least γ for each mistake. It suffices to bound $|w_i|$ to count the number of mistakes made by the algorithm. The following invariant maintained during updates: $w_i \rightarrow w_i + l(x_i)x_i$ where the unit vector $l(x_i)x_i$ makes an obtuse angle with w_i,

$$|w_{i+1}|^2 = |w_i + l(x_i)x_i|^2 \leq |w_i|^2 + 1 \tag{3}$$

If the total number of mistakes made by the algorithm is M combining equations (2) and (3) we have,

$$\sqrt{M} \geq |w_m| \geq w_m.w^* \geq \gamma.M \tag{4}$$

It follows that the number of mistakes M made by the perceptron algorithm is at most $1/\gamma^2$. □

The general case: The analysis of the perceptron algorithm assumed there was a hyperplane $w^*.x \geq 0$ separating points x_i with angular margin γ. The notion of the hinge loss TD_γ is introduced to handle the more general case. The hinge loss TD_γ is the minimum total distance through which the points x_i must be moved in order to make them separable by an angular margin γ.

For the general case where the points x_i are not separated by a hyperplane, the lower bound for the potential function (2) at the end of the process changes to,

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

The upper bound (3) continues to hold so the inequality $\sqrt{M} \geq \gamma M - TD_\gamma$ holds. Squaring and dropping the positive term TD_γ^2 on the right hand side,

$$M \geq \gamma^2 M^2 - 2\gamma M TD_\gamma \Rightarrow \frac{1}{\gamma^2} + \frac{2TD_\gamma}{\gamma} \geq M \tag{5}$$

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

Finding hyperplanes with large margins: Consider the variant of the perceptron algorithm that carries out updates when the current hypothesis fails to separate x_i with margin more than $\gamma/2$. The number of mistakes made by the modified perceptron algorithm is at most $8/\gamma^2$.

Each update increases the value of the potential function by γ as in (2) as points x_i are separated by an angular margin γ. We showed that $|w_{i+1}|^2 \leq |w_i|^2 + 1 \Rightarrow |w_{i+1}| \leq |w_i| + \frac{1}{2|w_i|}$ assuming that for all updates x_i makes an obtuse angle with w_i.

For updates made by the modified algorithm, moving x_i by distance $\gamma/2$ perpendicular to w^* produces a vector making an obtuse angle with w_i. The triangle inequality yields $|w_{i+1}| \leq |w_i| + \frac{1}{2|w_i|} + \frac{\gamma}{2}$, for $|w_i| \geq 2/\gamma$ we have $|w_{i+1}| \leq |w_i| + \frac{3\gamma}{4}$.

$$M\gamma \leq |w_m| \leq \frac{2}{\gamma} + \frac{3\gamma M}{4} \Rightarrow M \leq \frac{8}{\gamma^2} \tag{6}$$

The kernel trick: 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 is a general method to reduce to the classification problem to learning halfspaces in some implicit high dimensonal space.