

Advances in Statistical Learning Theory based on VC Dimension and Structured Risk Minimization

Sviluppi della teoria dell'apprendimento statistico basata sulla dimensione di Vapnik-Cervonenkis (VC) e sulla minimizzazione del rischio strutturato.

Michel Bera

KXEN Inc. Chief Scientist

<http://www.kxen.com>

michel.bera@kxen.com

Riassunto: Viene fornita una sintetica rassegna della teoria di Vapnik riguardo alla modellizzazione statistica, ed illustrata la sua principale applicazione nell'ambito dei cosiddetti sistemi d'apprendimento lineari generalizzati. A tale scopo, dopo aver definito la dimensione VC e la fondamentale disuguaglianza di Vapnik concernente la Minimizzazione del Rischio Empirico, viene impostata la strategia di Vapnik tendente a costruire, attraverso la Minimizzazione del Rischio Empirico, una procedura di stima di una generica funzione di previsione/classificazione che presenti, al contempo, un buon adattamento ai dati e una buona capacità di generalizzazione. Le nozioni di spazio osservazionale esteso, modello duale e spazi di Hilbert con “nuclei riproduttivi” vengono infine esaminate, al fine di fondare la stima di una funzione di classificazione nel contesto dei sistemi lineari generalizzati, attraverso la teoria delle “Macchine Vettoriali di Supporto”.

Keywords: Statistical Learning Theory, Support Vector Machines, Structured Risk Minimization, Modelization, Generalization.

1 Changing the model problem setup

1.1 A history of statistical inference : Fisher versus Vapnik

The basic idea of Vladimir Vapnik theory [Vapnik (1999, 1998)] is a redefinition of the build up of the problem of model setup.

As Kants put it, there are three steps to a scientific theory: setting of the problem, resolution of the problem, and proofs.

By “model” here we focus on the following problem:

- we are given L observations $(\vec{x}_i)$, $i = 1, \dots, L$ from $\mathbb{R}^n$, together with a supplementary variable y . y is the target.
- we want to set-up a model, that is find a function $\hat{y} = f(\vec{x})$ from $\mathbb{R}^n$ to $\mathbb{R}$, to estimate y given $\vec{x}$.
- we want $\hat{y}_i$ to be “close” to y_i on the given L observations (good fit on the “training set”) and we want, given a new observation $(\vec{x}, y)$ from the same probabilistic universe, the estimated value $\hat{y} = f(\vec{x})$ to remain “close” to the real y (that is, the model f generalizes well or is robust).

There are basically two ways to go to deal with such problems. A first way is to go and look for general theorems, that would hold true “whatever the data”, ie theorems

that do not depend of the probabilistic law underlying the data $(\vec{x}, y)$. A famous theorem “whatever the data” is Kolmogorov-Smirnov theorem, that gives the limit distribution of the random variable

$$D_L = \sqrt{L} \sup \{|F(\vec{x}) - F_L(\vec{x})|, \vec{x} \in \mathbb{R}^n\}$$

where

$$F_L(\vec{x}) = \frac{1}{L} \{\text{number of } \vec{x}_i \leq \vec{x} \mid i = 1, \dots, L\}$$

$(\vec{x}_i)$, $i = 1, \dots, L$ being a sample of L iid random variables of unknown distribution F , and $\{\vec{x}_i \leq \vec{x}\}$ being “true” when all n components of x_i are less than to the corresponding component of x . For every L , the law of D_L does not depend on F .

A second way, more simple, was designed by people such as Fisher and his disciples. The idea is this time to make hypothesis on the underlying distribution F for the data, to build statistics out of the given sample $\{(\vec{x}_i) \mid i = 1, \dots, L\}$. Under the hypothesis F (the *H0 hypothesis*: the data come from a specific distribution F , eg a log-normal or Poisson distribution) one can compute the probabilistic distribution of those statistics, and define tests for them, eg “are the values measured significantly away from zero?”.

