Lecture 21 - 25-05-2020

1.1 Pegasos in Kernel space

Objective function was

$$F_\lambda(w) = \frac{1}{m}\sum_{t=1}^{m} h_t(w) + \frac{1}{2}\|w\|^2 \quad w \in \mathbb{R}^d$$

$$w_{T+1} = \frac{1}{\lambda T}\sum_{t=1}^{T} y_{st}\, x_{st}\, I\{h_{st}(w_t) > 0\} \qquad s_1, ..., s_t \quad (realised\, draws\, in\, training)$$

$$K \qquad H_k = \{\sum_i \alpha_i k(x_i, \cdot), \alpha_i, x_i\} \qquad g \in H_k$$

$$F_\lambda = \frac{1}{m}\sum_{t=1}^{m} h_t(g) + \frac{1}{2}\|g\|^2 \qquad h_t(g) = [1 - y_t\, g(x_t)\,]_+$$

$$g_{T+1} = \frac{1}{\lambda T}\sum_{t=1}^{T} y_{st}\, k(x_st, \cdot)\, I\{h_{st}(g_t) > 0\}$$

where red part is v_{st}

1.2 Stability

A way to bound the risk of a predictor.
Controlling the variance error and leave to the user the job to minimise the bias.
Variance error is due to the fact that the predictor an algorithm generate from the training set will depends strongly on the training set itself. If we perturb the training set our predictor will change a lot.

Minimisation of training error $\Rightarrow$ predictor changes if training set if perturbed. $\Rightarrow$ risk of overfitting
Stability is the opposite since avoid overfitting when we perturbing the training set.

- S Training set $(x_t, y_t)...(x_m, y_m)$
- loss function ℓ

- distribution D

$h : X \to Y$ $\ell_D(h)$ risk of h
$z_t = (x_t, y_t)$ $\ell(h_t, z_t) = \ell(h(x_t), y_t)$

$$\hat{\ell}_s(h) = \frac{1}{m} \sum_{t=1}^{m} \ell(h, z_t)$$

Perturbation $z'_t = (x'_t, y'_t)$ also drawn from D
$S^{(t)}$**is** S **where** z_t **is replaced by** z'_t $h_s = A(S)$
A learning algorithm is ε-stable $(\varepsilon > 0)$ $h_s^{(t)} = A(S^{(t)})$

$$\ell(h_s^{(t)}, z_t) - \ell(h_s, z_t)$$

we expect this subtraction result to be positive.

$$\mathbb{E}\left[\, \ell(h_s^{(t)}, z_t) - \ell(h_s, z_t) \,\right] \le \varepsilon \qquad \forall t = 1, ...m$$

where $\mathbb{E}[\ ] \to s, z'_t$
z_t and z'_t come from D both

$$\mathbb{E}\left[\ell(h_s, z'_t) - \ell(h_s^{(t)}, z'_t)\right] \le \varepsilon$$

Theorem
If A is ε-stable, then

$$\mathbb{E}\left[\ell_D(h_s) - \hat{\ell}_s(h_s)\right] \le \varepsilon$$

Proof: S $z_t = (x_t, y_t)$ s' $z'_t = (x'_t, y'_t)$ D

$$\mathbb{E}\left[\hat{\ell}_s(h_s)\right] = \mathbb{E}\left[\frac{1}{m} \sum_{t=1}^{m} \ell(h_s, z_t)\right] = \frac{1}{m} \sum_{t=1}^{m} \mathbb{E}\left[\ell(h_s, z_t)\right] = \frac{1}{m} \sum_{t=1}^{m} \mathbb{E}\left[\ell(h_s^{(t)}, z'_t)\right]$$

$$\ell_D(h_s) = \mathbb{E}\left[\ell(h_s, z'_t) | S\right] = \frac{1}{m} \sum_{t=1}^{m} \mathbb{E}\left[\ell(h_s, z'_t)\right]$$

Average with respect to random draw of S
$\mathbb{E}\left[\ell_D(h_s)\right] = \frac{1}{m} \sum_{t=1}^{m} \mathbb{E}\left[\ell(h_s, z'_t)\right]$

$$\mathbb{E}\left[\ell_D(h_s) - \hat{\ell}_s(h_s)\right] = \frac{1}{m} \sum_{t=1}^{m} \mathbb{E}\left[\ell(h_s, z'_t) - \ell(h_s^{(t)}, z'_t\right] \le \varepsilon$$

A stable algorithm is not overfitting (but they still underfit!).
So if an *ERM* algorithm is ε-stable, it would be pretty good.

Theorem
If A is ε-stable and it approximates ERM in a class H:

$$\hat{\ell}_s \leq \min_{h \in H} \hat{\ell}_s(h) + \gamma \qquad \forall s,\ h_s = A(S)$$

for some $\gamma > 0$, then:

$$\mathbb{E}\left[\ell_D(h_s)\right] \leq \min_{h \in H} \ell_D(h) + \varepsilon + \gamma$$

