Regularisation and Support Vector Machines

Generalisation Theory

Constanza Uribe Óscar Alonso Juan Carlos Galeano

Department of Computer Science and Engineering
National University of Colombia

Machine Learning
2007-I

Outline

Introduction

- Learning from data (finding patterns in data) without controlling the generalisation error makes no sense. If our goal is to predict.
- Hence the learning machine looks for a model that does not fail in the entire problem domain.
- But we do not now the whole set of domain's instances, we only know some examples from which we have to extract the significant patterns.
- Then, the best we can do is to fix an acceptable level of error and then to bound the probability that the machine learner makes such error.

Introduction (cont.)

- Which factors have to be controlled to guarantee good generalisation.
- VC theory is the most appropriate to describe SVMs.
- VC theory place reliable bounds on the generalisation of linear classifiers and hence indicate how to control the complexity of linear functions in kernel spaces.

Probably Approximately Correct Learning

- Rates of uniform convergence, frequentist inference (statistics)
- PAC (computer science)
- Training and test data are generated i.i.d. according to an unknown but fixed distribution $\mathcal{D}$.
- Distribution over input/output pairings $(x, y) \in X \times \{-1, 1\}$

Probably Approximately Correct Learning(cont.)

- Natural measure of error is the probability that a randomly generated example is misclassified

$$err_{\mathcal{D}}(h) = \mathcal{D}\{(x, y) : h(x) \neq y\}$$

  where h is a classification function
- Such measure is known as *risk functional*
- Aim: to assert bounds on this error in terms of several quantities: number of training examples is perhaps the most crucial of those quantities
- PAC results presented as bounds on the number of examples required to obtain a particular level of error, a.k.a. *sample complexity of the learning problem*

Probably Approximately Correct Learning (cont.)

- Fixed inference rule for selecting a hypothesis h_S from the class H of classification rules at the learner's disposal based on

$$S = \{(x_1, y_2), \ldots, (x_\ell, y_\ell)\}$$

chosen i.i.d. according to $\mathcal{D}$

- $err_{\mathcal{D}}(h_S)$ as a random variable depending on the random selection of the training set.
- Aim: to bound the expected generalisation error. Expectation is taken over the random selection of training sets of a particular size ℓ

Probably Approximately Correct Learning (cont.)

- PAC bounds the tail δ of the distribution of $err_{\mathcal{D}}(h_S)$. So, the pac bound has the form $\epsilon = \epsilon(\ell, H, \delta)$ and asserts that with probability at least $1 - \delta$ over randomly generated training sets S of size ℓ the generalisation error of the selected hypothesis h_S will be bounded by

$$err_{\mathcal{D}}(h_S) \leq \epsilon(\ell, H, \delta)$$

  i.e. it is probably approximately correct (pac).
- It is equivalent to say that the probability that the training set give rise to a hypothesis with large error is small

$$\mathcal{D}^{\ell}\{S : err_{\mathcal{D}}(h_S) > \epsilon(\ell, H, \delta)\} < \delta$$

- This is a flavour of statistical test, the difference is that our bound should be *distribution free.*

Vapnik Chervonenkis Theory

- For a finite set of hypothesis it is not hard to obtain a bound in the form of inequality

$$\mathcal{D}^{\ell}\left\{S : err_{\mathcal{D}}\left(h_S\right) > \epsilon\left(\ell, H, \delta\right)\right\} < \delta$$

- Inference rule: to select any hypothesis h that is consistent with the training examples in S.
- Probability that all ℓ of the independent examples are consistent with h for which $err_{\mathcal{D}}\left(h\right) > \epsilon$ is bounded by

$$\mathcal{D}^{\ell}\left\{S : h \text{ consistent and } err_{\mathcal{D}}\left(h\right) > \epsilon\right\} \leq \left(1 - \epsilon\right)^{\ell} \leq \exp\left(-\epsilon\ell\right)$$

Vapnik Chervonenkis Theory (cont.)