As we shall see, one of the major model we shall focus on here is the linear model, used for classification between “yes” and “no”: $y \geq 0$ means “yes”, and $y < 0$ means “no”. The linear model has equation

$$f(x) = \langle \vec{w} | \vec{x} \rangle + b$$

where $\langle \vec{w} | \vec{x} \rangle$ is the scalar product of vector $\vec{x}$ and parameter vector $\vec{w}$ ($\vec{x}, \vec{w} \in \mathbb{R}^n$), and $b \in \mathbb{R}$ is a parameter.

In a very common model, named multilinear regression, we find that good estimates of $\vec{w}$ and b , if the $(\vec{x}, y)$ are supposed to follow a multinormal distribution F , will be derived from the variance-covariance matrix of the observations $\{(\vec{x}_i, y_i) \mid i = 1, \dots, L\}$; we will use tests for the validity of parameters $\vec{w} = \{w_j \mid j = 1, \dots, n\}$ and b (Fisher and Student tests, etc.) to finally decide of the existence of a proper model $y = \langle \vec{w} | \vec{x} \rangle + b$.

This second approach showed up to be very efficient, and very soon the General Theory of Statistics was built, solving in a very appropriate and efficient way many similar problems.

In the 1960s, the first digital computers started to store and number crunch growing amounts of data, with an increasing number of attributes describing those data. Very soon though, it became clear in the 70s that a strange phenomenon was occurring, to be named the *malediction of large number of attributes*, or *curse of dimensionality*. When models were built the Fisher way, although all the hypotheses and testing procedures were appropriately executed, giving good test results and green lights to apply the models that were built, those models did not produce good results on new data from the same probability space, they were *non robust*.

In the 1970s, new methods tried to overcome the dimension malediction:

- some of them (Factor/Principal Component Analysis) were based on specific properties of the multinormal distribution, eg. strong and elegant relationships between orthogonality and statistical independence. Surprisingly, though the underlying data distributions are seldom fully multinormal, such methods gave good results, but were mentioned as “empirical”, to remain nice, by some mathematicians.

- other methods (Neural Networks, Decision Trees) came from the fast growing field of Learning Theory. Those methods gave very often very good results, but they also had a hard time the same ways as PCAs coming out from the ghetto of “black boxes”, “numerical empirical algorithms” with no real mathematical proofs.

At the same time, on the Russian side, who had no computing power, important mathematical theories were discovered and developed in the 50s, 60s and 70s:

- ill-posed problems, by Tichonov, addressing problems of the kind

$$\int H(x, y)f(x)dx = g(y) \quad (1)$$

where H and g are known, f is to be found, and small variations of g can lead to huge variations of f . The theory of regularization suggests to solve an slightly different problem than (1): Minimize for $f \in F$, F being a family of functions, and λ a given “small” constant,

$$G(f) = \left\| \int H(x, y)f(x)dx - g(y) \right\|^2 + \lambda J(f)$$

where $J(f)$ is a penalty functional, with higher values for functions f showing a “hectic” behaviour (hence the term “regularization”).

- Vapnik - Chernovenkis (VC) dimension of a family of functions, and its relationship to model robustness through the principles of Empirical Risk Minimization and Structured Risk Minimization: these concepts and others they developed, such as the theory of stochastic ill-posed problems, open and lead since 1995, with the second edition of Vapnik’s book at Springer Verlag, the way to an entirely new theory: the Theory of Statistical Learning.

In modern Statistical Learning Theory, the “model problem” setup is taking again the “whatever the data” route, following the track opened by Kolmogorov in the 1930s.

1.2 VC dimension of a family of modeling functions, and VC Bounds for ERM

VC dimension is an integer that characterizes the separating power (complexity) of a family F of functions from $\mathbb{R}^n$ to $\mathbb{R}$. As we saw earlier, each f of F can be seen as a classifier.

