Practical Session 6

Constrained Optimization and SVM

1 Duality of finding maxima of a set

Problem 1: Given a set of variables $x_1, ..., x_N \in \mathbb{R}$, define an equation that finds the largest value in the set via minimization. Then, use the Lagrange dual function to derive a second, equivalent maximization problem.

The following equation finds the maximum via optimization:

$$\begin{aligned} &\text{minimize} && b \\ &\text{subject to} && b \geq x_i \;\Leftrightarrow\; x_i - b \leq 0, \quad i = 1, \ldots, N\,. \end{aligned}$$

Intuitively, we define an upper limit b and reduce it as much as possible while being greater or equal than all values in our set. We can now apply our recipe for solving a constrained optimization problem, with parameters $\boldsymbol{\theta} = b$ and Lagrange multipliers $\boldsymbol{\alpha} = \boldsymbol{w}$:

1. Calculate the Lagrangian:
$$L(b, \boldsymbol{w}) = b + \sum_{i=1}^{N} w_i(x_i - b)$$

2. Obtain the Lagrange dual function by solving $\boldsymbol{\nabla}_b L(b, \boldsymbol{w}) = 0$:
$$\begin{aligned} &\boldsymbol{\nabla}_b L(b, \boldsymbol{w}) = 1 - \sum_{i=1}^{N} w_i = 0 \\ \Leftrightarrow \quad &\sum_{i=1}^{N} w_i = 1 \end{aligned}$$
This condition holds if the parameter b is optimal. Using this, we obtain
$$g(\boldsymbol{w}) = L(b^*(\boldsymbol{w}), \boldsymbol{w}) = b^*(\boldsymbol{w}) + \sum_{i=1}^{N} w_i(x_i - b^*(\boldsymbol{w})) = \sum_{i=1}^{N} w_i x_i + b^*(\boldsymbol{w}) \left(1 - \sum_{i=1}^{N} w_i\right) = \sum_{i=1}^{N} w_i x_i$$

3. The dual problem is
$$\begin{aligned} &\text{maximize} && \sum_{i=1}^{N} w_i x_i \\ &\text{subject to} && \sum_{i=1}^{N} w_i = 1\,, \\ & && w_i \geq 0 \quad i = 1, \ldots, N\,. \end{aligned}$$

The dual problem we've found is another intuitive solution of our problem. We assign weights w_i to all elements in our set. By maximizing the sum $\sum_{i=1}^{N} w_i x_i$ while keeping the sum of all w_i's constant, the full weight will be assigned to the maximum element. The solution will then return the index of the maximum element.

Problem 2: Given a set of variables $x_1, ..., x_N \in \mathbb{R}$, define an equation that calculates the sum of the k largest values via maximization. Then, use the Lagrange dual function to derive a second, equivalent minimization problem.

Defining a minimization problem as we did above is not as obvious in this case. So, instead we start with a maximization problem similar to the one we derived before:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{N} w_i x_i \\ \text{subject to} \quad & \sum_{i=1}^{N} w_i = k\,, \\ & w_i - 1 \leq 0 \quad i = 1, \ldots, N\,, \\ & -w_i \leq 0 \quad i = 1, \ldots, N\,. \end{aligned}$$

Since all weights w_i are constrained to be between 0 and 1, the solution of this problem will find the indices of the largest elements and return their sum. This time, we will apply our recipe in the opposite direction to find the Lagrange dual problem, with parameters $\boldsymbol{\theta} = \boldsymbol{w}$ and Lagrange multipliers $\boldsymbol{\alpha} = (b, \boldsymbol{s}, \boldsymbol{t})$:

1. Calculate the Lagrangian:

$$L(\boldsymbol{w}, b, \boldsymbol{s}, \boldsymbol{t}) = -\sum_{i=1}^{N} w_i x_i + b\left(\sum_{i=1}^{N} w_i - k\right) + \sum_{i=1}^{N} s_i(w_i - 1) - \sum_{i=1}^{N} t_i w_i$$

   Note that since we are maximizing we need to flip the sign of the function (since the maximizer of $f(\boldsymbol{x})$ is the minimizer of $-f(\boldsymbol{x})$).

