References

1. Pattern Classification

   Duda, Hart, Stork

2. FOUNDATIONS OF MACHINE LEARNING [FOML]

   Mohri et al.

   [select portions]

$$\mathcal{D} = \left\{ (x^{(i)}, y^{(i)}) \;\middle|\; x^{(i)} \in \mathcal{X},\ y^{(i)} \in \{-1, 1\},\ i \in [N] \right\}$$

$$\mathcal{D} \sim P^{(N)}$$

Assume that there exists w^* such that

$$\text{sign}\left((w^*)^T x^{(i)}\right) = y^{(i)}.$$

$i \in [N]$.

$\|x^{(i)}\| \leq R$

$w^{(1)} = 0$

$i = 1, \ldots, N$

$$\Rightarrow y^{(i)} \left(w^{*T} x^{(i)}\right) > 0$$

$$\gamma = \min_i \frac{y^{(i)} (w^*)^T x^{(i)}}{\|w^*\|}.$$

$w^{(t)}$ be the current estimate.

Let $(x^{(i)}, y^{(i)}) \in \mathcal{D}$ such that

$$\text{sign}\left((w^{(t)})^T x^{(i)}\right) \neq y^{(i)}$$

$$\Rightarrow \overset{a)}{y^{(i)}}\left(w^{(t)T} x^{(i)}\right) < 0$$

$$w^{(t+1)} := \begin{cases} w^{(t)} + y^{(i)} x^{(i)} & \text{[update]} \\ w^{(t)} & \text{otherwise} \end{cases}$$

On a linearly separable dataset the perceptron algorithm terminates after making at most $\frac{R^2}{\gamma^2}$ updates

$$\frac{R^2}{\gamma^2} \leq \frac{R^2 \|w^*\|^2}{\min_i \{y_i (w^{*T} x^{(i)})\}^2}$$

Generalization error of the classifier returned by Perceptron?

$$\bar{w} = \sum_{k=1}^{M} y^{(i_k)} x^{(i_k)} \quad \Big| \quad i \in \{1, \ldots, N\}$$
$$\text{sign}(\bar{w}^T x^{(i)}) = y^{(i)}$$

$U : \{i_1, \ldots, i_M\}$ → Indices which are updated

$\mathcal{D}$ be a Sample of size N.

Linearly Separable [There exists $\gamma > 0$].

Let Perceptron Algorithm return a Classifier $h_{\mathcal{D}}^{(P)}$.

$$P\left(h_{\mathcal{D}}^{(P)}(x) \neq y\right) = R\left(h_{\mathcal{D}}^{(P)}\right)$$

$$\underset{\mathcal{D} \sim P^{[N]}}{E} R\left(h_{\mathcal{D}}^{(P)}\right)$$

Expected generalization error by a classifier returned by Perceptron algorithm acting on a <u>Linearly Separable sample</u> of N i.i.d. draws from P.

$$\mathop{E}_{\mathcal{D} \sim p^{[N]}} R\left(h_{\mathcal{D}}^{(p)}\right)$$

Hack:

1) $\bar{\mathcal{D}} \sim p^{(N+1)}$ Sample of size $N+1$.
2) Let it be linearly separable
3) Create $N+1$ Datasets from $\bar{\mathcal{D}}$ by removing ith datapoint.

$$\mathcal{D}^{(i)} = \bar{\mathcal{D}} - \{x^{(i)}, y^{(i)}\} \quad i = 1, \ldots, N+1$$

- Size of $\mathcal{D}^{(i)} = N$
- $\bar{\mathcal{D}}$ is linearly separable
- $\Rightarrow \mathcal{D}^{(i)}$ is linearly separable.

An algorithm A acting on a sample D of size m return a classifier $h_D^{(A)}$.

The Leave one out error of A on D is

$$\bar{R}_D^{LOO}(A) = \frac{1}{m} \sum_{i=1}^{m} 1_{\{h_{D^{(i)}}^{(A)}(x^{(i)}) \neq y^{(i)}\}}$$

$$D^{(i)} : D - \{(x^{(i)}, y^{(i)}\}$$

$$E_{D \sim p(m)} \bar{R}_D^{LOO}(A)$$

Expected LOO error on a random sample of size m.