We say that a set of L vectors of $\mathbb{R}^n$ $\{(\vec{x}_i), i = 1, \dots, L\}$ can be hashed by F if for every coloration of the L vectors in m “whites” and $L - m$ “blacks”, there exist an $f \in F$ that learns (classifies correctly) exactly the whites and the blacks, ie:

$$\begin{aligned} f(\vec{x}_i) &\geq 0 \text{ if and only if } \vec{x}_i \text{ is “white”,} \\ f(\vec{x}_i) &< 0 \text{ if and only if } \vec{x}_i \text{ is “black”} \end{aligned}$$

The set of functions F is said to have VC dimension h if:

- any set of h vectors of $\mathbb{R}^n$ can be shattered by F ,
- there exists a set of $h + 1$ vector that cannot be shattered by F .

For example, the set F of lines in $\mathbb{R}^2$ has a VC dimension of 3, the set of hyperplanes of $\mathbb{R}^n$ has a VC dimension of $n + 1$, but the special set of multilinear functions

$$f(x) = \langle \vec{w} | \vec{x} \rangle + b$$

with $\|\vec{w}\| < C$, has a VC dimension that depends on C and can take all values between 0 and n .

Important: the VC dimension of $F: \mathbb{R}^n \rightarrow \mathbb{R}$ is not equal to the number of parameters n . Vapnik gives examples in image processing where $n = 10^{21}$ and $h < 300$.

We are now ready for the first Vapnik inequality:

To build a model, we may imagine the f picked from family F as specified by a parameter $\vec{w}$:

$$f_{\vec{w}}(\vec{x}) = f(\vec{x}, \vec{w})$$

Let $\mathbf{z} = (\vec{x}, y)$, where $\vec{x} \in \mathbb{R}^n$ and $y \in \mathbb{R}$.

We shall describe the “quality”, ie the model (represented by $\vec{w}$) risk, by a cost function, a function $Q(\mathbf{z}, \vec{w})$. For instance, quality/cost could be evaluated by a least squares function such as

$$Q(\mathbf{z}, \vec{w}) = (f(\vec{x}, \vec{w}) - y)^2$$

or it could be a sum of misclassification costs on each of our L measures $\{\vec{x}_i | i = 1, \dots, L\}$, etc. Our Expected Risk Measure is then:

$$R(\vec{w}) = \int Q(\mathbf{z}, \vec{w}) dP(\mathbf{z})$$

where we integrate the risk over all possible values of $\mathbf{z} = (\vec{x}, y)$ under the unknown underlying probability distribution $P(\mathbf{z})$.

To build the model, we have only a sample of L events, $\{\mathbf{z}_i, i = 1, \dots, L\}$, and to find a good $\vec{w}$, we minimize the empirical risk, defined by:

$$E(\vec{w}) = \sum_{i=1}^L Q(\mathbf{z}_i, \vec{w})$$

Now we have a first Vapnik important inequality:

If $\vec{w}$ describes the functions f of a family F with VC dimension h , the following inequality holds “whatever the data”, with probability $1 - \eta$:

$$R(\vec{w}) < E(w) + \varphi\left(\frac{h}{L}, \eta\right)$$

where

$$\varphi^2\left(\frac{h}{L}, \eta\right) = \left[\frac{h}{L}(1 + \log(2)) - \frac{h}{L} \log\left(\frac{h}{L}\right) - \frac{1}{L} \log\left(\frac{\eta}{4}\right) \right]$$

As the formula shows, this inequality needs no specific knowledge of the probability distribution $P(\mathbf{z})$ or of the number of parameters n . Some consequences arise from this inequality:

- a model defined by a function $f(\vec{x}, \vec{w}) \in F$ will generalize in a *consistent* way (cost error on a test set converges to the cost error measured on the training set as its size L increases) if and only if F has a finite VC dimension;

- the speed of convergence is a function of h/L and does not depend of the number of attributes (this gives a new sense for problems such as 400 samples and 50000 attributes). $h/L \approx 1/20$ gives good precision;
- to build a model, one needs no longer to optimise the fit on the test set: the problem is now to optimise *both* the fit and the robustness, ie make the VC dimension h of F a controlled variable.

