Basic Statistical Learning Theory: Overnight problems

John Shawe-Taylor
School of Electronics and Computer Science
University of Southampton
jst@ecs.soton.ac.uk

September, 2004

Berder Island Summer School, September 2004

Covering Numbers

$\mathcal{F}$ a class of real functions defined on X and $\|\cdot\|_d$ a norm on $\mathcal{F}$, then

$$\mathcal{N}(\gamma, \mathcal{F}, \|\cdot\|_d)$$

is the smallest size set U_γ such that
for any $f \in \mathcal{F}$ there is a $u \in U_\gamma$ such that $\|f - u\|_d < \gamma$.

Covering Numbers cont.

For generalization bounds we need the γ-*growth function*,

$$\mathcal{N}^m(\gamma, \mathcal{F}) := \sup_{\mathbf{x} \in X^m} \mathcal{N}(\gamma, \mathcal{F}, \ell_\infty^{\mathbf{x}}).$$

where $\ell_\infty^{\mathbf{x}}$ gives the distance between two functions as the maximum difference between their outputs on the sample.

Covering numbers for linear functions

- For the case of linear functions there is a more direct route to bounding the covering numbers.

- We convert the $\gamma/2$ approximation on the sample problem into a classification problem, which is solvable with a margin of $\gamma/2$.

- It follows that if we use the perceptron algorithm to find a classifier, we will find a function satisfying the $\gamma/2$ approximation with just $8R^2/\gamma^2$ updates.

Covering numbers for linear functions

- This gives a sparse dual representation of the function. The covering is chosen as the set of functions with small sparse dual representations.

- Gives a bound on the size of the covering numbers of the form

$$\log_2 \mathcal{N}^{2m}(\gamma/2, \mathcal{F}) \leq k \log_2 \frac{e(2m+k-1)}{k} \text{ where } k = \frac{8R^2}{\gamma^2}.$$

Generalization of SVMs

For distribution with support in ball of radius R, (eg Gaussian Kernels $R = 1$) and margin γ, have bound:

$$\epsilon(m, \mathcal{L}, \delta, \gamma) = \frac{2}{m}\left(k \log_2 \frac{e(2m + k - 1)}{k} + \log_2 \frac{m}{\delta}\right)$$

where $k = \frac{8R^2}{\gamma^2}$.

Exercise: reconstruct the bound on the covering numbers!

$$\log_2 \mathcal{N}^m(\gamma, \mathcal{F}) \leq k \log_2 \frac{e(m+k-1)}{k}$$

$$\text{where } k = \frac{2R^2}{\gamma^2}.$$

- Overnight,
- can work in groups (all members get all the points),
- can ask for help from me but points will be deducted for each hint given to a group.

Three parts:

Exercise 1

1. Prove the perceptron convergence theorem that the number of updates of the perceptron algorithm is bounded by

$$\frac{R^2}{\gamma^2}$$

where $\|\mathbf{x}_i\| \leq R$ for all $i = 1, \ldots, m$ and γ is the margin of a correctly classifying hyperplane with normalised weight vector $\mathbf{w}^\star$ and no threshold.

Hint: Compute an upper bound on $\|\mathbf{w}_t\|^2$ the norm squared of the weight vector after t updates. Compute a lower bound on the value of $\langle \mathbf{w}_t, \mathbf{w}^\star \rangle$ and use the two bounds to show that t cannot grow indefinitely.

Value: 3 points.

Exercise 2

2. Show how the problem of guaranteeing that a weight vector is learnt that approximates $\langle \mathbf{w}_{\mathrm{SVM}}, \mathbf{x}_i \rangle$ to within $\pm\gamma/2$ for all i is converted to a classification problem.

   Hint: Add an extra dimension to the inputs to cater for the output value of the classifier to be approximated.

   Value: 5 points.

Exercise 3

3. Bound the number of weight vectors in the class:

$$\left\{ \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i : \alpha_i \in \mathbb{N}, \sum_{i=1}^{m} \alpha_i = B \right\}$$

Hint: This is a combinatorial question of how many ways you can place balls into pigeon holes.
Value: 5 points.