TEKNISKA HÖGSKOLAN
LINKÖPINGS UNIVERSITET

Forskarskola i medicinsk bioinformatik Course for CMI PhD programme in Medical Bioinformatics: Support Vector Machines Part I Lecture Notes 5: Regularization for SVM and Representer Theorem

8-9th of September, 2005
LiTH/Linköping University

Timo Koski
Department of mathematics
LiTH

This fifth set of lecture notes presents what is known as regularization theory as applied to SVM. Schölkopf and Smola (2002, p. 87) state that '..this may not be an easy digest for some of our readers..'. The quote is not inserted here in order to intimidate, but to point out that we are approaching a piece of SVM theory, where a relatively heavy mathematical machinery forms prerequisites, if a complete mastery of the subject is desired.

This lecture attempts only to summarize some of the most essential points about regularization and kernels needed to state the representer theorem, so that its contents can be appreciated.

OUTLINE OF TOPICS:

- regularization: ill-posedness
- inner product spaces of infinite dimensions
- new property of the kernel function, the existence of a reproducing kernel Hilbert space
- representer theorem

Literature

The lecture is mainly based on

- N. Christiani & J. Shawe-Taylor (2000): *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge.

- R. Herbrich (2002): *Learning Kernel Classifiers. Theory and Algorithms*. The MIT Press. Cambridge, Massachusetts, London, England.

- B. Schölkopf & A.J. Smola (2002): *Learning with Kernels*, The MIT Press, Cambridge, Massachusetts.
  (chapter 4.1–4.2)

Literature

Regularization for function approximation and RBF, e.t.c., is succinctly accounted for in

- T. Poggio & F. Girosi (1990): Networks for Approximation and Learning. *Proceedings of the IEEE*, 78, pp. 1481−1497.

Background

So far we have mainly discussed training algorithms like the perceptron algorithm, where the empirical risk R_{err} is zero after the algorithm has come to a halt.

Now we discuss learning algorithms, where we minimize

$$R_{\text{err}}(h) = \frac{1}{l} \sum_{i=1}^{l} \mid h(x_i) - y_i \mid$$

over the hypothesis space $\mathcal{H}$ for a given training set $\mathcal{S}$. The *loss function*, the so called $0-1$ loss function used in the preceding, was

$$\mid h(x) - y \mid = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}$$

It is, in fact, computationally difficult* to handle this loss function for minimization of $R_{\text{err}}(h)$. We need to modify the loss function.

*NP-hard

Hinge Loss

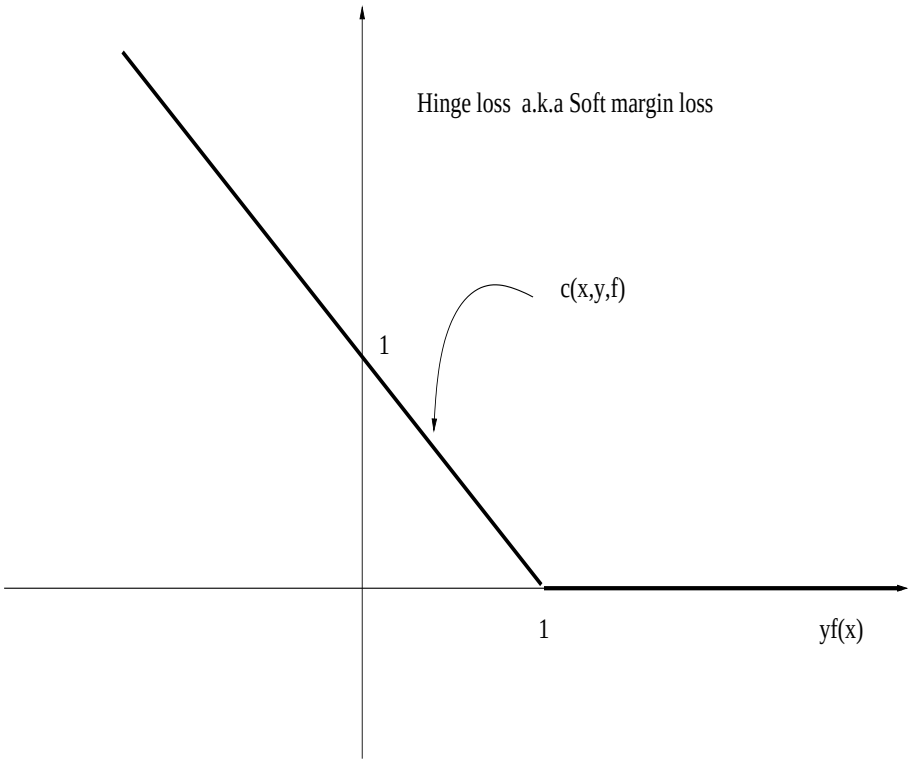
One approximation of $0-1$ loss function is the *hinge loss*[*] or *soft-margin loss* defined by

$$c(\mathbf{x}, y, f) = \begin{cases} 0 & \text{if } yf(\mathbf{x}) > 1 \\ 1 - yf(\mathbf{x}) & \text{otherwise} \end{cases}$$