1.3 Structured Risk Minimization: a new approach to build models

To make h a controlled variable, Vapnik suggests a methodology named Structured Risk Minimization (SRM). He defines a set of family of functions:

$$F_1 \subset F_2 \subset \dots \subset F_p$$

with corresponding VC dimensions $h_1 < h_2 < \dots < h_p$.

When building a best model f on the training set with $f \in F_a$, one gets VC dimension of level h_a to control the robustness of such a model. The richer the family F_a , the better the fit on the training set, but the higher h_a , hence the poorer the robustness (generalization ability). For a specific a^* , both effects are counterbalanced, giving the SRM-based best model.

2 The generalized linear problem

2.1 Extended spaces

It has been repeatedly said here that the number of attributes of a problem for which a model is built has little influence, if the VC dimension of the family of functions F remains controlled (and low).

At no moment was it mentioned that *correlations* between attributes would play a role: this is a new –and confusing– paradigm, due to the basic change of the model setup strategy.

We follow here mainly the notations of [Cristianini and Taylor (2000)].

This has a very important consequence: it becomes possible to design a mapping Ψ from the original data space (original manifold), here above $\mathbb{R}^n$, to a new manifold $\wp \subseteq \mathbb{R}^p$, and this with a very open mind.

$$\vec{u} = \{(u_j)|j = 1, \dots, p\} = \phi(\vec{x}) = \{\phi_1(\vec{x}), \dots, \phi_p(\vec{x})\}$$

In the new manifold, the final hope is to build a good model f^* in $\vec{u}$ using SRM, $y = f^*(u)$, and the final model will be:

$$f(\vec{x}) = f^*(\phi(\vec{x})) = (\phi \circ f^*)(\vec{x})$$

A very simple example would be to map the original attributes of $\vec{x} = \{(x_i), i = 1, \dots, n\}$ to a polynomial of those attributes; eg, go from two variables $\vec{x} = \{x_1, x_2\}$ ($n = 2$) to second degree polynomial, $\vec{u}$ becoming the vector with components $\{x_1, x_2, x_1^2, x_2^2, x_1x_2\}$ and $p = 5$.

But one of the most common formulations of models, the one that we shall study from now in this paper intends to build models where in the new manifold $\wp$, f is linear in the new attributes, ie we want to work with our good old model:

$$f^*(u) = \langle \vec{w} | \vec{u} \rangle + b$$

with $\vec{u}$ and $\vec{w} \in \mathbb{R}^p$.

2.2 Kernels and RHKS (Reproducing Hilbert Kernel Spaces)

As we said, we now look specifically for models that can be reduced to the linear regression model:

$$y = f(\vec{x}) = \langle \vec{w} | \vec{x} \rangle + b$$

If we take for cost function least squares: $Q(\mathbf{z}, \vec{w}) = (f(\vec{x}, \vec{w}) - y)^2$, then the empirical least square cost will be:

$$E(\vec{w}, b) = \sum_{i=1}^L Q(\mathbf{z}_i, \vec{w}) = \sum_{i=1}^L (y_i - \langle \vec{w} | \vec{x}_i \rangle - b)^2$$

Minimizing $E(\vec{w}, b)$ leads to the well known normal equations in matrix form (we use the symbol $'$ for transposition):

$$\mathbf{X1}' \mathbf{X1} W = \mathbf{X1}' Y$$

Where Y is the column vector $y' = (y_1, \dots, y_L)'$, $\mathbf{X1}$ is the $L \times n + 1$ matrix, where line i of $\mathbf{X1}$ is the line $(x'_i, 1)$ with $n + 1$ elements, W is the column vector $(\vec{w}', b)'$.

