Statistical Learning Theory: The Structural Risk Minimization Principle

Jose Principe, Ph.D.

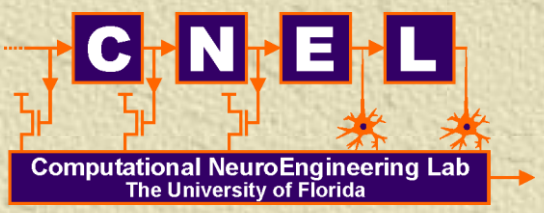
and

Sohan Seth

Distinguished Professor ECE, BME

Computational NeuroEngineering Laboratory and

principe@cnel.ufl.edu



Statistical Learning Theory

Now that we have an idea what is ERM (empirical risk minimization) we are ready to appreciate Vladimir Vapnik's contribution to Statistical Learning Theory.

Vapnik challenged the prevailing view in SLT about the compromise between machine capacity and generalization that moved the theory for many, many years (Akaike, MDL, Bayesian criteria).

Indeed all these criteria penalize the model order (machine capacity) to achieve better generalization.

Statistical Learning Theory

For instance, Akaike's criterion states that

$$\min_{l} AIC(l) = N \log J(l) + 2l$$

where l is the model order, N the number of samples and J(l) the MSE (empiric error) for model l.

MDL is similar (the penalty is slightly different)

$$\min_{l} AIC(l) = N \log J(l) + \frac{l}{2} \log(l)$$

For nonlinear systems the issue is much harder but it was known that the value of the parameters, not only their number was involved in the definition.

Structural Risk Minimization (SRM) Principle

Vapnik posed four questions that need to be addressed in the design of learning machines (LMs):

1. What are the necessary and sufficient conditions for consistency of a learning process.
2. How fast is the rate of convergence to the solution.
3. How can we control the generalization ability of the LM.
4. How can we construct an algorithm that implement these pre requisites.

Structural Risk Minimization (SRM) principle

Vapnik argues that the necessary and sufficient conditions of consistency (generalization) of the ERM principle depend on the capacity of the set of functions implemented by the learning machine.

He showed that the VC (Vapnik- Chervonenkis) dimension *h* provides a way to estimate an upper bound of the Bayes error.

h of a set of functions is defined as the maximum number of vectors that can be separated into two classes in all 2^h possible ways, using functions of the set.

For linear discriminants in R^N this is exactly N+1.

Structural Risk Minimization (SRM) principle

VC dimension of a learning machine is what counts for generalization. In fact he proved that the Risk of a learning machine f(x,w) of size *l* parametrized by *w* is bounded by

$$R_{VC}(w) \leq R_{emp}(w) + \Phi(\frac{N}{h})$$

The second term is a confidence interval and we see that what matters is the number of samples, N the training error and VC dimension (not on the number of parameters *l*).
This was not expected!
Notice that we can bound the generalization without any constraint of the size of the network by decreasing the error and minimizing the confidence interval: That is, search for the LM that has the smallest VC dimension.

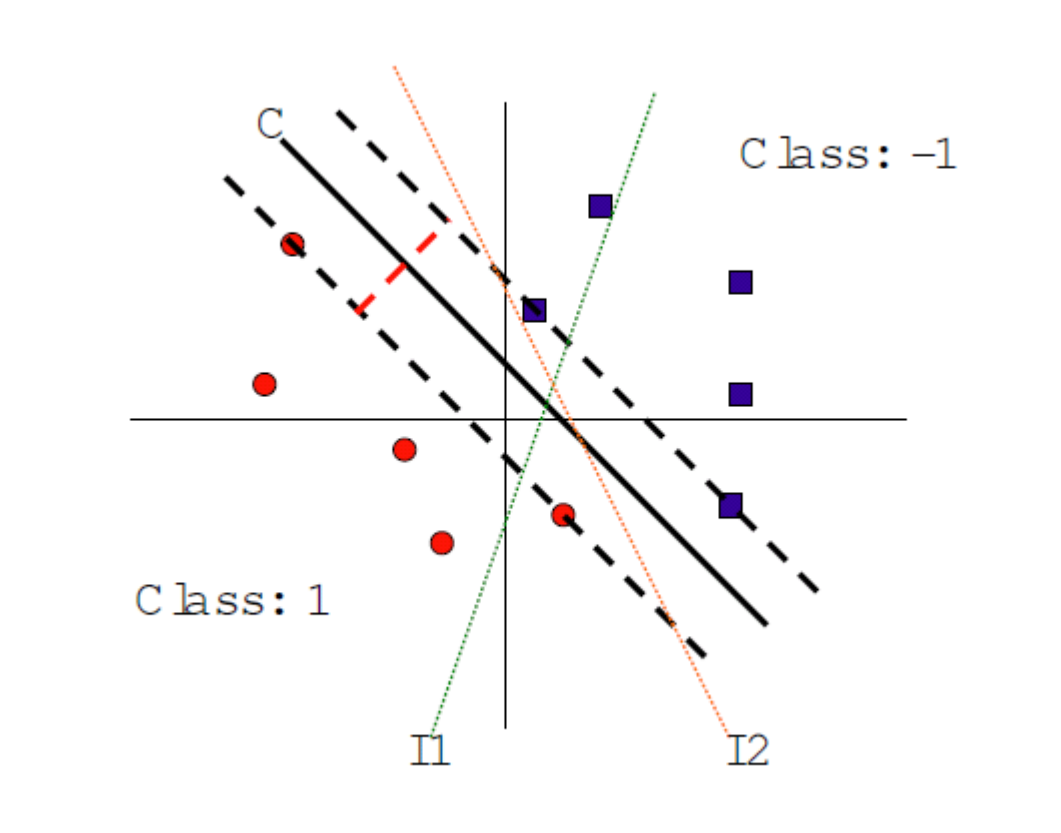
Structural Risk Minimization (SRM) principle

The problem is that the VC dimension is in general very difficult to compute for arbitrary classifiers…. But it is easy for hyperplanes.