2. Obtain the Lagrange dual function by solving $\boldsymbol{\nabla}_{\boldsymbol{w}} L(\boldsymbol{w}, b, \boldsymbol{s}, \boldsymbol{t}) = 0$:

$$\partial_{w_i} L(\boldsymbol{w}, b, \boldsymbol{s}, \boldsymbol{t}) = -x_i + b + s_i - t_i = 0$$

   This condition holds if the parameter $\boldsymbol{w}$ is optimal. Using this, we obtain

$$\begin{aligned} g(b, \boldsymbol{s}, \boldsymbol{t}) = L(\boldsymbol{w}^*(b, \boldsymbol{s}, \boldsymbol{t}), b, \boldsymbol{s}, \boldsymbol{t}) &= -\sum_{i=1}^{N} w_i^* x_i + b\left(\sum_{i=1}^{N} w_i^* - k\right) + \sum_{i=1}^{N} s_i(w_i^* - 1) - \sum_{i=1}^{N} t_i w_i^* \\ &= -kb - \sum_{i=1}^{N} s_i + \sum_{i=1}^{N} w_i^* \left(-x_i + b + s_i - t_i\right) \\ &= -kb - \sum_{i=1}^{N} s_i \end{aligned}$$

3. Since all t_i's are non-negative numbers and not important for the final solution, we can write the above condition as $-x_i + b + s_i = t_i \geq 0$. Using this, we can write the dual problem as

$$\begin{aligned} \text{maximize} \quad & -kb - \sum_{i=1}^{N} s_i \\ \text{subject to} \quad & b \geq x_i - s_i \quad i = 1, \ldots, N\,, \\ & s_i \geq 0 \quad i = 1, \ldots, N\,. \end{aligned}$$

Note that we don't have any constraints on the Lagrange multiplier b since it stems from an equality constraint. We can now simply flip the sign to obtain a minimization problem:

$$\begin{aligned} \text{minimize} \quad & kb + \sum_{i=1}^{N} s_i \\ \text{subject to} \quad & b \geq x_i - s_i \quad i = 1, \ldots, N\,, \\ & s_i \geq 0 \quad i = 1, \ldots, N\,. \end{aligned}$$

We have now found a problem formulation that is analogous to the one we used in the previous problem! Intuitively, b returns the value of the k'th largest element. The values s_i return the difference of the $k-1$ larger elements to this one and are 0 for all smaller or equal elements. Hence, the above sum is the sum of the k largest elements.

2 Constrained Optimization Toy Problem

Suppose we have 40 pieces of raw material. Toy A can be made of one piece material with 3 EUR machining fee. A larger toy B can be made from two pieces of material with 5 EUR machining fee.

Because distribution costs decrease with larger quantities, we can sell x pieces of toy A for $20 - x$ EUR each, and y pieces of toy B for $40 - y$ EUR each. From our experience, toy B is more popular than toy A; therefore, we will produce not more of toy A than of toy B.

To get the maximum profit, we want to calculate the amount of toy A and toy B that we should produce.

Problem 3: Write down the constrained optimization problem and the associated Lagrangian.

The constrained optimization problem is

$$\begin{aligned} \min f(x, y) &= -[x(20 - x) + y(40 - y) - 3x - 5y] = x^2 - 17x + y^2 - 35y \\ s.t.\ f_1(x, y) &= x + 2y - 40 \leq 0 \\ f_2(x, y) &= x - y \leq 0 \end{aligned}$$

and the associated Lagrangian is given by

$$L(x, y, \alpha_1, \alpha_2) = x^2 - 17x + y^2 - 35y + \alpha_1(x + 2y - 40) + \alpha_2(x - y)$$

with $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$.

Problem 4: Write down the Karush–Kuhn–Tucker (KKT) conditions for the above optimization problem.

Primal feasibility:

$$\begin{aligned} x + 2y - 40 &\leq 0 \\ x - y &\leq 0 \end{aligned}$$

Dual feasibility:

$$\begin{aligned} \alpha_1 &\geq 0 \\ \alpha_2 &\geq 0 \end{aligned}$$

Complementary slackness:

$$\begin{aligned} \alpha_1(x + 2y - 40) &= 0 \\ \alpha_2(x - y) &= 0 \end{aligned}$$

x, y minimize Lagrangian:

$$\begin{aligned} \frac{\partial L}{\partial x} &= 2x - 17 + \alpha_1 + \alpha_2 = 0 \\ \frac{\partial L}{\partial y} &= 2y - 35 + 2\alpha_1 - \alpha_2 = 0 \end{aligned}$$

Problem 5: Obtain the solution to the constrained optimization problem by solving the KKT conditions. Do not worry about non-integer production quantities.

