Prof. Dahleh, Prof. Mitter — Lecture 9
Scribed by Kostas Bimpikis — Th 3/8

We have seen in the previous lecture the *maximum margin hyperplane problem* can be expressed as:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\psi^{'}\psi \\ \text{subject to} & y_i(\psi^{'}x_i - b) \geq 1 \quad \forall i \end{array}$$

Let ψ° and b° denote the optimal solution of the above problem. Then it is straightforward to see that the margin is equal to $\frac{1}{|\psi^\circ|}$. Using duality we can conclude that the optimal hyperplane can be written as a linear combination of the data points, i.e.

$$\psi^\circ = \sum_{i=1}^{\ell} a_i^\circ y_i x_i$$

with

$$a_i^\circ(y_i(\psi^\circ x_i - b^\circ) - 1) = 0,$$

so essentially $a_i^\circ > 0$ only for those vectors that lie on the margin, i.e. $y_i(\psi^\circ x_i - b^\circ) - 1 = 0$, which we call the *support vectors.*

1 Statistical Properties of SVMs

Note that the solution to the dual of the maximum margin hyperplane problem is not necessarily unique and each dual solution defines a set of support vectors. Let K_ℓ denote the number of *essential* support vectors, i.e. the vectors that belong to the intersection of the support sets. Obviously, $K_\ell \leq n$. Finally, note that from above we can write:

$$\psi^{\circ\prime} x_j - b^\circ = \sum_{i=1}^{\ell} a_i^\circ y_i x_i' x_j - b^\circ = f(x, x_j) - b^\circ$$

Next, we define a mapping from the sequences of data points to an element of the model class, e.g. the set of separating hyperplanes. This effectively represent an algorithm. In particular,

$$\alpha_\ell : \mathcal{Z}^\ell \rightarrow \{ \text{ separating hyperplanes } \}$$

$$z_1 = (x_1, y_1), \cdots, z_\ell = (x_\ell, y_\ell) \mapsto \{ \text{ optimal hyperplane } \}$$

As before we can define the expected risk of that mapping as:

$$\mathbf{E}[R(\alpha_\ell)] = \mathbf{E}[L(z, \alpha_\ell(Z^\ell))] = \mathbf{E}_{Z_1,\cdots,Z_\ell}\mathbf{E}_{Z|Z_1,\cdots,Z_\ell}[L(Z, \alpha_\ell(Z_1, \cdots, Z_\ell))]$$

Note that the empirical risk is 0, since data is separable and we can always pick a hyperplane that classifies all data points perfectly.

First we show the following proposition,

Proposition 1. $\mathbf{E}[R(\alpha_\ell)] \leq \frac{\mathbf{E}[K_{\ell+1}]}{\ell+1}$

Proof. The proof is using the "leave one out one at a time" validation method. The main idea is that points far from the margin do not really matter and can be discarded. In particular, let $z_1, \cdots, z_{\ell+1}$ be a sequence of samples. Let z_{-i} denote the sequence that contains all but the i^{th} sample. Also let the loss function

$$L((x, y), \alpha_\ell(.)) = \begin{cases} 1 \text{ if } \alpha_\ell(.) \text{ misclassifies } (x, y) \\ 0 \text{ otherwise} \end{cases}$$

Finally, define the cross validation statistic as

$$\bar{\mathbf{L}}(z_1, \cdots, z_{\ell+1}) = \frac{1}{\ell+1} \sum_{i=1}^{\ell+1} L(z_i, \alpha_\ell(z_{-i}))$$

Then,

Lemma 1. $\mathbf{E}[R(\alpha_\ell)] = \mathbf{E}[\bar{\boldsymbol{L}}]$

Proof.

$$\mathbf{E}[\bar{\mathbf{L}}] = \frac{1}{\ell+1} \sum_{i=1}^{\ell+1} \mathbf{E}[L(z_i, \alpha_\ell(z_{-i}))] = \frac{1}{\ell+1}(\ell+1)\mathbf{E}[R(\alpha_\ell)] = \mathbf{E}[R(\alpha_\ell)]$$