- Assuming that all $|H|$ of the hypothesis have large error, the probability that one of them is consistent with S is at most

$$|H| \exp(-\epsilon\ell)$$

- This bounds the probability that a consistent hypothesis h_S has error greater than ϵ

$$\mathcal{D}^\ell \{S : h_S \text{ consistent and } err_\mathcal{D}(h) > \epsilon\} < |H| \exp(-\epsilon\ell)$$

Vapnik Chervonenkis Theory (cont.)

- In order to ensure the right hand side is less that δ, we set

$$\epsilon = \epsilon\left(\ell, H, \delta\right) = \frac{1}{\ell} \ln \frac{|H|}{\delta}$$

- This shows how the complexity (number of choices) of the function class H has a direct effect on the error bound.
- Major contribution of VC's theory was to extend such an analysis to infinite sets of hypothesis.

Vapnik Chervonenkis Theory (cont.)

- The key to bounding over and infinite set of functions is to bound the pac probability as

$$\mathcal{D}^{\ell}\left\{S : \exists h \in H : err_S\left(h\right) = 0, err_{\mathcal{D}}\left(h\right) > \epsilon\right\}$$
$$\leq 2\mathcal{D}^{2\ell}\left\{S\hat{S} : \exists h \in H \ : \ err_S\left(h\right) = 0,\ err_{\hat{S}}\left(h\right) > \epsilon\ell/2\right\}$$

  which follows from an application of Chernoff bounds provided $\ell > 2/\epsilon$

- Quantity on the right hand side is bounded by fixing the 2ℓ sample and counting different orders in which the points might have been chosen while still keeping all the errors in the second sample

Vapnik Chervonenkis Theory (cont.)

- Since each permutation is equally likely, the fraction of those permutations that satisfy the property is an upper bound of its probability.
- By only considering permutations that swap corresponding points from the first and second sample, we can bound the fraction by $2^{-\epsilon\ell/2}$ independently of the particular set of 2ℓ sample points.
- Considering errors over a finite set of 2ℓ sample points is that the hypothesis space becomes finite, since there cannot be more than $2^{2\ell}$ classification functions on 2ℓ points.

Vapnik Chervonenkis Theory (cont.)

- To obtain and union bound on the overall probability of the right hand side, all that is required is a bound on the size of the hypothesis space when restricted to 2ℓ points, a.k.a. the *growth function*

$$B_H(\ell) = \max_{(x_1,\ldots,x_\ell)\in X} \left|\left\{\left(h(x_1), h(x_2), \ldots, h(x_\ell)\right) : h \in H\right\}\right|$$

this quantity cannot exceed 2^ℓ since the sets over which the maximum is sought are all of the set of binary sequences of length ℓ

Vapnik Chervonenkis Theory (cont.)

- A set of points $\{x_1, \ldots .x_\ell\}$ for which the set $\{(h(x_1), h(x_2), \ldots, h(x_\ell)) : h \in H\} = \{-1, 1\}^\ell$ is said to be *shattered* by H.
- If there are sets of any size which can be shattered then the growth function is equal to 2^ℓ for all ℓ.

Vapnik Chervonenkis Theory (cont.)

- Final ingredient in the VC theory is the analysis of the case when there is a finite d which is the largest size of shattered set. In this case, the growth function can be bounded as follows for $\ell \geq d$

  $$B_H(\ell) \leq \left(\frac{e\ell}{d}\right)^d$$

  giving polynomial growth with exponent d (the VC dimension).

Vapnik Chervonenkis Theory (cont.)

- Putting this bound in the bound obtained for infinite set of functions we get

$$\mathcal{D}^{\ell}\left\{S : \exists \in H : err_S\left(h\right) = 0, err_{\mathcal{D}}\left(h\right) > \epsilon\right\} \leq 2\left(\frac{2e\ell}{d}\right)^{d} 2^{-\epsilon\ell/2}$$

resulting in a pac bound for any consistent hypothesis h of

$$err_{\mathcal{D}}\left(h\right) \leq \epsilon\left(\ell, H, \delta\right) = \frac{2}{\ell}\left(d \log \frac{2e\ell}{d} + log\frac{2}{\delta}\right)$$

provided $d \leq \ell$ and $\ell > 2/\epsilon$

Vapnik Chervonenkis Theory (cont.)