Here $y = \pm 1$ and $f(\mathbf{x}) = \ll \mathbf{w}, \mathbf{x} \gg + b$. Note that $yf(\mathbf{x}) < 0$, if there is a classification error. This is a continuous function in $yf(\mathbf{x})$.

[*]hinge= gångjärn in Swedish

Hinge Loss



Ill-posedness

Minimization of empirical error

$$\tilde{R}_{\text{err}}(f) = \frac{1}{l} \sum_{i=1}^{l} c(\mathbf{x}_i, y_i, f)$$

may have numerical problems. In particular, there is no unique minimum, if we try to find a set of weights $\mathbf{w}$ in an SVM and the bias b for minimization of empirical risk.

Hence we see that the solution is unstable or *ill-posed*: small changes in the training data $(\mathbf{x}_i, y_i)_{i=1}^{l}$ may lead to large changes in $(\mathbf{w}, b)$.

Regularizing

Regularization as used in several ill-posed problems of statistics and scientific computing would suggest for SVM the minimization of

$$R_{\text{reg}}(f) = \tilde{R}_{\text{err}}(f) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

where $\lambda > 0$ is the so-called regularization parameter. Now it will be shown that there is a way of re-designing this using the properties of the feature space, which then gives us the so-called *representer theorem* (to be stated later). This theorem provides an explicit representation for the minimizing f.

We need some prerequisites.

Prerequisites (D): Inner Product Spaces with Infinite Dimension

We must now consider infinite dimensional inner product spaces. These are also known as *(pre)-Hilbert spaces*. We give an example, this is an inner product space called l_2 space.

$\mathbf{X} = l_2$ is the vector space of vectors $\mathbf{x}$ of the form

$$\mathbf{x} = (x_1, x_2, \ldots), \quad x_i \in R, i = 1, 2, \ldots,$$

such that

$$\sum_{i=1}^{\infty} x_i^2 < \infty.$$

Here we take, as is natural, the vector space operations as

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, \ldots)$$

It is easily seen that $\mathbf{x} + \mathbf{y} \in l_2$.

$$\lambda\mathbf{x} = (\lambda x_1, \lambda x_2, \ldots),$$

Prerequisites (D): Inner Product Spaces with Infinite Dimension

The inner product in l_2 is defined by

$$\ll \mathbf{x}, \mathbf{y} \gg_{l_2} \stackrel{def}{=} \sum_{i=1}^{\infty} x_i y_i$$

This inner product can be seen to satisfy the axiomatic properties of an inner product as stated in the prerequisites (A) for lecture 1. Hence

$$\|\mathbf{x}\|_{l_2} = \sqrt{\ll \mathbf{x}, \mathbf{x} \gg} = \sqrt{\sum_{i=1}^{\infty} x_i^2}$$

is a norm on l_2.

Prerequisites (D): Weighted l_2, $l_2(\mu)$

An inner product space called a weighted l_2 space designated by $l_2(\mu)$ is the space of vectors $\mathbf{x}$ of the form

$$\mathbf{x} = (x_1, x_2, \ldots), \quad x_i \in R, i = 1, 2, \ldots,$$

such that

$$\sum_{i=1}^{\infty} \mu_i x_i^2 < \infty,$$

where

$$\mu_i > 0 \quad \text{for all } i, \quad \text{and} \quad \sum_{i=1}^{\infty} \mu_i < \infty$$

The inner product in $l_2(\mu)$ is defined by

$$\ll \mathbf{x}, \mathbf{y} \gg_{l_2(\mu)} \stackrel{def}{=} \sum_{i=1}^{\infty} \mu_i x_i y_i$$

and consequently

$$\|\mathbf{x}\|_{l_2(\mu)} = \sqrt{\ll \mathbf{x}, \mathbf{x} \gg}_{l_2(\mu)} = \sqrt{\sum_{i=1}^{\infty} \mu_i^2 x_i^2}$$

is a norm on $l_2(\mu)$.