□

Finally, note that if $L(z_i, \alpha_\ell(z_{-i})) = 1$ then z_i has to belong to the set of essential support vectors. Thus, we conclude that $\mathbf{E}[R(\alpha_\ell)] = \mathbf{E}[\bar{\mathbf{L}}] \leq \frac{\mathbf{E}[K_{\ell+1}]}{\ell+1}$ □

2 SVM Extensions via Kernals

Now we are ready to define *support vector machines* as a simply a mapping from $\mathbf{R}^n$ to $\mathbf{R}^m$ for the data points, where typically $m > n$. Namely,

$$\phi : \mathbf{R}^n \to \mathbf{R}^m$$

$$\text{such that } (x_i, y_i) \mapsto (\phi(x_i), y_i)$$

We can rewrite ψ° as

$$\psi^\circ = \sum_{i=1}^{\ell} a_i^\circ y_i \phi(x_i)$$

and

$$\psi^\circ \phi(x_j) - b^\circ = \sum_{i=1}^{\ell} a_i^\circ y_i \phi(x_i)\phi(x_j) - b^\circ$$

Define $\phi(x_i)\phi(x_j)$ as $K(x_i, x_j)$, the kernel function. One question that arises naturally at this point is which kernels best separate the data. Also, given a kernel function $K(x_i, x_j)$, does there exist a mapping ϕ such that $K(x_i, x_j) = \phi(x_i)\phi(x_j)$?

The answer to the second question is given by *Mercer's Theorem.* More precisely, suppose x is mapped to some Hilbert space:

$$\phi(x) = (\phi_1(x), \phi_2(x), \cdots).$$

Theorem 1. *(Mercer's)*
A continuous symmetric function $K(u, v)$ in $L_2(C)$, C compact, can be expanded as:

$$K(u, v) = \sum_{k=1}^{\infty} a_k \phi_k(u) \phi_k(v)$$

where $a_k > 0$, if and only if

$$\int_C \int_C K(u, v) g(u) g(v) \mathrm{d}u \mathrm{d}v \geq 0,$$

for all $g \in L_2(C)$.

Followingly, we give a few examples of kernel functions:

- $K(u, v) = [u'v + 1]^d$ (polynomial function)
- $K(u, v) = exp(-\gamma|u - v|^2)$ (radial function)
- $K(u, v) = \frac{1}{1+exp(cu'v+1)}$ (segmoidal function)

3 Extensions of VC Theory to General (Bounded) Loss Functions

In this section we will consider the case of non-binary loss functions (e.g. regression problem). The main assumpion is that the loss function $L(z,a)$ is bounded, i.e. $b_1 \leq L(z,a) \leq b_2, \forall z$. Then, similarly we can define:

$$B^\Lambda(z^\ell) = \{L(z_1,a), \cdots, L(z_\ell,a)\} \subseteq [a,b]^\ell$$

which is a sequence of real numbers.

To define a similar notion as the VC dimension, we consider ϵ-covers of the $B^\Lambda(z^\ell)$ object, i.e. $B^\Lambda(z^\ell)$ is contained in $\bigcup_i \text{Ball}_\epsilon(r_i)$. Associated with each $B^\Lambda(z^\ell)$ and ϵ is the minimal ϵ-cover (smallest number of such balls) which we denote by $\hat{N}(\epsilon, z^1, \cdots, z^\ell)$.

Similarly with the classification case we can define the *annealed entropy* as:

$$H^\Lambda{}_{annl}(\epsilon,\ell) = \ln \mathbf{E}[N^\Lambda(\epsilon, Z_1, \cdots, Z_\ell)]$$

and the growth function as:

$$G^\Lambda(\epsilon,\ell) = \sup_{z_1,\cdots,z_\ell} \ln N^\Lambda(\epsilon, z_1, \cdots, z_\ell)$$

The results that follow are similar to the indicator function case.