Duality: a very important concept (well known to the statisticians that deal with principal component analysis and correspondence analysis, where *factors* in the attribute space correspond to *factors* in the event space) is the concept of duality: for all the problems that we examine here, there is a way to express the problem in the world of attributes (dimension n), and a *dual* way to express it in the world of events (dimension L). Such an approach allows to consider problems with a very large number of attributes, as long as L (hence the dual problem) remains low.

Ridge Regression: ridge regression is suggested when the $\mathbf{X1}' \mathbf{X1}$ is not of full rank. Instead of minimizing the least square loss function:

$$E(\vec{w}, b) = \sum (y_i - \langle \vec{w} | \vec{x}_i \rangle - b)^2$$

we minimize

$$E_{\text{ridge}} = E(\vec{w}, b) + \lambda \|\vec{w}\|^2$$

where $\|\vec{w}\|^2 = \langle \vec{w}, \vec{w} \rangle$ and λ penalizes “large” $\vec{w}$ as measured by $\|\vec{w}\|^2$.

It can be shown that this problem can be separated into two parts, one that will find b in terms of λ and the L means of x_i , and a *centered* problem, with $b = 0$ and each x_{ij} being replaced by $x_{ij} - \text{mean}\{x_{ij} | j = 1, \dots, n\}$.

Let us define $\mathbf{X}$ as a $L \times n$ matrix where the line i of $\mathbf{X}$ is equal to x'_i . We can make a *dual* correspondence between $\vec{w}$ and $\vec{\alpha}$ by the following equation:

$$\vec{w} = \mathbf{X}' \vec{\alpha}$$

The minimum solution of E_{ridge} in $\vec{w}$ and b can be obtained by minimizing the dual equation, where G (G for Gram) is the symmetric $L \times L$ matrix $\mathbf{X} \mathbf{X}'$, that we recognize as the scalar product matrix of the L centered vectors $\vec{x}_i$.

$$E_{\text{ridge}}^*(\vec{w}) = \lambda \vec{\alpha}' G \vec{\alpha} + \vec{\alpha}' G' G \vec{\alpha} - 2 y' G \vec{\alpha} + y' y$$

We obtain as solution for predictor $f(\vec{x})$:

$$f(\vec{x}) = y' (\lambda I_L + G)^{-1} q$$

where $q_i = \langle \vec{x} | \vec{x}_i \rangle$ and $G_{ij} = \langle \vec{x}_i | \vec{x}_j \rangle$ with $i, j \in (1, \dots, L)$.

What is important here is that the model solution depends only of scalar products in $\mathbb{R}^n$. We are now ready to introduce the notion of Kernel.

Let us imagine again a mapping from the original attribute space $\mathbb{R}^n$ to an extended space or feature space $\mathbb{R}^p$, in the form

$$\vec{x} \rightarrow \vec{u} = \phi(\vec{x}) = (\phi_1(\vec{x}), \dots, \phi_p(\vec{x}))$$

Sometimes one looks for *dimensionality reduction*, ie find a ϕ that will allow $p < n$. On other times, it may prove a better strategy to go for a much larger space, as long as the problem becomes linear, and can be efficiently solved through linear models and efficient VC control through SRM.

Kernels are made to help project the original problem into a feature space where the problem becomes linear:

$$f(\vec{x}) = \sum_{i=1}^p w_i \phi_i(\vec{x}) + b$$

We saw before that linear learning (learning with models such as linear regression or ridge regression, and, we shall see, Support Vector Machines) can be expressed in the dual space, the space of L events, with solutions that can be expressed through the scalar products, now in the feature space, of the form $\langle \phi(\vec{x}_i) | \phi(\vec{x}_j) \rangle$ or $\langle \phi(\vec{x}_i) | \phi(\vec{x}) \rangle$, these scalar products being evaluated in $\mathbb{R}^n \times \mathbb{R}^n$ (which is important to take into account if n is huge, for computational complexity and cache/data base timing reasons).

