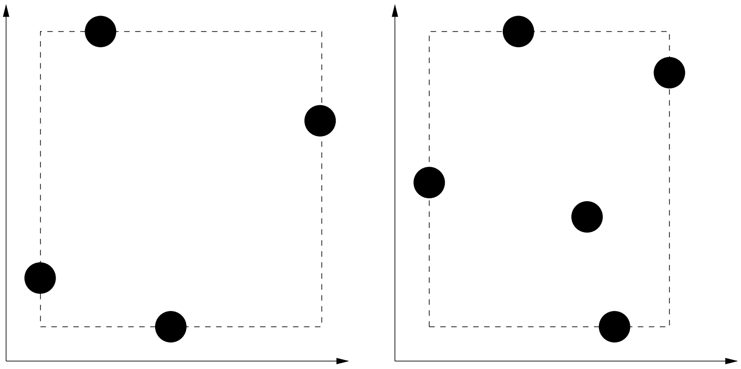
T-61.5030 Advanced course in neural computing

Solutions for exercise 2

1. We first show that the VC-dimension is at least four and then show it is less than five.



Pick four points $P_1, \ldots, P_4$ and let $\mathbf{w}_{\text{opt}}$ represent the smallest rectangle which includes the points. (The four points need to be chosen so that each one lies alone in one of the sides of the rectangle.) We can now implement any classification for the points by moving the side inwards by ϵ for those points which belong to class 0 and outwards for points in class 1. Therefore the VC-dimension is at least four. Note that classification is not possible for any four points. For the VC-dimension it suffices to have all classifications for some four points.

Now consider any five points and the rectangle $\mathbf{w}_{\text{opt}}$. As the rectangle has only four sides, at least one of the points has to be either inside the rectangle or share a side with another point. Order the points so that this point is P_5. (There can be more than one point like this.) It is now obvious that it is impossible to classify $P_1, \ldots, P_4$ into class 1 and the point P_5 into class 0. Since this holds for any five points, VC-dimension is less than five.

2. We first show that the VC-dimension is at least m and then show that it is less than $m+1$, hence the VC-dimension is m.

Suppose there are m input vectors $\mathbf{x}_i$ (whose dimension is m). Denote the classification of input $\mathbf{x}_i$ by d_i so that $d_i = 1$ if $\mathbf{x}_i$ belongs to class 1 and $d_i = -1$ if $\mathbf{x}_i$ belongs to class 0. The linear classifier should satisfy $\mathbf{x}_i^T\mathbf{w} > 0$ if $d_i = 1$ and $\mathbf{x}_i^T\mathbf{w} < 0$ if $d_i = -1$. By denoting $\mathbf{y}_i = d_i\mathbf{x}_i$ we can write this simply as $\forall i : \mathbf{y}_i^T\mathbf{w} > 0$. This is satisfied if $\mathbf{Y}\mathbf{w} = \mathbf{1}$, where $\mathbf{Y} = [\mathbf{y}_1\,\mathbf{y}_2 \ldots \mathbf{y}_m]^T$ and $\mathbf{1} = [1\,1 \ldots 1]^T$. By definition, the first element of $\mathbf{x}_i$ equals to one. We shall now choose x_i such that also the ith element equals to one (for instance $\mathbf{x}_1 = [1\,0\,0 \ldots 0]^T$ and $\mathbf{x}_3 = [1\,0\,1\,0\,0 \ldots 0]^T$). It is fairly easy to see that $\det(\mathbf{Y}) = \prod_i d_i = \pm 1$, and thus $\mathbf{Y}$ is invertible and $\mathbf{Y}\mathbf{w} = \mathbf{1}$ has a nontrivial solution $\mathbf{w} = \mathbf{Y}^{-1}\mathbf{1}$ for any choice of d_i. Therefore the VC-dimension is at least m.

