1. Exercise 8.11
   1. We can see that for , the expression can be rewritten as

, which is the minimization. We can also rewrite as . Since and we have the matrix multiplied by our vector . We see that , therefore our product depends solely on and , which gives us **.** Therefore, we have the equivalent constraint. Hence, we have a standard QP-problem.

* 1. and , therefore

. This is a symmetric matrix, and we then get , which means the matrix is positive semi-definite.

1. Let us have the following data set: . . This gives us

. We then have to solve the dual problem: subject to the constraint . Differentiation yields:. Solving the system gives us . This proves that the reverse of the KKT complementary slackness property does not hold.

1. is the system. Subtracting (i) from (ii) and adding (i) to (iii) gives us for our system. Plotting the data points gives us a linear separator at the line .
2. The dual problem expands into: . Substituting our dataset yields which simplifies to subject to the constraint and . To minimize this function under our constraints, we use Lagrange multipliers. . We can eliminate by making the substitution , resulting in We then get the following derivatives: . Solving these for gives us which satisfies the constraint, .
3. Exercise 8.15
   1. Let . In terms of K,
   2. Define . This then gives us
   3. Assume are kernels but their sum is not. Therefore, a transform s.t. . This is a contradiction because we were able to find a transform. This means that are kernels if are kernels.
4. These are the results for accuracy and total number of support vectors for each model:

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | # Support Vectors |
| Cost = 0.01 | 88.6792% | 1112 |
| Cost = 0.1 | 95.9906% | 424 |
| Cost = 1 | 95.9906% | 163 |
| Cost = 2 | 95.9906% | 131 |
| Cost = 3 | 96.2264% | 117 |
| Cost = 5 | 96.2264% | 102 |
| Linear Kernel | 95.9906% | 162 |
| Polynomial Kernel | 95.5189% | 543 |
| Radial Basis Function Kernel | 95.9906% | 163 |

As we can see, as the cost parameter increases, our margin decreases as well as the number of support vectors and the accuracy increased. It appears that there could be overfitting when our cost parameter goes up to 2,3, and 5. Furthermore, we no longer see significant decreases in the total number of support vectors beyond a cost of 1. As for the kernel models, the linear kernel is best because our data is linearly separable. The polynomial kernel shows overfitting and is clearly not a good choice because we have a significant increase in support vectors. As for radial basis, there is next to no difference. Therefore, a linear kernel is best because it is the simplest kernel that effectively classifies our data without risking overfitting.