Notes on SVM

Shasha Liao
Georgia Tech

October 25, 2020

1 Classification

- Decision boundary: $w^T x + b = 0$
- Positive class: (x, y) with $w^T x + b \geq c$ and $y > 1$.
- Negative class: (x, y) with $w^T x + b \leq -c$ and $y > 1$.
- Margin: $\gamma = \frac{2c}{\|w\|}$, the distance between two hyperplanes $w^T x + b = c$ and $w^T x + b = -c$.
- Correct classification condition: $(w^T x_i + b) y_i \geq c$ for all data points (x_i, y_i), $i = 1, 2, ..., m$.

2 Maximum margin classifier

Find decision boundary, i.e. w and b to maximize the margin.

$$\max_{w,b} \gamma = \frac{2c}{\|w\|} \text{ s.t. } (w^T x_i + b) y_i \geq c \text{ for all } i. \tag{1}$$

This becomes an optimization problem with constraints. And the solution is invariant under the scaling of c. So we set $c = 1$ and the problem (1) is equivalent to the following constrained convex quadratic programming problem:

$$\min_{w,b} \|w\| \text{ s.t. } (w^T x_i + b) y_i \geq 1 \text{ for all } i = 1, 2, ..., m. \tag{2}$$

- Only a few of the constrains are relevant → **support vectors**
- Kernel methods are introduced for nonlinear classification with nonlinear decision boundary.

To solve the optimization problem (2), we need to use Lagrangian multipliers.

2.1 Lagrangian Duality

The Lagrangian function:

$$L(w, b, \alpha) = \frac{1}{2} w^T w + \sum_{i=1}^{m} \alpha_i (1 - y^i (w^T x^i + b)). \tag{3}$$

The KKT conditions:

$$\frac{\partial L}{\partial w} = 0, \quad \frac{\partial L}{\partial b} = 0, \quad \alpha_i (1 - y^i (w^T x^i + b)) = 0, \alpha_i \geq 0, \forall i.$$

The last equation implies that all the points that are not on the margin will have the corresponding $\alpha_i = 0$. The first equation gives

$$w = \sum_{i=1}^{m} \alpha_i y^i x^i,$$

this tells us that w is a linear combination of the support vectors, i.e., the vectors on the margins.

From the KKT conditions we can solve for optimal w^* and b^*, plug in them into L we get the dual Lagrangian problem:

$$\begin{aligned} \max_{\alpha} L(w^*, b^*, \alpha) = \sum_{i}^{m} \alpha_i - \frac{1}{2} \sum_{i,j}^{m} \alpha_i \alpha_j y^i y^j ((x^i)^T x^j) && (4) \\ \text{s.t. } \alpha_i \geq 0, i = 1, 2, ..., m && (5) \\ \sum_{i}^{m} \alpha_i y^i = 0. && (6) \end{aligned}$$

This is a constrained quadratic programming. $L(w^*, b^*, \alpha)$ is nice and convex, and the global maximum can be found. For more details, see the slides. The optimal α_i can be found by solving the dual Lagrangian problem.

After that, we can find w, b in terms of α_i for i with x_i on the margins, i.e., x_i being support vector.

For a new test point z, compute

$$w^T z + b = \sum_{i \in \text{support vectors}} \alpha_i y^i ((x^i)^T z) + b.$$

If $w^T z + b > 0$, z is predicted to be in the positive class. Otherwise, z is predicted to be in the negative class.

3 Generalized SVM

- Kernelized SVM

  Since the Lagrangian and the decision rule only depends on the pairwise inner products of the points, we can replace all the inner products with nonlinear kernel functions to obtain nonlinear decision boundary.

- Soft-margin SVM

  When the data is not linearly separable, we allow a few points to violate the hard margin constraint and solve the following optimization problem:

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

  Question: can $\xi^i \geq 1$?