We start by observing that the KKT complementary slackness conditions demand that either $\alpha_1 = 0$ or $x + 2y = 40$ (i.e. production is limited by available resources) and $\alpha_2 = 0$ or $x = y$ (i.e. production is limited by our desire to not produce more of toy A than toy B). We *guess* that the production will be limited by available resources and not by our desire to produce more of toy A then toy B, thus

$$\begin{aligned} x + 2y - 40 &= 0 \\ \alpha_2 &= 0\,. \end{aligned}$$

Solving the first condition for x gives $x = 40 - 2y$.

Substituting this expression for x together with $\alpha_2 = 0$ into the KKT minimization conditions ($\partial L/\partial x = 0$, $\partial L/\partial y = 0$) and solving for y and then x gives

$$y = 16.1$$
$$x = 7.8\,.$$

We now have to check the correctness of our assumptions by verifying the remaining KKT conditions. Since we explicitly used $x + 2y = 40$, the first primal feasibility condition is satisfied with equality. We explicitly check that $x \leq y$ is satisfied. We assumed $\alpha_2 = 0$, thus the second dual feasibility condition is satisfied. To check that $\alpha_1 \geq 0$, we solve the KKT minimization conditions for α_1, insert our values for x, y and obtain $\alpha_1 = 1.4 > 0$.

Consequently our guess was correct and we obtained a solution for the constrained optimization problem from the KKT conditions. If any constraint would have been violated, we would have to try a new combination of active/inactive constraints. Since there are 2^N combinations of active/inactive constraints for N constraints, the method shown here is only effective for reasonably small N.

3 Concrete SVM Example

You are given a data set with data from a single feature x in $\mathbb{R}$ and corresponding labels $y \in \{+1, -1\}$. Data points for $+1$ are at $-3, -2, 3$ and data points for -1 are at $-1, 0, 1$.

Problem 6: Can this data set in its current feature space be separated using a linear separator? Why/why not?

The convex hulls intersect. Thus it is not linearly separable.

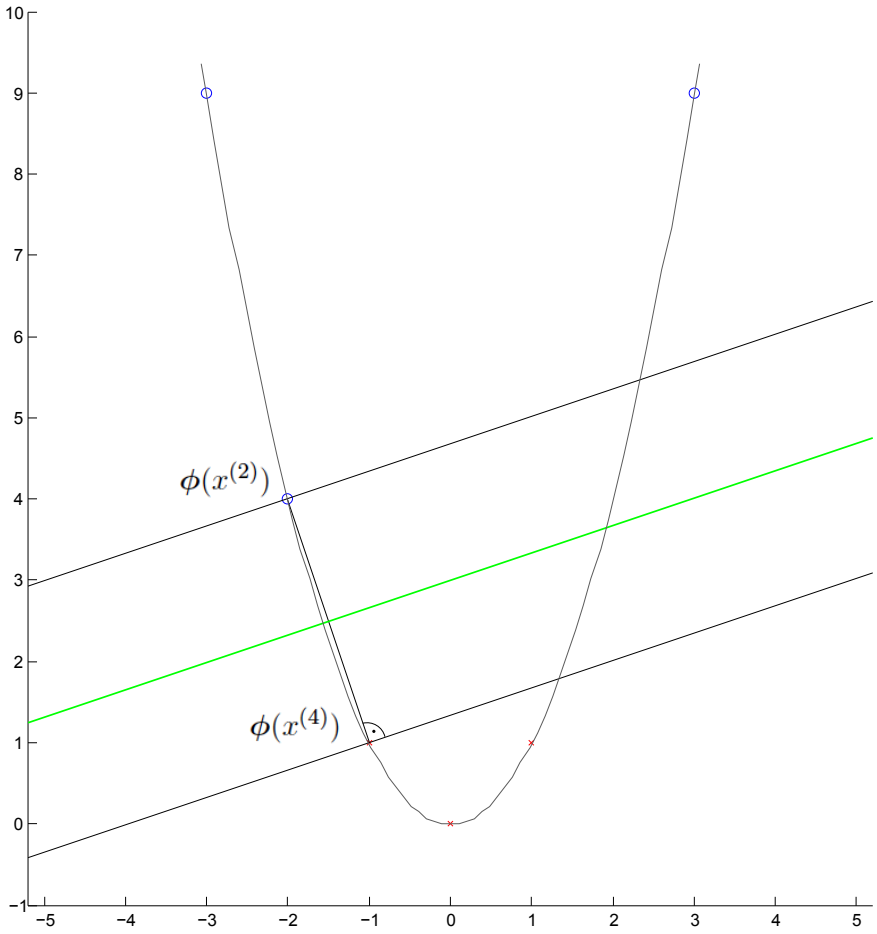
Now, we define a simple feature map $\boldsymbol{\phi}(x) = (x, x^2)$ that transforms points in $\mathbb{R}$ to points in $\mathbb{R}^2$.