Inner Product Spaces with Infinite Dimension *

In mathematical terms we have introduced here two *pre-Hilbert spaces* , l_2 and $l_2(\mu)$, as discussed in greater length in

- D.G. Luenberger: *Optimization by Vector Space Methods.* Wiley Professional Paperbacks, John Wiley annd Sons, 1998, chapter 3

Statements about l_2 are found on pp. 29−31. A pre-Hilbert space becomes a Hilbert space, as soon as it is completed with limits of all convergent sequences.

*These references are included as information for those interested, and are not part of the course syllabus

Feature Space

Let X be an input space, and let ψ be a feature map from X to the feature space $\mathcal{F} = l_2(\mu)$

$$\psi : X \mapsto l_2(\mu).$$

Consequently

$$\psi(\mathbf{x}) = (\psi_1(\mathbf{x}), \psi_2(\mathbf{x}), \ldots)$$

The kernel is by Mercer's theorem given by

$$K(\mathbf{x}, \mathbf{y}) = \ll \psi(\mathbf{x}), \psi(\mathbf{y}) \gg_{l_2(\mu)}$$

Thus

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \mu_i \psi_i(\mathbf{x}) \psi_i(\mathbf{y})$$

by definition of the inner product $\ll \cdot, \cdot \gg_{l_2(\mu)}$.

Next we modify the construction presented in Christiani & Shawe-Taylor (2000) as needed for the representer theorem.

Representer theorem: a construction (1)

Let us consider a set of functions from X to real numbers

$$\mathbf{H}_a = \{\phi : X \mapsto R \mid \phi(\mathbf{x}) = \sum_{i=1}^{\infty} a_i \psi_i(\mathbf{x})\}$$

where $a = (a_1, a_2, \ldots)$ is a sequence such that

$$\sum_{i=1}^{\infty} \frac{a_i^2}{\mu_i} < \infty \quad (*)$$

Take, e.g. $a_i = \mu_i k_i$, where $(k_1, k_2, \ldots) \in l_2(\mu)$. We can make $\mathbf{H}_a$ to a vector space by defining

$$(\phi_1 + \phi_2)(\mathbf{x}) = \phi_1(\mathbf{x}) + \phi_2(\mathbf{x})$$

and

$$(\lambda\phi_1)(\mathbf{x}) = \lambda\phi_1(\mathbf{x}).$$

Then we can introduce an inner product in $\mathbf{H}_a$ as

$$\langle \phi_1, \phi_2 \rangle_{\mathbf{H}_a} \stackrel{def}{=} \sum_{i=1}^{\infty} \frac{a_i b_i}{\mu_i}$$

if $\phi_1(\mathbf{x}) = \sum_{i=1}^{\infty} a_i \psi_i(\mathbf{x})$ and $\phi_2(\mathbf{x}) = \sum_{i=1}^{\infty} b_i \psi_i(\mathbf{x})$.

Then let us recall the kernel:

$$K(\mathbf{z}, \mathbf{x}) = \sum_{i=1}^{\infty} \mu_i \psi_i(\mathbf{z}) \psi_i(\mathbf{x})$$

The notation

$$K(\mathbf{z}, \cdot)$$

means that we have fixed $\mathbf{z}$ and consider $K(\mathbf{z}, \cdot)$ as a function of the second argument in X. Now we show that

$$K(\mathbf{z}, \cdot) \in \mathbf{H}_a.$$

Representer theorem: a construction (2)

In order to see that

$$K(\mathbf{z},\cdot)\in\mathbf{H}_a,$$

we msut check $(*)$ above. This foloows, since

$$K(\mathbf{z},\cdot)=\sum_{i=1}^{\infty}\underbrace{\mu_i\psi_i(\mathbf{z})}_{:=a_i(\mathbf{z})}\psi_i(\cdot)=$$

$$=\sum_{i=1}^{\infty}a_i(\mathbf{z})\,\psi_i(\cdot)$$

and

$$\sum_{i=1}^{\infty}\frac{(a_i(\mathbf{z}))^2}{\mu_i}=\sum_{i=1}^{\infty}\mu_i\psi_i^2(\mathbf{z})<\infty$$

since $\psi(\mathbf{z})\in l_2(\mu)$. Then we can compute

$$\langle K(\mathbf{z},\cdot),K(\mathbf{x},\cdot)\rangle_{\mathbf{H}_a}=\sum_{i=1}^{\infty}\frac{a_i(\mathbf{z})\,a_i(\mathbf{x})}{\mu_i}$$

$$=\sum_{i=1}^{\infty}\frac{\mu_i\psi_i(\mathbf{z})\,\mu_i\psi_i(\mathbf{x})}{\mu_i}=\sum_{i=1}^{\infty}\mu_i\psi_i(\mathbf{z})\,\psi_i(\mathbf{x})=K(\mathbf{z},\mathbf{x})$$

