Lecture Note 14: SVM and SMO

Professor: Zhihua Zhang *Scribes: Han Xu, Yubo Xie*

1 SVM for Non-separable Case

1.1 Primal Problem

Recall from last lecture that we have learned a modified maximum margin idea that allows for mislabelled samples. If there exists no hyperplane that can split the "yes" and "no" samples, the *soft margin* method will choose a hyperplane that splits the samples as cleanly as possible, while still maximizing the distance to the nearest cleanly splitted samples. The method introduces non-negative slack variables ξ_i, which measures the degree of misclassification of the data $\mathbf{x}_i$. That is,

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0.$$

The optimization problem can be expressed as

$$\begin{aligned} \min\ & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i{}^p \\ \text{s.t.}\ & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0, \end{aligned}$$

where C is the tradeoff and we take $p = 1$. This problem is equivalent to

$$\min\ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b))_+,$$

where $(1-z)_+ = \max(z, 0)$, which is called hinge loss function. Dividing the objective function by C, the problem is further equivalent to

$$\min\ \sum_{i=1}^{n}(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b))_+ + \frac{1}{2C}\|\mathbf{w}\|^2.$$

For any supervised learning problem, we can generate a *regularization function* which is composed of a *loss function* by adding a *penalty function*, i.e., $L(x, y|\theta) + P(\theta)$. In non-separable SVM problems, if we use 0-1 loss function to represent whether a node is separated correctly or not, we need to enumerate every node in each step, which is NP-hard. Therefore, we use a continuous convex function $(1-z)_+$ to replace it.

1.2 Dual Problem

We have the following form for general optimization problems:

$$\begin{aligned} \inf \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \le 0, \text{ for } i \in I \\ & h_j(x) = 0, \text{ for } j \in J \\ & x \in C. \end{aligned}$$

For a convex optimization, it must be satisfied that
(1) $f(x)$ is a convex function of x.
(2) The constrain conditions must form a convex set.
For a linear optimization, it must be satisfied that all the functions f, g, and h are linear functions. If any of f, g, or h is non-linear, it is a non-linear optimization.

Considering the problem

$$\begin{aligned} \inf \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \le 0, \ i = 1, 2, \ldots, m \\ & x \in C. \end{aligned}$$

We have the KKT (Karush-Kuhn-Tucker) condition:

> **KKT Condition** Suppose the problem above has a local minimizer $\hat{x}$ in C. If the function f and g_i (for $i \in I(\hat{x})$) are differentiable at $\hat{x}$, and $\langle \nabla g_i(\hat{x}), d \rangle < 0$ for all indices i in the active set $I(\hat{x})$, then there is a Lagrange multiplier vector for $\hat{x}$. We call a vector $\boldsymbol{\lambda} \in \mathbb{R}^m_+$ a Lagrange multiplier vector for $\hat{x}$ if $\hat{x}$ is a critical point of the Lagrangian $L(x, \boldsymbol{\lambda}) = f(x) + \sum_{i=1}^m \lambda_i g_i(x)$ and complementary conditions hold: $\lambda_i = 0$ for $i \in I(\hat{x})$.

The active set above is defined as $I(\hat{x}) = \{i \mid g_i(\hat{x}) = 0\}$. Back to our SVM problem, we have the Lagrangian

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i [y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^n \beta_i \xi_i.$$

Take derivatives:

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i = \mathbf{0} \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \\ \frac{\partial L}{\partial b} &= -\sum_{i=1}^n \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^n \alpha_i y_i = 0 \\ \frac{\partial L}{\partial \xi_i} &= C - \alpha_i - \beta_i = 0 \quad \Rightarrow \quad \alpha_i + \beta_i = C \end{aligned}$$

And

$$\begin{aligned} \forall\, i \ \ \alpha_i [y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i] = 0 \quad &\Rightarrow \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) = 1 - \xi_i \ \vee \ \alpha_i = 0 \\ \forall\, i \ \ \beta_i \xi_i = 0 \quad &\Rightarrow \quad \beta_i = 0 \ \vee \ \xi_i = 0 \end{aligned}$$

Take them back, and we get the dual problem:

$$\max \ W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{s.t.} \ \ 0 \leq \alpha_i \leq C \ \wedge \ \sum_{i=1}^{n} \alpha_i y_i = 0.$$

We can replace $\mathbf{x}_i^T \mathbf{x}_j$ with kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. What's more, we can replace $y_i y_j \mathbf{x}_i^T \mathbf{x}_j$ with kernel function $K((\mathbf{x}_i, y_i), (\mathbf{x}_j, y_j))$. This problem is a *quadratic programming* (QP) problem, which can be solved by SMO.

2 Sequential Minimal Optimization (SMO)

The parameters of the maximum margin hyperplane are derived by solving the optimization. There exist several specialized algorithms for quickly solving the QP problem that arises from SVMs, mostly relying on heuristics for breaking the problem down into smaller, more manageable chunks. A common method is Platt's sequential minimal optimization (SMO) algorithm, which breaks the problem down into 2-dimensional sub-problems that may be solved analytically, eliminating the need for a numerical optimization algorithm.

Recall the coordinate ascent algorithm. To maximize $W(\alpha_1, \alpha_2, \ldots, \alpha_m)$, we take turns to estimate α_i. When estimating α_i, hold all other α_j's $(j \neq i)$ fixed. However, this is not suitable for our problem because if all other α_j's $(j \neq i)$ are fixed, $\alpha_i y_i = -\sum_{j \neq i} \alpha_j y_j$ is also fixed.

Therefore, we change two α_i's at a time. Take α_1 and α_2 for example:

$$\alpha_1 y_1 + \alpha_2 y_2 + \sum_{i=3}^{n} \alpha_i y_i = 0.$$

Letting $r = -\sum_{i=3}^{n} \alpha_i y_i$, we have

$$\alpha_1 y_1 + \alpha_2 y_2 = r,$$
$$\alpha_2 = y_2 r - \alpha_1 y_1 y_2.$$

Then the optimization function W becomes a quadratic function of α_1 with constrains $0 \leq \alpha_1 \leq C$ and $0 \leq y_2 r - \alpha_1 y_1 y_2 \leq C$ when α_i $(i = 3, \ldots, n)$ are fixed, which is much easier to estimate.