A *kernel* K will be a function from $\mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$, such that:

$$K(\vec{x}_1, \vec{x}_2) = \langle \phi(\vec{x}_1) | \phi(\vec{x}_2) \rangle$$

We can generalize using Hilbert space theory, which will lead us to the useful notion of Reproducing Kernel Hilbert Spaces (RKHS).

A very important theorem used here is *Mercer's theorem*. We suppose that $\vec{x}_1$ and $\vec{x}_2$ belong to a compact subspace Ω of $\mathbb{R}^n$, and that K is a continuous symmetric function such that the integral operator $T_k : L_2(\Omega) \rightarrow L_2(\Omega)$,

$$(T_k f)(\cdot) = \int_{\Omega} K(\cdot, \vec{x}) f(\vec{x}) d\vec{x}$$

is positive, that is

$$\int \int_{\Omega \times \Omega} K(\vec{x}_1, \vec{x}_2) f(\vec{x}_1) f(\vec{x}_2) d\vec{x}_1 d\vec{x}_2 > 0$$

is valid for all $f \neq 0$ with $f \in L_2$, ie: $\int_{\Omega} f^2(\vec{x}) d\vec{x} < \infty$

Then $K(\vec{x}_1, \vec{x}_2)$ has an expansion as a uniformly convergent series

$$K(\vec{x}_1, \vec{x}_2) = \sum_{k=1}^{\infty} \lambda_k \phi_k(\vec{x}_1) \phi_k(\vec{x}_2)$$

with $\phi_k(\vec{x})$ being the normalised ($\|\phi_k\|_{L_2} = 1$) eigen-functions of T_k , with positive eigen-value λ_k .

In such an environment, when can generalize the primal and dual representation of our model (primal representation, feature space size)

$$f(\vec{x}) = \sum_{k=1}^{\infty} \lambda_k \Psi_k \phi_k(\vec{x}) + b$$

and (dual representation, sample size)

$$f(\vec{x}) = \sum_{i=1}^L \alpha_i y_i K(\vec{x}, \vec{x}_i) + b$$

with $\Psi_k = \sum_{i=1}^L \alpha_i y_i \phi_k(\vec{x}_i) + b$. The important part here is that we can define kernel functions K , associated to specific mappings of the original attributes in a convenient feature space, and obtain the solution for best linear models by computing only in the dual space of size L , even if this mapping occurs from a feature space of infinite dimension.

All these types of equations will not surprise once again mathematicians with good practice of PCA world, where primal/dual factor representations are well known (see also Schölkopf *et al.* (1998)).

Most commonly used kernels are the following, and there are numerous ways in the literature to build new ones:

- Linear: $K(\vec{x}_1, \vec{x}_2) = \langle \vec{x}_1 | \vec{x}_2 \rangle$
- Polynomial: $K(\vec{x}_1, \vec{x}_2) = (\langle \vec{x}_1 | \vec{x}_2 \rangle + 1)^d$
- Radial Basis Functions: $K(\vec{x}_1, \vec{x}_2) = \exp(-\gamma \|\vec{x}_1 - \vec{x}_2\|^2)$
- Sigmoid: $K(\vec{x}_1, \vec{x}_2) = \tanh(\gamma \|\vec{x}_1 - \vec{x}_2\| + c)$

3 Direct approach and SVMs

Let us suppose that we have a set of L points from $\mathbb{R}^n$ with m “whites” and $L - m$ “blacks”, that are separable by a linear hyperplane classifier of the form

$$f(\vec{x}) = \text{sign}(\langle \vec{w} | \vec{x} \rangle + b)$$

We shall first suppose (hard margin SVM) that “whites” and “blacks” can be separated exactly by a hyperplane. There is an infinite number of possible hyperplanes, and if we take one of such, we can imagine parallel hyperplanes distant from the original one at $+\delta$ and $-\delta$ (forming a hyper-corridor or an kind of three layer mille-feuilles, Vapnik calls it a δ -margin separating hyperplane), such that the corridor separates still (for small enough δ) whites and blacks.

