Basic Statistical Learning Theory: Overnight problems solutions

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

Exercise: reconstruct the bound on the covering numbers!

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

where $k = O\left(\frac{R^2}{\gamma^2}\right)$.

- 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.

Solution: First observe:

$$\begin{aligned}
\|\mathbf{w}_{t+1}\|^2 &= \langle \mathbf{w}_t + y_i \mathbf{x}_i, \mathbf{w}_t + y_i \mathbf{x}_i \rangle \\
&= \|\mathbf{w}_t\|^2 + 2y_i \langle \mathbf{w}_t, \mathbf{x}_i \rangle + \|\mathbf{x}_i\|^2 \\
&\leq \|\mathbf{w}_t\|^2 + R^2 \\
&\leq (t+1)R^2.
\end{aligned}$$

Secondly

$$\begin{aligned}\langle \mathbf{w}^\star, \mathbf{w}_{t+1} \rangle &= \langle \mathbf{w}^\star, \mathbf{w}_t \rangle + y_i \langle \mathbf{w}^\star, \mathbf{x}_i \rangle \\ &\geq \langle \mathbf{w}^\star, \mathbf{w}_t \rangle + \gamma \\ &\geq (t+1)\gamma.\end{aligned}$$

Putting the two together we have:

$$\begin{aligned}t^2\gamma^2 &\leq \langle \mathbf{w}^\star, \mathbf{w}_t \rangle^2 \\ &\leq \|\mathbf{w}_t\|^2 \leq tR^2,\end{aligned}$$

implying that

$$t \leq \frac{R^2}{\gamma^2}.$$

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.

   Solution Let the normalised SVM solution be $\mathbf{w}_{\mathrm{SVM}}$ with output $y_i = \langle \mathbf{w}_{\mathrm{SVM}}, \mathbf{x}_i \rangle$ on the i-th training example. Consider the weight vector $\hat{\mathbf{w}}^s tar = (\mathbf{w}_{\mathrm{SVM}}, -1)$. If we augment $\mathbf{x}_i$ by an extra component to create two new classification vectors

$$\mathbf{x}_i^a = (\mathbf{x}_i, y_i + \gamma/2),$$

   a negative example $y_i^a = -1$

and $\mathbf{x}_i^b = (\mathbf{x}_i, y_i - \gamma/2)$,

a positive example $y_i^b = +1$.

First note that:

$$y_i^a \langle \hat{\mathbf{w}}^\star, \mathbf{x}_i^a \rangle = -\langle \mathbf{w}_{\mathrm{SVM}}, \mathbf{x}_i \rangle + y_i + \gamma/2 = \gamma/2$$

while $y_i^b \langle \hat{\mathbf{w}}^\star, \mathbf{x}_i^b \rangle = \langle \mathbf{w}_{\mathrm{SVM}}, \mathbf{x}_i \rangle - y_i + \gamma/2 = \gamma/2$

Hence, taking into account that $\|\hat{\mathbf{w}}^\star\|^2 = 2$, it follows that $\hat{\mathbf{w}}^\star$ has margin $\gamma/(2\sqrt{2})$ on the classification vectors, so that the perceptron algorithm will run in at most $8\hat{R}^2/\gamma^2$, where $\hat{R}$ is the radius of the transformed vectors.
If we take into account that

$$|y_i| = |\langle \mathbf{w}_{\mathrm{SVM}}, \mathbf{x}_i \rangle| \leq \|\mathbf{x}_i\| \leq R,$$

and similarly for γ, we obtain

$$\|\mathbf{x}_i^a\|^2 = \|\mathbf{x}_i\|^2 + y_i^2 + \gamma|y_i| + \gamma^2/4 \leq 3.25R^2.$$

Making

$$\frac{8\hat{R}^2}{\gamma^2} = \frac{26R^2}{\gamma^2}.$$

Once we run the perceptron algorithm we will find a vector $(\mathbf{w}, w_n)$ with a sparse dual representation. I claim that $\mathbf{w}^\ddagger = -\mathbf{w}/w_n$ is a $\gamma/2$ approximation of $\mathbf{w}_{\mathrm{SVM}}$:

$$\begin{aligned} \langle \mathbf{w}^\ddagger, \mathbf{x}_i \rangle &= -\frac{1}{w_n}\langle \mathbf{w}, \mathbf{x}_i \rangle \\ &< y_i + \gamma/2, \\ \text{since } -\langle \mathbf{w}, \mathbf{x}_i \rangle - w_n(y_i + \gamma/2) &> 0, \qquad (1) \\ \text{and } \langle \mathbf{w}^\ddagger, \mathbf{x}_i \rangle &= -\frac{1}{w_n}\langle \mathbf{w}, \mathbf{x}_i \rangle \end{aligned}$$

$$> \quad y_i - \gamma/2,$$

since $\langle \mathbf{w}, \mathbf{x}_i \rangle + w_n(y_i - \gamma/2) > 0.$ (2)

Note that the above computations make the assumption that $w_n < 0$. This follows from adding equations (1) and (2) to obtain,

$$-w_n \gamma > 0,$$

which since $\gamma > 0$ implies $w_n < 0$[1].

[1] This final observation was taken from Mingrui Wu's solution!

Exercise 3

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

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

Note this is not actually the covering set - we need to perform the division by the last component described above – but this will not change the number of weights in the class.

Solution: Consider m pigeon holes into which we must place B balls. We must count the number of ways this can be done. Imagine that the balls are black and that we add one white ball to each of

the pigeon holes except the last - i.e. $m-1$ white balls.

I claim there is a one to one correspondence between these configurations and the possible allocations of the black and white balls to $m+B-1$ linear positions:

if we have a pigeon hole configuration, simply transfer the balls in order to the first linear positions with the white ball last, then continue with the second pigeon hole and so on. This gives an allocation of the balls to the linear positions.

given an allocation to the linear positions transfer the balls in order to the pigeon holes moving to the next hole each time a white ball is encountered.

Hence, the total number of ways this can be done

is:

$$\begin{aligned}\binom{m+B-1}{B} &\leq \sum_{i=0}^{B}\binom{m+B-1}{i} \\ &\leq \left(\frac{e(m+B-1)}{B}\right)^{B}\end{aligned}$$

as required.

Generalization of SVMs

Putting these results together we obtain the bound (note that the dual representation was in the artificial classification set which had $4m$ elements - 2 for each of the $2m$ points of the training and ghost sample):

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

$$\epsilon(m, \mathcal{L}, \delta, \gamma) = \frac{2}{m}\left(k \log_2 \frac{e(4m + k - 1)}{k} + \log_2 \frac{m}{\delta}\right)$$

where $k = \frac{26R^2}{\gamma^2}$.