$$E_{\mathcal{D} \sim p(m)} \bar{R}^{LOO}_{\mathcal{D}}(A)$$

$$= \frac{1}{m} E_{\mathcal{D} \sim p(m)} \sum_{i=1}^{m} 1_{\{h^{(A)}_{\mathcal{D}^{(i)}}(x^{(i)}) \neq y^{(i)})\}}$$

$$E_{\mathcal{D} \sim p(m)} = E_{\mathcal{D}^{(i)} \sim p(m-1)} E_{x^{(i)}, y^{(i)} \sim p}$$

IID

$$E_{X^{(i)}, Y^{(i)} \sim p} 1_{\{h^{(A)}_{\mathcal{D}^{(i)}}(x^{(i)}) \neq y^{(i)}\}}$$

$$= P\left(h^{(A)}_{\mathcal{D}^{(i)}}(x^{(i)}) \neq y^{(i)}\right)$$

$$= R\left(h^{(A)}_{\mathcal{D}^{(i)}}\right)$$

$$E_{\mathcal{D} \sim p^{(m)}} \bar{R}^{LOO}_{\mathcal{D}}(A)$$

$$= \frac{1}{m} \sum_{i=1}^{m} E_{\mathcal{D}^{(i)} \sim p^{(m-1)}} R\left(h^{(A)}_{\mathcal{D}^{(i)}}\right)$$

$$= E_{\mathcal{D} \sim p^{(m-1)}} R\left(h^{(A)}_{\mathcal{D}}\right)$$

Average Generalization error of a classifier derived by an Algorithm A acting on a sample of size $m-1$

$$E_{\mathcal{D} \sim p^{(m-1)}} R\left(h_{\mathcal{D}}^{(A)}\right)$$

= Expected LOO error of A on sample m.

$$\mathbb{E}_{\mathcal{D} \sim p(m)} \bar{R}^{\text{LOO}}_{\mathcal{D}}(A)$$

$$= \mathbb{E}_{\mathcal{D} \sim p(m-1)} R\left(h^{(A)}_{\mathcal{D}}\right)$$

[Reading : Sec 5.2.4.
Lemma 5.3.
FOML]

Perceptron

$$E_{\mathcal{D}\sim p(N)} R\left(h_{\mathcal{D}}^{(P)}\right)$$

$$= E_{\mathcal{D}\sim p(N+1)} \bar{R}_{\mathcal{D}}^{LOO}(P)$$

Consider linearly separable sample $\mathcal{D}$ of size $N+1$

Let the perceptron make M updates using the indices from

$$U = \{i_1, i_2, \ldots, i_M\}$$

$$h_{\mathcal{D}}^{(p)}(x) = \operatorname{sign}\left(\bar{w}_D^T x\right)$$

$$\bar{w}_D = \sum_{k=1}^{M} y^{(i_k)} x^{(i_k)}$$

$$\mathcal{D}^{(j)} = \mathcal{D} \setminus \{x^{(j)}, y^{(j)}\}$$

forall $j \in [N]$

$$\operatorname{sign}\left(\bar{w}_D^T x^{(j)}\right) = y^{(j)}$$

$$1_{\{h_{\mathcal{D}}^{(p)}(x^{(j)}) \neq y^{(j)}\}} = 0$$

(I)

if $j \notin u$

$$h^{(p)}_{\mathcal{D}^{(j)}}(x) = h^{(p)}_{\mathcal{D}}(x)$$

$$\Rightarrow \mathbb{1}\{h^{(p)}_{\mathcal{D}^{(j)}}(x^{(j)}) \neq y^{(j)}\} = 0$$

(II)

Because of (I)

If $j \in u$

$$\mathbb{1}\{h^{(p)}_{\mathcal{D}^{(j)}}(x^{(j)}) \neq y^{(j)}\} \leq 1$$

(III)

$$\bar{R}^{LOO}_{\mathcal{D}}(P) =$$

$$\frac{1}{N+1} \sum_{j=1}^{N+1} 1_{\{h^{(P)}_{\mathcal{D}^{(j)}}(x^{(j)}) \neq y^{(j)}\}}$$

$$\leq \frac{1}{N+1} \left(\sum_{j \notin U} 0 + \sum_{j \in U} 1 \right) = \frac{M}{N+1}$$

π π

$$E_{\mathcal{D} \sim p^{(N)}} R\left(h_{\mathcal{D}}^{(P)}\right)$$

$$\leq E_{\mathcal{D} \sim p^{(N+1)}} \frac{\min\left(M(\mathcal{D}), \frac{R^2(D)}{\gamma^2(\mathcal{D})}\right)}{N+1}$$

[Theorem 8.9
FOML]