However, there is only one widest corridor, and the central layer, the median of the widest corridor separating whites and blacks is called the *hard margin support vector machine*.

Important Vapnik lemma : the VC dim of hyperplanes $y = \langle \vec{w} | \vec{x} \rangle + b$ with margin δ and description vectors $\{\vec{x}_i | i = 1, \dots, L\}$ such that $\|\vec{x}_i\| \leq R$ (R is the radius of the smallest sphere containing the $\vec{x}_i$), is bounded by

$$h \leq \min(R^2/\delta^2, L) + 1$$

Once again, this lemma shows that VC dimension of our classifier here does not depend of the number of parameters n , and that the widest the corridor (the larger the δ), the lower the corresponding h .

The SVM approach is straightforward: we shall look at a way to maximize δ given the L constraints that all L points remain outside the δ -margin separating hyperplanes. This leads to now well known Vapnik equations:

$$\text{Minimize } J = \frac{1}{2} \|\vec{w}\|^2$$

subject to $y_i(\langle \vec{w} | \vec{x} \rangle + b) \geq 1$, for $i = 1, \dots, L$ To solve such a problem we look for the saddle-point of the Lagrange functional:

$$L(\vec{w}, b, \vec{\alpha}) = \frac{1}{2} \|\vec{w}\|^2 - \sum_{i=1}^L (\alpha_i (\langle \vec{w} | \vec{x}_i \rangle + b) y_i - 1)$$

where the α_i are Lagrange multipliers. Differentiating L in b and $\vec{w}$, we find that the solution in the saddle point defined by $\vec{w}^*$, b^* and $\vec{\alpha}^*$ will satisfy: $\vec{w}^* = \sum_{i=1}^L y_i \alpha_i^* \vec{x}_i$ besides, according to Kuhn-Tucker equations at the saddle-point:

$$\alpha_i^* ((\langle \vec{x}_i | \vec{w}^* \rangle + b^*) y_i - 1) = 0, \text{ for } i = 1, \dots, L$$

which means that either $\alpha_i = 0$, or (**support vector**): $(\langle \vec{x}_i | \vec{w}^* \rangle + b^*) y_i = 1$. The sum for $\vec{w}^*$ is then carried only over the subset of support vectors $\text{SV} \subseteq [1, \dots, L]$.

We now get to the dual (in the sense of Kuhn-Tucker but also in the sense of duality that was described before) optimisation to solve, ie to maximise the dual functional:

$$W(\vec{\alpha}) = \sum_{i=1}^L \alpha_i - \frac{1}{2} \sum_{i,j=1}^L y_i y_j \alpha_i \alpha_j \langle \vec{x}_i | \vec{x}_j \rangle$$

under the constraints $\alpha_i \geq 0$ and $\sum y_i \alpha_i = 0$. This is a positive semi-definite quadratic program. Its solution is unique, with

$$\|\vec{w}^*\|^2 = \sum_{i,j \in \text{SV}} \alpha_i^* \alpha_j^* \langle \vec{x}_i | \vec{x}_j \rangle$$

$$f(\vec{x}) = \text{sign} \left(\sum_{i \in \text{SV}} y_i \alpha_i^* \langle \vec{x}_i | \vec{x} \rangle + b^* \right)$$

We note that here again, as in multilinear regression and ridge regression, the solution exists, is unique, and can be expressed in terms only of at most L inner products $\langle \vec{x}_i | \vec{x} \rangle$, whatever the number n of attributes. When data are non-linearly separable (ie whatever the hyperplane, there will be misclassifications), Vapnik suggests to solve (soft margin SVM):

$$\text{Minimize } J = \frac{1}{2} \|\vec{w}\|^2 + C \sum_i \xi_i$$

subject to: $y_i(\langle \vec{w} | \vec{x} \rangle + b) \geq 1 - \xi_i$, $\xi_i \geq 0$, for $i = 1, \dots, L$. C is a parameter that will control trade-off between margin and error, in a kind of SRM way, but this time we have three items to balance: fit, misclassification and robustness.

