Machine Learning, Tutorial 5
University of Bern

Mehdi Noroozi (noroozi@inf.unibe.ch)

24/10/2018

Constraint Optimization and SVM

1. Consider the constrained minimisation problem below. Solve it using the KKT conditions.

$$\begin{aligned} \min_{x,y} \quad & x^2 + y^2 \\ \text{subject to} \quad & (x-3)^2 + y^2 \leq 4 \end{aligned}$$

Solution
The Lagrangian of the problem is $L(\alpha, x, y) = x^2 + y^2 + \alpha((x-3)^2 + y^2 - 4)$. The KKT conditions are:

$$\begin{aligned} \text{primal feasibility:} \quad & (x-3)^2 + y^2 \leq 4 \\ \text{dual feasibility:} \quad & \alpha \geq 0 \\ \text{complementary slackness:} \quad & \alpha((x-3)^2 + y^2 - 4) = 0 \\ \text{gradient of Lagrangian vanishes:} \quad & \tfrac{\partial L}{\partial x} = 2x + 2\alpha(x-3) = 0 \\ & \tfrac{\partial L}{\partial y} = 2y + 2\alpha y = 0 \end{aligned}$$

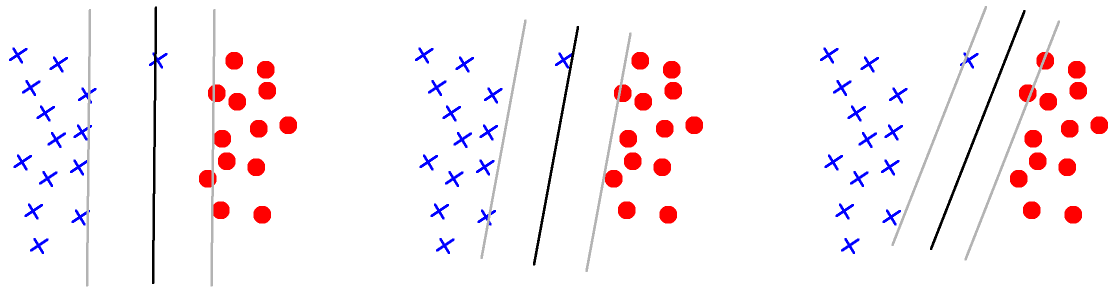
Because $1 + \alpha \geq 1 + 0 > 0$, the only way to satisfy the last equation $(1+\alpha)y = 0$, if we set $y = 0$. According to the complementary slackness, either α or the other term is 0. If $\alpha = 0$, then $x = 0$ (from the 4^{th} equation). When $x = 0$ and $y = 0$, the primal feasibility is not satisfied, therefore $a > 0$, therefore $(x-3)^2 - 4 = 0$. This leads to $x = 1$ or $x = 5$, but only $x = 1$ is feasible. The solution is therefore $x = 1$ and $y = 0$.

2. Let $\{x^{(i)}, y^{(i)}\}_{i=1}^m$ be a set of m training examples of feature vectors $x^{(i)}$ and corresponding labels $y^{(i)}$. We consider binary classification, and assume $y^{(i)} \in \{-1, +1\}$ for $i = 1, \ldots, m$. The following is the primal formulation of L^2-SVM, a variant of the standard Support Vector Machine (SVM) formulation obtained by squaring the hinge loss:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^m \xi_i^2 \\ \text{s.t.} \quad & y^{(i)}(w^\top x^{(i)} + b) \geq 1 - \xi_i, \ i = 1, \ldots, m, \\ & \xi_i \geq 0, \ i = 1, \ldots, m, \end{aligned}$$

where $w,\ b$ are parameters of the usual SVM binary classifier model, ξ are slack variables, and $C > 0$ is a tuning parameter.