- Remark: For infinite set of hypotheses the problem of overfitting is avoidable and the measure of complexity that should be used is the VC dimension.
- Remark: The size of the training set required to ensure good generalisation scales linearly with this quantity in the case of consistent hypothesis.
- Remark: VC theory provides a distribution free bound on generalisation of a consistent hypothesis.

Vapnik Chervonenkis Theory (cont.)

- Remark: for a hypothesis class with high VC dimension there exist input probability distributions which will force the learner to require a large training set to obtain a good generalisation (VC dimension charaterises learnability in the pac sense)
- Remark: It is possible that a class with high VC dimension is learnable if the distribution is benign. An essential fact for the performance of SVMs, which are designed to take advantage of such benign distributions

Vapnik Chervonenkis Theory (cont.)

- To apply the theory to linear machines we have to calculate the VC dimension of a linear function class L in $\Re^n$ in terms of n, that is determine what is the largest number d of examples that can be shattered by L
- Proposition:
  - Given any set S of $n+1$ training examples in general position there exist a function in L that consistently classifies S, whatever the labeling of the training points in S
  - For any set of $\ell > n+1$ inputs, there is at least one classification that cannot be realised by any function in L.

Vapnik Chervonenkis Theory (cont.)

- So far, the theory only applies when the hypothesis is consistent with the training data.
- The theory can be adapted to allow for a number of errors in the training set by counting the permutations which have no more errors on the left hand size

Vapnik Chervonenkis Theory (cont.)

- The resulting bound on generalisation error is given by

$$err_{\mathcal{D}}(h) \leq \epsilon(\ell, H, \delta) = \frac{2k}{\ell} + \frac{4}{\ell}\left(d \log \frac{2e\ell}{d} + log\frac{2}{\delta}\right)$$

where k is the number of errors on the classification of the training set.

Vapnik Chervonenkis Theory (cont.)

- A learning algorithm should seek to minimise the number of training errors since everything else has been fixed by the choice of H (empirical risk minimisation)
- This bound can be used to chose the hypothesis h_i for which the bound is minimal that is, the reduction in the number of errors (first term) outweighs the increase in capacity (second term)
- This induction strategy is known as structural risk minimisation.

Margin-Based Bounds on Generalisation

- Consider using a class $\mathcal{F}$ of real-valued functions on an input space X for classification by thresholding at 0.
- The *margin of an example* $(x_i, y_i) \in X \times \{-1, 1\}$ *with respect to a function* $f \in \mathcal{F}$ is the quantity

$$\gamma_i = y_i f(x_i)$$

- $\gamma_i > 0$ implies correct classification

Margin-Based Bounds on Generalisation

- the margin $m_S(f)$ of f is the minimum of the margin distribution
- $m_S > 0$ if f correctly classifies S
- The *margin of a training set* S with respect with the class $\mathcal{F}$ is the maximum margin over all $f \in \mathcal{F}$
- If we are considering linear function class we assume that the margins are geometric (weight vector has unit norm)

Maximal Margin Bounds

- A large γ can reduce the size of the function space.
- Generalisation performance can be approximated by a function whose output is within $\gamma/2$ on the points of double sample.
- A $\gamma -$ *cover* of $\mathcal{F}$ with respect to a sequence of inputs $S = \{x_1, \dots, x_\ell\}$ is a finite set of functions B such that for all $f \in \mathcal{F}$ there exists $g \in B$ such that

$$\max_{1 \leq i \leq \ell} \left(\left| f\left(x_i\right) - g\left(x_i\right) \right| \right) < \gamma$$

- $\mathcal{N}\left(\mathcal{F}, S, \gamma\right)$ is the smallest cover
- $\mathcal{N}\left(\mathcal{F}, \ell, \gamma\right) = \max\limits_{S \in X^l} \mathcal{N}\left(\mathcal{F}, S, \gamma\right)$ are the covering numbers

