Exercise session notes - Week 4

Support Vector Machine + Duality

1 Support Vector Machine

We first review the definition of a Support Vector Machine (SVM).
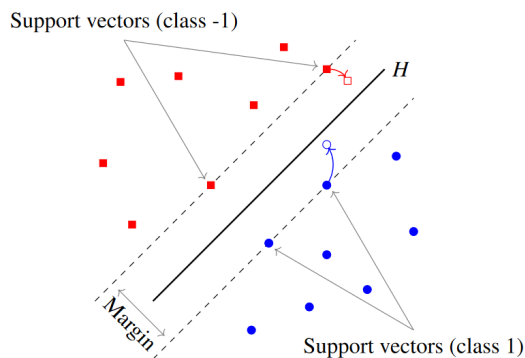
Definition. Support Vector Machine
<u>Input.</u> points $x_1, \ldots, x_m \in \mathbb{R}^n$ with labels $y_1, \ldots, y_m \in \{-1, 1\}$.
<u>Output.</u> Hyperplane (w, b) separating points with different labels, i.e.

$$
\begin{aligned}
w^\top x_i + b &\geq 1 \quad \forall i \text{ with } y_i = 1 \\
w^\top x_i + b &\leq -1 \quad \forall i \text{ with } y_i = -1
\end{aligned}
$$

maximizing the margin between the set of points.

In the lecture we saw that if the points are linearly separable, then the margin is given by $\frac{2}{\|w\|}$. Therefore we can model the problem as the following convex program.

$$
\begin{aligned}
\min \quad & \|w\|^2 && \leftarrow \text{convex objective} \\
& y_i(w^\top x_i + b) \geq 1 && \leftarrow \text{linear constraint} \\
& w, b \text{ free}
\end{aligned}
$$

Often the set of points is not linearly separable, and in this case we can make use of Feature Maps. Assume we have the following set of points
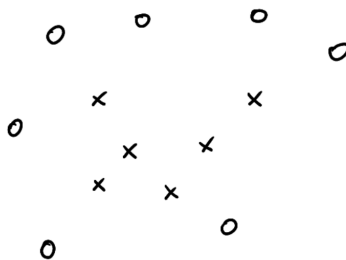
Note that these points are not linearly separable but we see that the crosses are concentrated within some range and circles are outside of that range. Therefore we apply the feature map $(x, y) \mapsto (x, y, x^2 + y^2)$ and obtain a new set of points that are now linearly separable and we can apply the SVM algorithm.
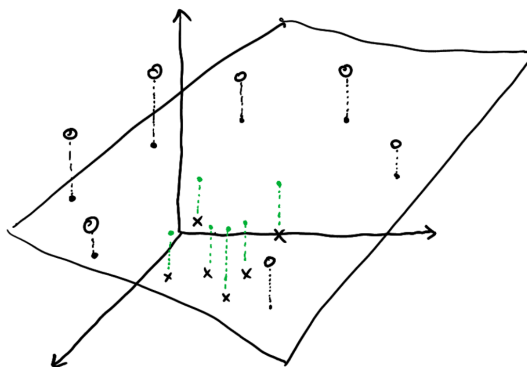
Note that although there always exists a feature map that makes the points linearly separable it is usually very hard to find one.

2   Duality

Given is a <u>general</u> mathematical program (P)

$$\min \{f(x) \mid g_i(x) \leq 0, \ h_j(x) = 0\}$$

For such a program we define the **Lagrangian**

$$L(x, \lambda, \nu) = f(x) + \sum \lambda_i g_i(x) + \sum \nu_j h_j(x)$$

with Lagrange multipliers $\lambda_i \geq 0$, $\nu_j \in \mathbb{R}$. Note that the Lagrangian is affine in λ, ν and for any point x feasible for (P) we have

$$L(x, \lambda, \nu) \leq f(x) \qquad \text{for } \lambda \geq 0, \nu \text{ free}$$

We define the **Lagrange dual function**

$$\widehat{L}(\lambda, \nu) = \inf_x \ L(x, \lambda, \nu)$$

and note that this function is <u>concave</u> (as infimum of concave functions). Finally we define the **Lagrange dual program** (D)

$$\max \left\{ \widehat{L}(\lambda, \nu) \ \Big| \ \lambda \geq 0, \ \nu \text{ free} \right\}$$

Let's do some simple calculations: For any feasible point x for (P) and any feasible point λ, ν for (D) it holds

$$\widehat{L}(\lambda, \nu) \;\leq\; L(x, \lambda, \nu) \;\leq\; f(x)$$

In particular, since the the two sides of the inequality are independent, we can squash the values and we get

$$\max_{\lambda, \nu} \; \widehat{L}(\lambda, \nu) \;\leq\; \min_{x} \; f(x)$$

This just says that if the two programs are feasible, the optimal value of (D) gives us a lower bound on the optimal value of (P). This property is a special case of the Weak Duality Theorem. Moreover note that for any general mathematical program (P) its Lagrange dual program (D) is always a convex program, therefore it can be solved efficiently. In particular for any (also NP-hard) problem we can find efficiently a lower bound on the optimal value. Next week we will see the general statement of Weak Duality Theorem and the Strong Duality Theorem.