Problem 7: After applying $\boldsymbol{\phi}$ to the data, can it now be separated using a linear separator? Why/why not? (Plotting the data may help you with your answer.)

If you plot the data you can easily see that there exists plenty of room to draw a line that separates the two sets.

Problem 8: *Construct* a maximum-margin separating hyperplane (i.e. you do not need to solve a quadratic program). Clearly mark the support vectors. Also draw the resulting decision boundary in the feature space $\boldsymbol{\phi}(x) = (x, x^2)$. Is it possible to add another point to the training set in such a way, that the hyperplane *does not* change? Why/why not?

We know that the margin must at least touch one point of every class otherwise it could not be maximal. Obviously the margin cannot be smaller than the smallest distance between every pair of points from both classes. From the plot (see below) we see that the points $\phi(x^{(2)})$ and $\phi(x^{(4)})$ are closest together. The margin will thus touch both these points. We now have to decide on the angle. The margin, that is the distance between the hyperplane that goes through $\phi(x^{(2)})$ and the hyperplance that goes through $\phi(x^{(4)})$, will be maximal if the hyperplane are orthogonal to the line connecting $\phi(x^{(2)})$ with $\phi(x^{(4)})$. (This can easily be seen from the Pythagorean theorem.)



By plotting the function $f(x) = x^2$ we see that it crosses the decision boundary at approx. $x = 1.9$ and approx. $x = -1.6$. Therefore the points between 1.9 and -1.6 in the original feature space are labeled -1, all others are labeled $+1$.

If you add a new training point in such a way that its feature vector does not lay closer to the above hyperplane than the two support vectors, nothing changes.

Problem 9: For this specific training set write down the SVM optimization problem, the Lagrangian, the Lagrange dual function and the dual problem.

The primal optimization problem is

$$\begin{aligned}
\text{minimize} \quad & f_0(w_1, w_2, b) = \frac{1}{2}(w_1^2 + w_2^2) \\
\text{subject to} \quad & \forall i : y^{(i)}(\boldsymbol{w}^T \boldsymbol{\phi}(x^{(i)}) + b) - 1 \geq 0 \\
\text{that is subject to} \quad & -3w_1 + 9w_2 + b - 1 \geq 0 \\
& -2w_1 + 4w_2 + b - 1 \geq 0 \\
& 3w_1 + 9w_2 + b - 1 \geq 0 \\
& w_1 - w_2 - b - 1 \geq 0 \\
& -b - 1 \geq 0 \\
& -w_1 - w_2 - b - 1 \geq 0\,.
\end{aligned}$$

The Lagrangian is

$$\begin{aligned}
L(w_1, w_2, b, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) = & \frac{1}{2}(w_1^2 + w_2^2) - \\
& \alpha_1 [-3w_1 + 9w_2 + b - 1] - \alpha_2 [-2w_1 + 4w_2 + b - 1] - \\
& \alpha_3 [3w_1 + 9w_2 + b - 1] - \alpha_4 [w_1 - w_2 - b - 1] - \\
& \alpha_5 [-b - 1] - \alpha_6 [-w_1 - w_2 - b - 1]
\end{aligned}$$

The Lagrange dual function is

$$g(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) = \sum_{i=1}^{6} \alpha_i - \frac{1}{2} \sum_{i=1}^{6} \sum_{j=1}^{6} y_i y_j \alpha_i \alpha_j (\phi(x^{(i)})_1 \phi(x^{(j)})_1 + \phi(x^{(i)})_2 \phi(x^{(j)})_2)$$

with

$$\boldsymbol{\phi}(x^{(1)}) = \begin{pmatrix} -3 \\ 9 \end{pmatrix} \quad \boldsymbol{\phi}(x^{(2)}) = \begin{pmatrix} -2 \\ 4 \end{pmatrix} \quad \boldsymbol{\phi}(x^{(3)}) = \begin{pmatrix} 3 \\ 9 \end{pmatrix} \quad \boldsymbol{\phi}(x^{(4)}) = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \quad \boldsymbol{\phi}(x^{(5)}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \boldsymbol{\phi}(x^{(6)}) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The dual problem is

