Support Vector Machines

Questions

1. Explain why minimizing the norm of the vector of weights maximizes the margin.

   The distance of a point x_n from the separating hyperplane is: $\frac{t_n(w^T x_n + w_0)}{\|w\|}$. The numerator is also the constraint of the SVM for each point: $t_n(w^T x_n + w_0) \geq 1$, which is satisfied at the equality (that is, the constraint is active) for support vectors. Therefore, we can substitute the numerator with 1, and we get that the distance of a support vector from the hyperplane is $\frac{1}{\|w\|}$. It follows that minimizing the norm of **w** maximizes the margin.

2. What do the constraints in the optimization problem represent?

   There is one constraint per point in the dataset. Each constraint imposes that the corresponding point is on the correct side of the separating boundary, and not on the boundary itself.

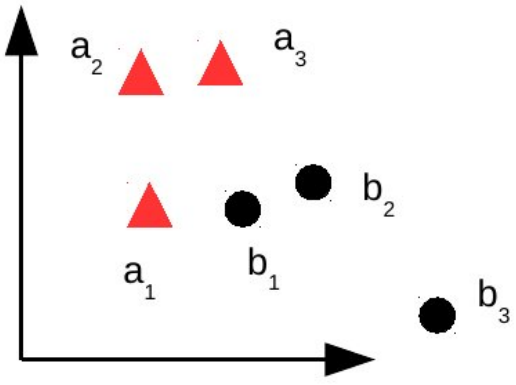
3. What is the role of slack variables? What do they achieve?

   The slack variables achieve a so called *soft* margin, because they allow some points to be on the wrong side of the classification boundary. The average distance of the misclassified points from the boundary is minimized together with the norm of **w**.

4. What is the difference between a soft-margin and a hard-margin svm?

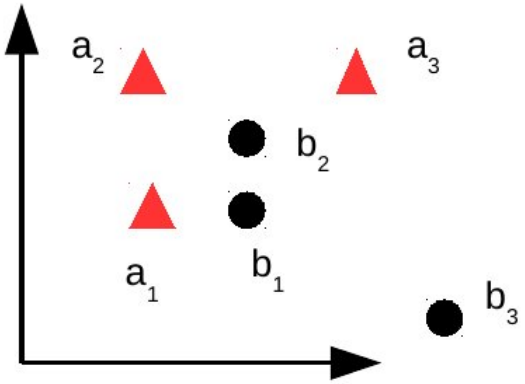
   A soft margin SVM, through the slack variables, allows points to be misclassified, and therefore a solution always exists. For a hard margin SVM, on the other hand, all constraints on correct classification must be fulfilled, therefore if the dataset is not linearly separable there is no hard-margin SVM able to classify that dataset.

5. Given the dataset below, determine if a hard-margin SVM can separate the classes, and, if that is the case, identify the support vectors:



Yes, the dataset is linearly separable. The point a_1 and b_1 are definitely support vectors, while a_3 and b_2 might also be.

6. Same as the question before:



This dataset is not linearly separable, therefore a hard-margin SVM cannot classify it.

7. What options do you have with support vector machines if the dataset is not linearly separable?

   Soft margin with slack variables, and projecting the dataset to a higher dimensional space through a set of basis functions. The two are not mutually exclusive.

8. What is the kernel trick and what does it achieve?

   The kernel trick allows us to compute the dot product of certain feature vectors much faster than by using the general definition of dot product. In the dual formulation of SVMs the input vectors, projected through a set of basis functions, only appear in dot products. Therefore, we can use very high dimensional projections (indeed even with infinite dimensions!) implicitly through the kernel trick, when computing the dot product explicitly would not be possible.

9. Why are kernels useful?

   Given some initial kernel functions, that have been discovered for specific basis functions, it is possible to engineer kernels that correspond to very high dimensional projections, without ever having to explicitly compute the coordinates of the projected points. A data set that is not linearly separable in its original dimensionality, when projected to a higher dimensional space, may become linearly separable. With enough many dimensions, every dataset is linearly separable.

10. Consider the dataset: $\{\langle 0,0,-1\rangle, \langle 0,1,1\rangle, \langle 1,0,1\rangle\}$ where the last element of each vector is the class $t \in \{-1,1\}$. We want to use a linear (non-kernel) support vector machine classifier to specify the decision boundary in the form of $y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0$. Let a_1, a_2, a_3 denote the Lagrange multipliers for the constraints on x_1, x_2, and x_3 respectively.

    1. Plot the data points and derive the decision boundary by inspecting the data. What can be said about the Lagrange multipliers?