Olvi Mangasarian suggests a penalty for misclassifications in the form $C \sum \xi_i^2$ which allows a very fast computation. Soft Margin optimisation problem has a dual problem in the same way, though more complicated:

$$W(\vec{\alpha}) = \sum_{i=1}^L \alpha_i - \frac{1}{2} \sum_{i,j=1}^L y_i y_j \alpha_i \alpha_j \langle \vec{x}_i | \vec{x}_j \rangle$$

subject to $0 \leq \alpha_i \leq C$ and $\sum y_i \alpha_i = 0$. we still have a problem where in the dual form we only involve inner products of the type $\langle \vec{x}_i | \vec{x}_j \rangle$. The solution of such a problem is still unique, but now the influence of training samples is different.

To understand the role of C , we can say that $\vec{\alpha}$ expresses the “influence” of training sample i :

- “influence” is limited by C ;
- $0 \leq \alpha_i \leq C \Leftrightarrow$ SV with $\xi_i = 0$ (“good SV”);
- $\alpha_i = C \Leftrightarrow$ SV with $\xi_i > 0$ (“bad SV”, may be misclassified);
- $\alpha_i = 0$ training samples with no influence to the result.

We can generalize SVMs to extended spaces defined by kernels in the same way we did it with multilinear regression and ridge regression. We obtain as dual problem again and the solution is given by $f(\vec{x}) = \text{sign} \left(\sum_{i \in \text{SV}} y_i \alpha_i^* K(\vec{x}_i, \vec{x}) + b^* \right)$

Though this is not the discussion here, we can say that SVMs perform well when the data are with little noise, and a lot of attributes. But in our experience, good ridge regressions in proper extended spaces with the right kernels give results that cannot be differentiated in quality (fit, robustness) in a statistical sense.

4 Conclusion

In this paper, we gave:

1. a new inequality “whatever the data” gives a mathematical sense to model robustness notion (generalization capacity), and suggests SRM method to control generalization/fit trade-off for model optimization;
2. Kernel function theory in dual data space allows problem linearization in new manifolds;
3. A new equation controls specific linear models robustness, either the direct way (SVM) or the indirect way (ridge regression, Kernel PCA/PLS).

References

- Cristianini N. and Taylor J.S. (2000) *Support Vector Machines and other kernel-based learning methods*, Cambridge University Press.
- Hastie T., Tibshirani R. and Friedman J. (2001) *The Elements of Statistical Learning*, Springer.
- Mangasarian O. (2000) Generalized support vector machines, in: *Advances in Large Margin Classifiers*, Smola A., Bartlett P., Schölkopf B. and Schuurmans D., eds., MIT Press, 135–146.
- Schölkopf B., Burges C. and et al (1999) *Advances in Kernel Methods*, MIT Press.
- Schölkopf B., Mika S., Smol A., Rätsch G. and Müller K.R. (1998) Kernel pca pattern reconstruction via approximate pre-images, in: *Proceedings of the 8th International Conference on Artificial Neural Networks, Perspectives in Neural Computing*, Niklasson L., Bodén M. and Ziemke T., eds., Springer Verlag., 147–152.
- Smola A., Bartlett P. and et al. (2000) *Advances in Large Margin Classifiers*, MIT Press.
- Vapnik V. (1998) *Statistical Learning Theory*, Wiley.
- Vapnik V. (1999) *The Nature of Statistical Learning Theory*, Springer-Verlag.