   (a) Show that removing the last set of inequality constraints $\{\xi_i \geq 0\}_{i=1}^m$ does not change the optimal solution of the above primal formulation.
   Hint: Consider how the constraints $y^{(i)}(w^\top x^{(i)} + b) \geq 1 - \xi_i$ change when ξ_i is positive or negative.
   Solution: Let (w^*, b^*, ξ^*) be the optimal solution to the problem without the last set of constraints. It suffices to show that $\xi_i^* \geq 0, \forall i$. Suppose $\exists \xi_j^* : \xi_j^* < 0$. Then we have $y^{(j)}(w^\top x^{(j)} + b) \geq 1 - \xi_j^* > 1$, implying that $\xi_j' = 0$ is a feasible solution that results in a smaller objective value. This contradicts the assumption that ξ_j^* is optimal.



a) b) c)

(b) After removing the last set of inequality constraints we arrive at the simplified problem:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 \\ \text{s.t.} \quad & y^{(i)}(w^\top x^{(i)} + b) \geq 1 - \xi_i, \ i = 1, \ldots, m. \end{aligned}$$

Write the dual formulation of this problem.
Hint: Recall that the dual formulation involves the Lagrangian of the primal problem, and that the Lagrangian is some linear combination of the objective function and of the constraints.
Solution: The Lagrangian is given by

$$L(w, b, \xi, \alpha) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m}\xi_i^2 - \sum_{i=1}^{m}\alpha_i(y_i(w^T x_i + b) - 1 + \xi_i), \tag{1}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers.

3. The regularized Support Vector Machine minimizes the following objective,

$$\begin{aligned} \min_{\xi,w,b} \quad & \tfrac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, ..., m \\ & \xi_i \geq 0, \qquad\qquad i = 1, ..., m. \end{aligned}$$

The figures above show different solutions $a)$, $b)$ and $c)$ obtained with different C parameters.

Which solution belongs to $C = 0.01$, $C = 1$ and to $C = 100$? Briefly justify your answer.
Solution.
$C = 0.01$ belongs to $a)$, $C = 1$ to $b)$ and $C = 100$ to $c)$. With higher C we penalize the slack more and we have smaller margin.

4. The following image shows some training data $x_i \in R^2$, and $y_i \in \{-1, +1\}$. Circles represent the positive, and crosses represent the negative examples.

- Which training points are likely to be support vectors?
  Solution The most likely support vectors are the ones in the green boxes in the image. They are the closest to the other class, and they provide the widest margin.
- We add another positive training example to the training set. What happens to w and the margins, if we place the training example at location 1,2 or 3?
  Solution The training point in location 1 is inside the margin, so the inequality $y_1(w^T x_1 + b) \geq 1$ does not hold. The solution w and b changes, and x_1 will be a support vector. The point in location 2 is at the margin, $y_2(w^T x_2 + b) \geq 1$ holds, therefore w and b will remain the same. x_2 will might become a support vector, and the weights of support vectors might change. In case of location 3, the training point is outside the margin, so the solution will not change, and neither will the support vectors.

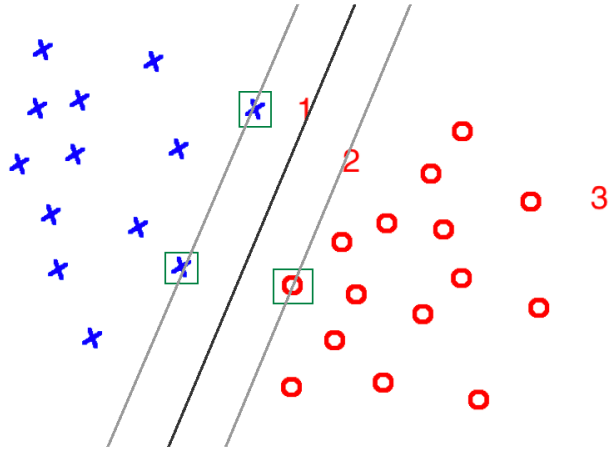


Figure 1: Train data.