Support Vector Machine

Primal Form

- A hyperplane that separates a n-dimensional space into two half-spaces.
- Prediction rule: $y = sign(\omega^T x + b)$
- Margin
  - Geometric margin (≥ 0): $\gamma^{(i)} = y^{(i)}((\frac{\omega}{||\omega||})^T x^{(i)} + \frac{b}{||\omega||})$
  - Whole training set, the margin is $\gamma = \min\limits_i \gamma^{(i)}$
- Goal: Learn ω and b that achieves the maximum margin $\max\limits_{\omega,b} \min\limits_i \gamma^{(i)}$

$$\begin{aligned} &\max_{\gamma,\omega,b} \ \gamma \\ &s.t.\ y^{(i)}(\omega^T x^{(i)} + b) \geq \gamma ||\omega||, \quad \forall i \end{aligned}$$

- Scaling (ω, b) such that $\gamma||\omega|| = 1$, the problem becomes

$$\begin{aligned} &\max_{\omega,b} \ \frac{1}{||\omega||} \Leftrightarrow \min_{\omega,b} \ \omega^T \omega \Leftrightarrow \min_{\omega,b} \frac{1}{2}||\omega||^2 \\ &s.t.\ y^{(i)}(\omega^T x^{(i)} + b) \geq 1, \qquad \qquad \forall i \end{aligned}$$

- $\max\limits_{\omega,b} \ \frac{1}{||\omega||}$ is equivalent to $\min\limits_{\omega,b} \ \omega^T \omega$

  Def. The primal problem

$$\begin{aligned} &\min_{\omega,b} \frac{1}{2}||\omega||^2 \\ &s.t.\ y^{(i)}(\omega^T x^{(i)} + b) \geq 1, \quad \forall i \end{aligned}$$

Duality of SVM

- The Lagrangian problem for SVM

$$\min_{\omega,b,\alpha} \mathcal{L}(\omega, b, \alpha) = \frac{1}{2}||\omega||^2 + \sum_{i=1}^{m} \alpha_i (1 - y^{(i)}(\omega^T x^{(i)} + b))$$

- The Lagrangian dual problem for SVM is $\max_\alpha \mathcal{G}(\alpha) = \inf_{\omega,b} \mathcal{L}(\omega, b, \alpha)$

$$\begin{aligned} \max_{\alpha} \quad & \mathcal{G}(\alpha) = \textstyle\sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j \left(x^{(i)}\right)^T x^{(j)} \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^m \alpha_i y^{(i)} = 0 \\ & \alpha_i \geq 0 \quad \forall i \end{aligned}$$

- Proof.
  - $\frac{\partial}{\partial \omega} \mathcal{L}(\omega, b, \alpha) = \omega - \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} = 0$ and $\frac{\partial}{\partial b} \mathcal{L}(\omega, b, \alpha) = \sum_{i=1}^m \alpha_i y^{(i)} = 0$
  - $\mathcal{L}$ is a convex function.

- It suffices **Slarter's Condition**. Thus, the problem can be solved by QP solver (MATLAB, …)
- Since we have the solution α^* for the dual problem, we can calculate the solution for the primal problem.

$$\omega^* = \sum_{i=1}^{m} \alpha^* y^{(i)} x^{(i)} b^* = y^{(i)} - \omega^{*T} x^{(i)}, \text{ if } \alpha^* > 0$$

- For robustness, the optimal value for b is calculated by taking the averages across all b^*

$$b^* = \frac{\sum_{i:\alpha_i^*>0} \left(y^{(i)} - \omega^{*T} x^{(i)}\right)}{\sum_{i=1}^{m} \mathbf{1}\left(\alpha_i^* > 0\right)}$$

- However, according to **Complementary Slackness**, $\alpha_i^* \left[1 - y^{(i)}\left(\omega^{*T} x^{(i)} + b^*\right)\right] = 0$.
- α_i^* is non-zero only if $x^{(i)}$ lies on the margin, i.e., $y^{(i)}\left(\omega^{*T} x^{(i)} + b^*\right) = 1$. (**Support Vector**, $\mathcal{S}$).