Proof

$$\mathbb{E}\left[\ell_D(h_s)\right] = \mathbb{E}\left[\ell_D(h_s) - \hat{\ell}_s(h_s)\right] + \mathbb{E}\left[\hat{\ell}_s(h_s) - \hat{\ell}_s(h^*)\right] + \mathbb{E}\left[\ell_s(h^*)\right]$$
$$h^* = arg\min_{h \in H}\ \ell_D(h)$$
$$\mathbb{E}\left[\hat{\ell}(h^*)\right] = \ell_D(h^*) \ \longrightarrow\ \mathbb{E}\left[\frac{1}{m}\sum_t \ell(h^*, z_t)\right] = \frac{1}{m}\sum_t \mathbb{E}\left[\ell(h^*, z_t)\right]$$

where **red** is $\ell_D(h^*)$

$\ell(\cdot, z)$ is a convex function $\ell(w, z)$
$\exists L > 0 \qquad |\ell(w,z) - \ell(z,z)| \leq L\|w - w'\|$
$z = (x, y)$
In the case of SVM, $\ell(w,z) = \left[y\, w^T\, x\right]_+$ $\exists L > 0 \quad \forall z\ \forall w, w'$

$$|\ell(w,z) - \ell(w',z)| \leq L\|w - w'\|$$

where *ell* is **Lipschitz**

Theorem
Let ℓ be convex, Lipschitz and differentiable.
Consider $A \qquad A(S) = w_s$ where

$$w_s = arg\min_{w \in \mathbb{R}^d}\left(\hat{\ell}_s(w) + \frac{\lambda}{2}\|w\|^2\right)$$

If ℓ is hinge loss, then A is SVM.
then A is $\frac{(2\,L)^2}{\lambda m}$-stable $\qquad \forall \lambda > 0$

Proof
Fix $\lambda > 0 \qquad F_s(w) = \hat{\ell}_s(w) + \frac{\lambda}{2}\|w\|^2$

$$w_s = arg\min_{w \in \mathbb{R}^d} F_s(w) \qquad w_s^{(t)} = arg\min_{w \in \mathbb{R}^d} F_s^{(t)}(w)$$

$$\ell(w_s, z_t') - \ell(w_s^{(t)}, z_t') \leq \varepsilon \qquad \forall s, z_t' \; \forall t$$

Use Lipschtiz

$$|\ell(w_s, z_t') - \ell(w_s^{(t)}, z_t')| \leq L \|w_s - w_s^{(t)}\|$$

$w = w_s, \;\; w' = w_s^{(t)}$

$$F_s(w') - F_s(w) = \hat{\ell}(w') - \hat{\ell}(w) + \frac{1}{2}\|w'\|^2 - \frac{\lambda}{2}\|w\|^2 \;\; =$$

$$= \;\hat{\ell}_s^{(t)}(w') - \hat{\ell}_s^{(t)} + \frac{1}{m}\left(\ell(w', z_t) - \ell(w, z_t)\right) - \frac{1}{m}\left(\ell(w', z_t') - \ell(w, z_t')\right) + \frac{\lambda}{2}(\|w'\|^2 - \|w\|^2) \;\; =$$

$$= \;\; {\color{red} F_s^{(t)}(w') - F_S^{(t)}(w)} + \frac{1}{m}(\ell(w', z_t) - \ell(w, z_t)) - \frac{1}{m}(\ell(w', z_t') - \ell(w, z_t')) \;\; \leq$$

where **red** is ≤ 0

$$\leq \;\; |\frac{1}{m}\ell(w', z_t) - \ell(w, z_t)| + \frac{1}{m}|\ell(w', z_t') - \ell(w, z_t')| \leq$$

—— MANCAAAAAAA —-

$$F_s(w) - F_s(w') \leq \frac{2\,L}{m}\|w - w'\|$$

F_s is λ-SC $\qquad F_s(w') \geq F_s(w) + \nabla F_s(w)^T(w' - w) + \frac{\lambda}{2}\|w - w'\|^2$ Since w is minimiser of F_s the gradiant $\nabla F_s(w)^T = 0$ Therefore:

$$F_s(w') - F_s(w) \geq \frac{1}{2}\|w - w'\|^2$$

$$\frac{\lambda}{2}\|w - w'\|^2 \geq \frac{2\,L}{m}\|w - w'\| \Rightarrow \|w - w'\| \leq \frac{4\,L}{\lambda\, m}$$

$$\ell(w_s, z_t') - \ell(w_s^{(t)}, z_t') \leq \frac{4\,L^2}{\lambda\, m}$$

We now know the stability of the SVM.

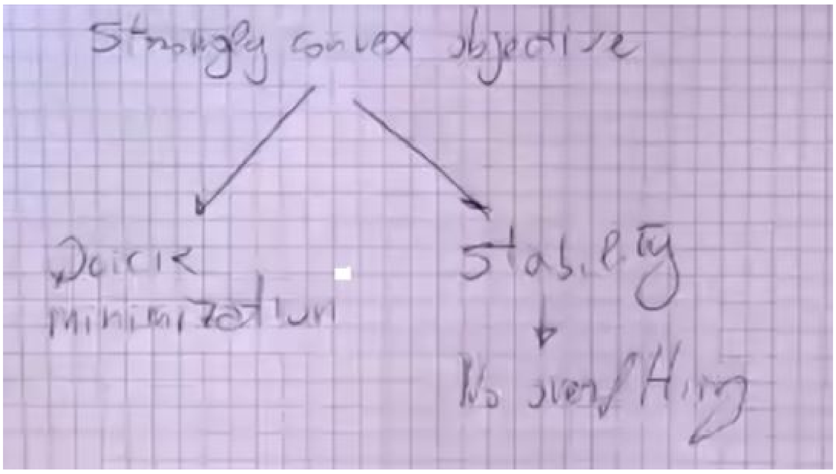
Strongly convex objective
Quick minimization
Stability
No overfitting

Figure 1.1: