1 Support Vector Machines

Support vector machine is a kernelized optimal margin linear classifier [1]. For binary classification problem $\mathcal{Y} = \{-1, +1\}$, we are interested in finding a linear decision bondary, parameterized by $w \in \mathbb{R}^d, b \in \mathbb{R}$, that separates the training data points by maximizing the worst case distance (margin) of each data point to the decision boundary. We first assume that training set can be linearly separated. Given dataset $\{(x_i, y_i)\}_{i=1}^n$, we are interested in solving the following quadratic programming problem,

$$\begin{aligned} &\min_{w,b} \frac{1}{2} \|w\|^2 \\ &\text{s.t. } y_i(w^T x_i + b) \geq 1 \quad i = 1, 2, \cdots, n \end{aligned}$$

To derive the dual problem, we write the Lagrangian,

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^n \alpha_i \left[1 - y_i(w^T x_i + b)\right]$$

where $\alpha = \{\alpha_i\}_{i=1}^n$ are the dual variables. Solve for $\inf_{w,b} \mathcal{L}(w, b, \alpha)$ to arrive at the dual objective. In particular, first order optimality condition gives $w = \sum_{i=1}^n \alpha_i y_i x_i$ and it must be that $0 = \sum_{i=1}^n \alpha_i y_i$. Therefore,

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j \\ &\text{s.t. } \alpha_i \geq 0 \quad i = 1, 2, \cdots, n && \text{(dual feasibility)} \\ &\sum_{i=1}^n \alpha_i y_i = 0 && \text{(from } \nabla_b \mathcal{L} = 0\text{)} \end{aligned}$$

The dual can be solved more efficiently than the primal problem using coordinate descent. The decision rule is linear w.r.t support vectors (those x_i right on margin with $\alpha_i > 0$)

$$\hat{y}(x) = \text{sgn}\left(\sum_{i=1}^n \alpha_i y_i x_i^T x + b\right) \quad \text{for} \quad b = y_i - \sum_{j=1}^n \alpha_j y_j x_j^T x_i$$

for any support vector x_i. We observe that optimization as well as prediction uses input vectors via dot products only. We are motivated to use feature mapping $\phi : \mathcal{X} \to \mathcal{H}$ to map input vectors to a higher dimensional possibly infinite feature space in hope that the lifted space is linearly separable. The kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ allows us to compute dot products $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$ efficiently and represents a notion of similarity over two instance of aribtrary objects, e.g. vectors in $\mathbb{R}^n$, graphs, texts. We can substitute k whenever inner product is used and arrive at a optimal margin classifier over implicitly defined nonlinear feature mapping ϕ. In case when training dataset is not linearly separable, we can introduce slack variable $\{\xi_i\}_{i=1}^n$ where $x_i \geq 0$ to relax the inequality constraints and penalize misclassified or within margin points with $C \sum_{i=1}^n \xi_i$ for some $C \in \mathbb{R}$. In this case, we have the following Lagrangian,

$$\mathcal{L}(w, \xi, \alpha, \beta) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha \left(1 - y_i(\langle w, \phi(x_i) \rangle_{\mathcal{H}} + b) - \xi_i\right) + \sum_{i=1}^n \beta_i \xi_i$$

First order condition $0 = \frac{\partial}{\partial \xi_i}\mathcal{L} = C - \alpha_i + \beta_i$ together with dual feasibility $\beta_i \geq 0$ yield $\alpha_i \leq C$ for all $i = 1, 2, \cdots, n$. Therefore, we optimize for the following dual problem,

$$\begin{aligned} \min_{\alpha} \ & \frac{1}{2}\alpha^T Q \alpha - 1^T \alpha \\ \text{s.t. } & y^T \alpha = 0 \\ & 0 \leq \alpha_i \leq C \quad i = 1, 2, \cdots, n \end{aligned}$$

where $Q \in \mathbb{R}^{n \times n}$ with $Q_{ij} = \alpha_i \alpha_j k(x_i, x_j)$, with optimal decision rule as

$$\hat{y}(x) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b\right)$$

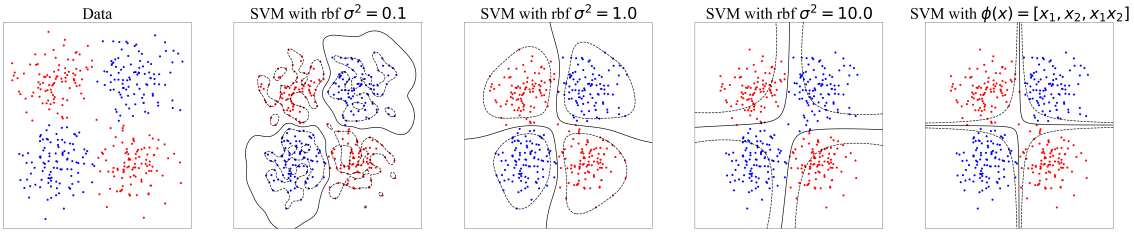

Figure 1: SVM on simulated 2D Gaussian with varying kernel

2 Kernel Method

References

[1] Bernhard Boser, Isabelle Guyon, and Vladimir Vapnik. “A Training Algorithm for Optimal Margin Classifiers”. In: (1992).