So Vapnik and colleagues worked with linear machines, not in the input space, but in a special infinite dimensional space of functions called a Reproducing Kernel Hilbert Space (RKHS). It can be shown that for some RKHS, linear classifiers are universal classifiers (remember Cover's Theorem?) and they are called Support Vector Machines.

What one needs to do is to find a way to maximize the margin ($\alpha = df(x,w)$) of the linear classifier.

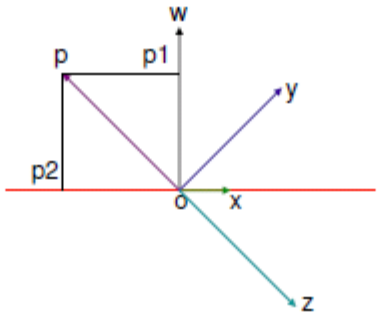
Structural Risk Minimization (SRM) principle



Construct decision surface where the distance between the closest samples to the boundary (the margin) is maximized

SVM Design

Basic geometry:



Equation of the hyperplane

$$g(\mathbf{x}) = \mathbf{w}^\top(\mathbf{x} - \mathbf{x}_0) = \mathbf{w}^\top\mathbf{x} + b = 0$$

Sides of the hyperplane

$$\mathbf{w}^\top\mathbf{y} + b > 0 \text{ and } \mathbf{wz} + b < 0$$

Projection on the hyperplane

$$\mathbf{p} = \mathbf{p}_1 + \mathbf{p}_2 \text{ where } \mathbf{p}_1 = r\frac{\mathbf{w}}{||\mathbf{w}||}$$

SVM Design

Classification Given samples and corresponding class labels i.e. $\{\mathbf{x}_i, d_i\}_{i=1}^{n}$

$$d_i = 1 \text{ if } \mathbf{w}^\top \mathbf{x}_i + b \geq 1$$
$$d_i = -1 \text{ if } \mathbf{w}^\top \mathbf{x}_i + b \leq -1$$

i.e.

$$d_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$$

Margin

$$\frac{2}{||\mathbf{w}||}$$

Problem

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^\top \mathbf{w} \text{ such that } d_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \, \forall \, i$$

SVM Design

Use Lagrange multipliers

$$\max_{(\alpha_1,\ldots,\alpha_n)} \min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^\top\mathbf{w} - \sum_{i=1}^{n} \alpha_i[d_i(\mathbf{w}^\top\mathbf{x}_i + b) - 1] \text{ such that } \alpha_i \geq 0$$

- Derivative with respect to $\mathbf{w}$ is zero i.e.

$$\frac{\partial J}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i d_i \mathbf{x}_i$$

  $\mathbf{w}$ is in the span of the samples

- Derivative with respect to b is zero i.e.

$$\frac{\partial J}{\partial b} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i d_i$$

SVM Design

KKT condition

$$\alpha_i[d_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1] = 0$$

Only α_i's with $d_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$ can take nonzero values

Sparsity! Support vectors!

Dual problem

$$\max_{(\alpha_i,\ldots,\alpha_n)} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^\top \mathbf{x}_j \text{ such that } \sum_{i=1}^{n} \alpha_i d_i = 0, \alpha_i \geq 0$$

Quadratic optimization problem

SVM Design

Nonlinear decision surface Use similar ideas as in RBF.

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i d_i \varphi(\mathbf{x})$$

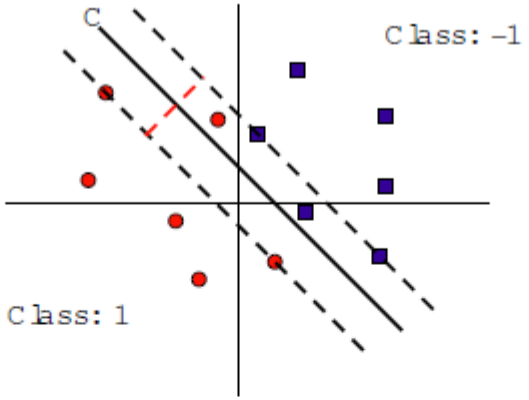
But wait! Note that, α depends on $\mathbf{x}$ through the inner product $< \mathbf{x}|\mathbf{y} >_1 = \mathbf{x}^\top \mathbf{y}$.

Specify the inner product without specifying the nonlinear functions explicitly. For example,

$$< \varphi(\mathbf{x})|\varphi(\mathbf{y}) >_2 = (\mathbf{x}^\top \mathbf{y})^2$$

For this example if $\mathbf{x} = [x_1, x_2]$ then $\varphi(\mathbf{x}) = [x_1^2, x_2^2, \sqrt{2}x_1 x_2]$

SVM Design



Slack variable

$$d_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

$0 < \xi_i \leq 1$:Correct classification but inside margin

$\xi_i > 1$:On the wrong side!

SVM Design

Primal problem

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_{i=1}^{n}\xi_i \text{ such that } d_i(\mathbf{w}^\top\mathbf{x}_i + b) \geq 1\,\forall\, i$$

C acts as regularizer.

Dual problem

$$\max_{(\alpha_1,\ldots,\alpha_n)} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j d_i d_j \mathbf{x}_i^\top\mathbf{x}_j \text{ s.t. } \sum_{i=1}^{n}\alpha_i d_i = 0, 0 \leq \alpha_i \leq C$$