Suppose now that there are $m+1$ input vectors. Now the vectors $\mathbf{x}_i$ are necessarily linearly dependent and the equation $\mathbf{x}_{m+1} = a_1\mathbf{x}_1 + \ldots + a_m\mathbf{x}_m$ has a solution (given that we can choose any of the $\mathbf{x}_i$ to be $\mathbf{x}_{m+1}$). It is now impossible to realise the following classification: $d_{m+1} = -1$ and for $i \neq m+1$, $d_i = 1$ if $a_i > 0$ and $d_i = -1$ otherwise. This is because

$$\mathbf{x}_{m+1}^T \mathbf{w} = \sum_{i=1}^{m} a_i \mathbf{x}_i^T \mathbf{w} = \sum_{i=1}^{m} a_i d_i \mathbf{y}_i^T \mathbf{w} > 0$$

since $a_i d_i \geq 0$ and $\mathbf{y}_i^T \mathbf{w} \geq 0$. Now $\mathbf{y}_{m+1}^T \mathbf{w} = -\mathbf{x}_{m+1}^T \mathbf{w} < 0$ which is against the definition $\mathbf{y}_i^T \mathbf{w} \geq 0$. This shows that for any $m+1$ input vectors $\mathbf{x}_i$ there exists a classification which cannot be realised and the VC-dimension is thus less than $m+1$.

3. From equation (2.97) we have that the inequality $P(w) < \nu(w) + \epsilon\sqrt{P(w)}$ holds with probability $1-\alpha$, where α is the right hand side of equation (2.97). Since ϵ is a function of N, h and α and $P(w)$ is a function of $\nu(w)$ and ϵ, this can be written as $P(w) < \nu(w) + \epsilon_1(N, h, \alpha, \nu(w))$.

   Since the exponent in the factor $e^{-\epsilon^2 N/4}$ in equation (2.97) is $-\epsilon^2 N/4$ whereas that in equation (2.95) of the Haykin's book is $-\epsilon_0^2 N$, it follows that the same α is achieved if $\epsilon = 2\epsilon_0$ and thus

$$\frac{P(w) - \nu(w)}{\sqrt{P(w)}} < \epsilon = 2\epsilon_0 \;\Rightarrow\; P(w) - \nu(w) < 2\epsilon_0\sqrt{P(w)}.$$

   If we define $\epsilon_1 = 2\epsilon_0\sqrt{P(w)}$, we then have $\epsilon_1 = 2\epsilon_0\sqrt{\nu(w) + \epsilon_1}$ and squaring both sides yields $\epsilon_1^2 = 4\epsilon_0^2(\nu(w) + \epsilon_1)$. This is a second order equation with respect to ϵ_1 and solving for ϵ_1 yields

$$\epsilon_1 = \frac{1}{2}\left[4\epsilon_0^2 \pm \sqrt{16\epsilon_0^4 + 16\epsilon_0^2 \nu(w)}\right] = 2\epsilon_0^2\left(1 + \sqrt{1 + \frac{\nu(w)}{\epsilon_0^2}}\right),$$

   where we retained only the positive root.

4. Note: the description of the classifier in the exercise is imprecise. If the classifier has no adjustable parameters, it always gives the same results and is not able to classify even a single input in two different ways. The following description is better: the classifier classifies all girls to class C_1 and all boys to class C_2, and C_1 and C_2 can be chosen by training (they are parameters of the classifier). If the classifier is trained on the sample of 100 teenagers, the choice that gives the smallest training error is C_1 = 'prefers strawberry cake' and C_2 = 'prefers chocolate cake'.

   (a) A training set having one boy and one girl can be classified in all four ways, thus the VC-dimension is at least two. If the training set has more than one boy, both will always be assigned the same classification. The same holds for girls. Since a training set having three or more youngsters necessarily has at least two boys or two girls, the VC-dimension is at most two. Combining the results, the VC-dimension $h = 2$.

   (b) From the problem description we have $\alpha = 0.001$, $N = 100$ and $v = 0.1$ which yields $\epsilon_0 \approx 0.43$ and $\epsilon_1 \approx 0.82$ by equations (2.96) and (2.99) in Haykin's book. This means that with probability 99.9% the classification error for future samples is less than $v + \epsilon = 0.53$.

   (c) The equations are valid as long as the underlying class probabilities are constant. The first group probably has mobile phones and they inform the next group thus altering the probability of choosing cakes. No bounds can be given for the classification error.