Maximal Margin Bounds

- The theorem cam be reformulated using the covering numbers

$$\mathcal{D}^{\ell}\left\{S : \exists f \in F \ : \ err_S(f) = 0,\ m_S(f) \geq \gamma,\ err_{\mathcal{D}}(f) > \epsilon\right\}$$
$$\leq 2\mathcal{D}^{2\ell}\left\{S\hat{S} : \exists f \in F \ : \ err_S(f) = 0,\ m_S(f) \geq \gamma,\ err_{\hat{S}}(f) > \epsilon\ell/2\right\}$$

- By a similar analysis, the right hand side of the inequality can be bounded by

$$\leq 2\left|B\right| 2^{-\epsilon\ell/2} \leq 2\mathcal{N}\left(\mathcal{F}, 2\ell, \gamma/2\right) 2^{-\epsilon\ell/2}$$

Maximal Margin Bounds

- The, we get the result

$$err_{\mathcal{D}}(f) \leq \epsilon(\ell, F, \delta, \gamma) = \frac{2}{\ell}\left(\log \mathcal{N}(\mathcal{F}, 2\ell, \gamma/2) + log\frac{2}{\delta}\right)$$

  provided $\ell > 2/\epsilon$

- The bound on $\log \mathcal{N}(\mathcal{F}, \ell, \gamma)$ represents a generalisation of the bound on the growth function required for the VC theory.
- The corresponding quantity we shall use to bound the covering numbers will be a real-valued generalisation of the VC dimension known as the fat-shattering dimension.

Maximal Margin Bounds (cont.)

- A set of points $x_1, \ldots, x_\ell$ is $\gamma -$ *shattered* by $\mathcal{F}$ if there exists real numbers r_i such that for every binary classification $b \in \{-1, 1\}^\ell$ there exists $f_b \in \mathcal{F}$, such that

$$f_b(x_i) = \begin{cases} \geq r_i + \gamma, & b_i = 1 \\ < r_i - \gamma, & b_i = -1 \end{cases}$$

- The fat-shattering dimension at scale γ is the size of the largest $\gamma -$ *shattered* subset of X (a.k.a. scale-sensitive VC dimension)
- Clearly, the larger the value of γ, the smaller the size of set that can be shattered since the restrictions placed on the functions that can be used become stricter.

Margin Percentile Bounds

- It includes the case when a hypothesis is not fully consistent with the training data.

$$err_{\mathcal{D}}(f) \leq \frac{k}{\ell} + \sqrt{\frac{c}{\ell}\left(\frac{R^2}{M_{s,k}(f)^2} log^2 \ell + log\frac{1}{\delta}\right)}$$

, where k/ℓ is the number of allowed errors, and $M_{s,k}(f)$ is the k/ℓ percentile of $M_s(f)$.

- It suggest that we can obtain the best generalisation performance by minimising the number of margin error, where we define a training point to be a $\gamma - margin$ error if it has margin less than γ.

Soft Margin Bounds

- Consider using a class $\mathcal{F}$ of real-valued functions on an input space $\mathcal{X}$ for classification by thresholding at 0. We define the margin slack variable of an example $(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}$ with respect to a function $f \in \mathcal{F}$ and target margin γ to be the quantity

$$\epsilon\left(\left(x_i, y_i\right), f, \gamma\right) = \epsilon_i = max\left(0, \gamma - y_i f\left(x_i\right)\right)$$

Soft Margin Bounds (cont)

- Consider thresholding real-valued lineal functions L with unit weight vectors on an inner product space $\mathcal{X}$ and fix $\gamma \in R^+$. There is a constant c such that for any probability distribution D in $\mathcal{X}$ x $\{-1, 1\}$ with support in a ball of radious R around the origin, with probability $1 - \delta$ over ℓ random examples S , any hypothesis $f \in \mathcal{L}$ has error no more than

$$err_{\mathcal{D}}(f) \leq \frac{c}{\ell}\left(\frac{R^2 + \|\epsilon\|_2^2}{\gamma^2} \log \frac{2e\ell}{d} + log\frac{1}{\delta}\right)$$

where ϵ is the slack vector with respect to f and γ