1. (d) We minimize
$$\frac{1}{2}\mathbf{w}^T\mathbf{w}$$
subject to the constraint
$$y_n\left(\mathbf{w}^T\mathbf{x}_n + b\right) \geq 1$$
The Lagrangian is
$$\frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_n \alpha_n\left(y_n\left(\mathbf{w}^T\mathbf{x}_n + b\right) - 1\right)$$
and the primal problem requires differentiating with respect to the α_n's, which gives the constraint
$$y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$$
If we re-absorb the b into the vector $\mathbf{w}$ (making it a $d+1$ dimensional vector), we have the mapping
$$\frac{1}{2}\mathbf{w}^T\mathbf{w} \to \frac{1}{2}\mathbf{w}^T\mathbf{P}\mathbf{w}, \qquad y_n(\mathbf{w}^T\mathbf{x}_n + b) \to y_n\mathbf{w}^T\mathbf{x}_n$$
where $\mathbf{P}$ is the $(d+1)\times(d+1)$ matrix
$$\mathbf{P} = \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d\times d}\end{pmatrix}, \qquad \mathbf{I}_{d\times d} \text{ is the identity matrix.}$$
Thus we have the $d+1$ variable quadratic programming problem minimizing with respect to $\mathbf{w}$
$$\frac{1}{2}\mathbf{w}^T\mathbf{P}\mathbf{w}$$
subject to the constraint
$$y_n\mathbf{w}^T\mathbf{x}_n = 1.$$

2. (a) 0 vs all had the lowest in-sample accuracy. Interestingly, it was hard to distinguish the others from the pack. Yet, they had a lower E_{in}.

3. (a) 1 vs all had the lowest E_{in}. It was hard to distinguish the rest from the pack.

4. (c) 0 vs all has 2180 SV and 1 vs all has 386 SV, so the difference is about 1800.

5. (d) The $C = 1$ case has 2 more correct classification points in-sample than the other C's.

6. (b) At $Q = 2$ the number of support vectors is 76 while at $Q = 5$ the number of support vectors is 25.

7. (b) $C = 0.001$ has the most lowest-E_{CV} cases at 28 cases out of 100 iterations.

8. (c) Averaging E_{CV} for $C = 0.001$ gives $E_{\text{CV}} = 0.00478$.

9. (e) $C = 10^6$ had the lowest $E_{\text{in}} = 0.00064$.

10. (c) $C = 100$ had the lowest $E_{\text{out}} = 0.019$.