2. Write the Lagrangian, apply the optimality conditions, and express the vector **w** in terms of the data points.

   All three points are support vectors, since it can be easily seen that removing one would change the decision boundary. This implies that all three Lagrange multipliers are non-zero.

   The write the Lagrangian we first note that the constraints in the form $t_i(\boldsymbol{w}^T x_i + w_0) \geqslant 1$ can be rewritten as $1 - t_i(\boldsymbol{w}^T x_i + w_0) \leqslant 0$, that is $g(x_i) \leqslant 0$ The Langrangian is:

   $L(\boldsymbol{w}, \boldsymbol{a}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + a_1(1 - t_1(\boldsymbol{w}^T x_1 + w_0)) + a_2(1 - t_2(\boldsymbol{w}^T x_2 + w_0)) + a_3(1 - t_3(\boldsymbol{w}^T x_3 + w_0))$.

   The optimality condition that allows us to express w in terms of a is $\nabla_w L(\boldsymbol{w}, \boldsymbol{a}) = 0$, from which:

   $\boldsymbol{w} - a_1 t_1 x_1 - a_2 t_2 x_2 - a_3 t_3 x_3 = 0$,

   and therefore:

   $\boldsymbol{w} = -a_1\langle 0,0\rangle + a_2\langle 0,1\rangle + a_3\langle 1,0\rangle$. To compute w_0 we can use the constraint of any support vector (so in this case any of the three points), for instance:

   $\boldsymbol{w}^T\langle 0,1\rangle + w_0 = 1$ from which: $w_0 = 1 - \boldsymbol{w}^T\langle 0,1\rangle$

11. Consider the dataset: $\{\langle -1,0\rangle, \langle 1,0,1\rangle, \langle 2,0,1\rangle\}$. Compute the value of the three Lagrange multipliers corresponding to each point.

    I'll denote the weight vector with $\boldsymbol{w_a} = \langle w_0, w_1, w_2\rangle$, and the gradient of the decision boundary with $\boldsymbol{w} = \langle w_1, w_2\rangle$. The third point is not a support vector, therefore $\lambda_3 = 0$. If we plot the dataset, it is clear that the optimal decision boundary has equation $x = 0$, therefore the corresponding weight vector is $\boldsymbol{w_a} = \langle 0, 1, 0\rangle$, or any vector parallel to it, that is, that can be obtained by it through a multiplication by a constant c. For instance, the vector $\langle 0,2,0\rangle$ would give the same decision boundary, since it corresponds to the equation $2x = 0$.

    We first need to identify this constant by looking at any constraint corresponding to a support vector, for example, for the first point:

    $1 + (\begin{pmatrix} c & 0 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \end{pmatrix}) = 0$,

    which leads to $c = 1$, and therefore no scaling of the weight vector is necessary (since multiplying all its elements by 1 changes nothing).

    We know from the previous question, that the application of the optimality condition leads to the expression of the vector of weights in terms of the support vectors and the classes:

    $\boldsymbol{w} = -\lambda_1\langle -1,0\rangle + \lambda_2\langle 1,0\rangle$.

    We also know that by applying the optimality condition on the derivative with respect to w_0 (see slide 7 in deck SVM part 3) we obtain:

    $-\lambda_1 + \lambda_2 + \lambda_3 = 0$. Since $\lambda_3 = 0$ We then have two equations in two unknowns and can solve for the Lagrange multipliers:

$$\begin{cases} \lambda_1+\lambda_2=1 \\ -\lambda_1+\lambda_2=0 \end{cases}$$

From which $\lambda_1=0.5, \lambda_2=0.5$ and we already knew that $\lambda_3=0$

12. Consider the dataset: $\{\langle -1,0,-1\rangle, \langle 1,0,1\rangle, \langle 2,0,-1\rangle\}$, that is, the same as the previous question, with class of the last point inverted. This dataset is not linearly separable, therefore we need to introduce slack variables. Assuming that the decision boundary is still $x=0$, what is the value of the slack variables ξ_1, ξ_2, ξ_3 , corresponding to each point?

    The first and second point are on the correct side of the decision boundary, therefore: $\xi_1=0, \xi_2=0$. For the third point, we can compute its slack variable from the constraint $t_i(\boldsymbol{w}^T x_i+w_0)\geqslant 1-\xi_i$,which for the third point gives the minimum value of ξ_3 as:

    $-1\left(\begin{pmatrix}1 & 0\end{pmatrix}\begin{pmatrix}2\\0\end{pmatrix}\right)=1-\xi_3$.

    We obtain $\xi_3=3$. Note that this is exactly the distance between x_1 and x_3, that is, between x_3 and the margin of the side it belongs to, since the norm of **w** is 1 (see slide 8 in the deck SVM part 2).