Representer theorem: a construction (3)

If $\phi \in \mathbf{H}_a$, then

$$\langle K(\mathbf{z}, \cdot), \phi \rangle_{\mathbf{H}_a} = \sum_{i=1}^{\infty} \frac{a_i(\mathbf{z})\, a_i}{\mu_i}$$

$$= \sum_{i=1}^{\infty} \frac{\mu_i \psi_i(\mathbf{z})\, \mu_i a_i}{\mu_i} = \sum_{i=1}^{\infty} a_i \psi_i(\mathbf{z}) = \phi(\mathbf{z})$$

i.e.,

$$\langle K(\mathbf{z}, \cdot), \phi \rangle_{\mathbf{H}_a} = \phi(\mathbf{z}).$$

We say that ϕ is **reproduced** by the inner product in $\mathbf{H}_a$.

We have now outlined a theorem that there is for any kernel $K(\mathbf{x}, \mathbf{y})$ on $X \times X$ an inner product space, called RKHS*, here found as $\mathbf{H}_a$, such that

1. $K(\mathbf{x}, \cdot) \in \mathbf{H}_a$,
2. the inner product $\langle \cdot, \cdot \rangle_{\mathbf{H}_a}$ is reproducing.

It holds also by the above that

$$\langle K(\mathbf{x}, \cdot), K(\mathbf{y}, \cdot) \rangle_{\mathbf{H}_a} = K(\mathbf{y}, \mathbf{x}) = K(\mathbf{x}, \mathbf{y})$$

*reproducing kernel Hilbert space, c.f. Luenberger (1998), p. 73.

Representer theorem

But, now we see that the e.g. dual function of soft margin classifier

$$f_{SVM}(\mathbf{x}) = \sum_{i=1}^{l} y_i \alpha_i^* K(\mathbf{x}_i, \mathbf{x})$$

belongs to $\mathbf{H}_a$. it can also be shown that $\mathbf{H}_a$ is the closure w.r.t. $\langle \cdot, \cdot \rangle_{\mathbf{H}_a}$ of sums of the form

$$\sum_{i=1}^{l} a_i K(\mathbf{x}_i, \mathbf{x})$$

Representer theorem for L1SVM

Each minimizer of the regularized risk

$$R_{\text{reg}}(f) = \tilde{R}_{\text{err}}(f) + \frac{1}{2}\|f\|^2_{\mathbf{H}_a}$$

can be written in the form

$$f^*(\mathbf{x}) = \sum_{i=1}^{l} a_i^* K(\mathbf{x}_i, \mathbf{x})$$

It holds that

$$\|f\|^2_{\mathbf{H}_a} = \sum_{i=1}^{l}\sum_{j=1}^{l} a_i^* a_j^* K(\mathbf{x}_i, \mathbf{x}_j)$$

The proof is based on the properties of orthogonal projection on linear subspaces in Hilbert spaces, and is found, e.g., in Schölkopf and Smola (2002), pp. 90−91. The regularized SVM is known as L1SVM.

Representer theorem

The formulation of the representer theorem may seem a bit abstract and artificial, especially when wondering about the construction of $\mathbf{H}_a$, and $\langle \cdot, \cdot \rangle_{\mathbf{H}_a}$. It turns out, however, that in many cases the space and the inner product can be found in an explicit form, i.e., without infinite sums. Examples of this, and examples of the representer theorem are found in

- G. Wahba (1998): Support Vector Machines, Reproducing Kernel Hilbert Spaces and the Randomized GACV. *Department of Statistics, University of Wisconsin, Technical Report NO. 984rr*.
  `http://www.stat.wisc.edu/~wahbda/ftp1`

The webpage of Grace Wahba (URL above) contains several technical reports about RKHS and SVM.

Regularization: Consistency ?

We have spent a whole deal of time providing conditions (VC-dimension) for consistency of the empirical risk error minimizer. Is the regularized estimator consistent? The answer can be yes, under some precise and rigorous conditions on kernels, see

- I. Steinwart (2005): Consistency of Support Vector Machines and other Regularized Kernel Classifiers. *IEEE Transactions on Information Theory*, 51 (1), pp. 128−142.