$$\begin{aligned}
\text{maximize} \quad & g(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6) \\
\text{subject to} \quad & \alpha_1 + \alpha_2 + \alpha_3 = \alpha_4 + \alpha_5 + \alpha_6 \\
& \alpha_1 \geq 0 \\
& \alpha_2 \geq 0 \\
& \alpha_3 \geq 0 \\
& \alpha_4 \geq 0 \\
& \alpha_5 \geq 0 \\
& \alpha_6 \geq 0\,.
\end{aligned}$$

Problem 10: Write down the KKT conditions for this training set explicitly and verify that the maximum-margin hyperplane you constructed satisfies them.

$$
\begin{aligned}
-3w_1 + 9w_2 + b - 1 &\geq 0 && \text{primal feasibility}\\
-2w_1 + 4w_2 + b - 1 &\geq 0 && \text{primal feasibility}\\
3w_1 + 9w_2 + b - 1 &\geq 0 && \text{primal feasibility}\\
w_1 - w_2 - b - 1 &\geq 0 && \text{primal feasibility}\\
-b - 1 &\geq 0 && \text{primal feasibility}\\
-w_1 - w_2 - b - 1 &\geq 0 && \text{primal feasibility}\\
\alpha_1 &\geq 0 && \text{dual feasibility}\\
\alpha_2 &\geq 0 && \text{dual feasibility}\\
\alpha_3 &\geq 0 && \text{dual feasibility}\\
\alpha_4 &\geq 0 && \text{dual feasibility}\\
\alpha_5 &\geq 0 && \text{dual feasibility}\\
\alpha_6 &\geq 0 && \text{dual feasibility}\\
\alpha_1(-3w_1 + 9w_2 + b - 1) &= 0 && \text{complementary slackness}\\
\alpha_2(-2w_1 + 4w_2 + b - 1) &= 0 && \text{complementary slackness}\\
\alpha_3(3w_1 + 9w_2 + b - 1) &= 0 && \text{complementary slackness}\\
\alpha_4(w_1 - w_2 - b - 1) &= 0 && \text{complementary slackness}\\
\alpha_5(-b - 1) &= 0 && \text{complementary slackness}\\
\alpha_6(-w_1 - w_2 - b - 1) &= 0 && \text{complementary slackness}\\
w_1 &= -3\alpha_1 - 2\alpha_2 + 3\alpha_3 + \alpha_4 - \alpha_6 && \partial L/\partial w_1 = 0\\
w_2 &= 9\alpha_1 + 4\alpha_2 + 9\alpha_3 - \alpha_4 - \alpha_6 && \partial L/\partial w_2 = 0\\
\alpha_1 + \alpha_2 + \alpha_3 &= \alpha_4 + \alpha_5 + \alpha_6 && \partial L/\partial b = 0
\end{aligned}
$$

Obviously the primal feasibility conditions are fulfilled because the "hyperplane" separates the dataset.

Because only $\boldsymbol{\phi}(x^{(2)})$ and $\boldsymbol{\phi}(x^{(4)})$ are support vectors, all other αs must be forced to zero by the complementary slackness conditions:

$$\alpha_1 = \alpha_3 = \alpha_5 = \alpha_6 = 0$$

The condition $\partial L/\partial b$ then reduces to

$$\alpha_2 = \alpha_4$$

and the weights are given by

$$
\begin{aligned}
w_1 &= -\alpha_2\\
w_2 &= 3\alpha_2\,.
\end{aligned}
$$

From the lecture we know that the size of the margin is given by $m = 2/||\boldsymbol{w}||$ and thus $||\boldsymbol{w}|| = 2/m$. From the plot we know that $m = ||\boldsymbol{\phi}(x^{(4)}) - \boldsymbol{\phi}(x^{(2)})|| = \sqrt{1^2 + 3^2} = \sqrt{10}$, thus we have

$$||\boldsymbol{w}|| = \sqrt{\alpha_2^2 + 3^2\alpha_2^2} = \sqrt{10}\alpha_2 = \frac{2}{\sqrt{10}}$$

which gives

$$\alpha_2 = \alpha_4 = \frac{2}{10} = 0.2\,.$$

Since this is obviously larger than zero all KKT conditions are fulfilled.