$$\therefore \omega = \sum_{s \in \mathcal{S}} \alpha_s y^{(s)} x^{(s)}$$

Kernel

- Basic idea: mapping data to higher dimensions where it exhibits linear patterns.
- Each kernel K has an associated feature mapping $\phi : \mathcal{X} \to \mathcal{F}$ from input to feature space.
  - e.g., quadratic mapping $\phi : x \to \{x_1^2, x_2^2, \cdots, x_1 x_2, \cdots, x_1 x_n, \cdots, x_{n-1} x_n\}$
- Kernel $K(x,z) = \phi(x)^T \phi(z)$, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ **takes two inputs and gives their similarity** in $\mathcal{F}$.

  Thereom. **Mercer's Condition**.

  For K to be a kernel function if K is a positive definite function.

$$\int \int f(x) K(x,z) f(z) dx dz > 0 \forall f, \ s.t. \ \int_{-\infty}^{\infty} f^2(x) dx < \infty$$

- Composing rules
  - Direct sum $K(x,z) = K_1(x,z) + K_2(x,z)$
  - Scalar product $K(x,z) = \alpha K_1(x,z)$
  - Direct product $K(x,z) = K_1(x,z) K_2(x,z)$

  Def. Kernel Matrix.

$$K_{i,j} = K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)})$$

Example Kernel

- Linear (trivial) Kernal $K(x,z) = x^T z$
- Quadratic Kernel $K(x,z) = \left(x^T z\right)^2$ or $\left(1 + x^T z\right)^2$
- Polynomial Kernel (of degree d) $K(x,z) = \left(x^T z\right)^d$ or $\left(1 + x^T z\right)^d$
- Gaussian Kernel $K(x,z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$
- Sigmoid Kernel $K(x,z) = \tanh\left(\alpha x^T + c\right)$

Applicable Algorithm

- SVM, linear regression, etc.
- K-means, PCA, etc.
- For SVM

$$\begin{array}{ll} \max_\alpha & \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i,j=1}^m y^{(i)}y^{(j)}\alpha_i\alpha_j\left(x^{(i)}\right)^T x^{(j)} \\ \text{s.t.} & \sum_{i=1}^m \alpha_i y^{(i)} = 0 \\ & \alpha_i \geq 0 \quad \forall i \end{array} \Rightarrow \begin{array}{ll} \max_\alpha & \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i,j=1}^m y^{(i)}y^{(j)}\alpha_i\alpha_j K_{i,j} \\ \text{s.t.} & \sum_{i=1}^m \alpha_i y^{(i)} = 0 \\ & \alpha_i \geq 0 \quad \forall i \end{array}$$

Soft Margin

- Relax the constraints from $y^{(i)}(\omega^T x^{(i)} + b) \geq 1$ to $y^{(i)}(\omega^T x^{(i)} + b) \geq 1 - \xi_i$
- $\xi_i \geq 0$ is called slack variable

  Def. **Soft Margin SVM**

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

- C is a hyper-parameter that controls the relative weighting between $\frac{1}{2}||\omega||^2$ for **larger margins** and $\sum_{i=1}^m \xi_i$ for **fewer misclassified examples**.
- Lagrangian function

$$\mathcal{L}(\omega, b, \xi, \alpha, r) = \frac{1}{2}\omega^T\omega + C\sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i\left[y^{(i)}\left(\omega^T x^{(i)} + b\right) - 1 + \xi_i\right] - \sum_{i=1}^m r_i\xi_i$$

- KKT conditions (the optimal values of ω, b, ξ, α, and r should satisfy the following conditions)
  - $\nabla_\omega \mathcal{L}(\omega, b, \xi, \alpha, r) = 0 \Rightarrow \omega^* = \sum_{i=1}^m \alpha_i^* y^{(i)} x^{(i)}$
  - $\nabla_b \mathcal{L}(\omega, b, \xi, \alpha, r) = 0 \Rightarrow \sum_{i=1}^m \alpha_i^* y^{(i)} = 0$
  - $\nabla_{\xi_i} \mathcal{L}(\omega, b, \xi, \alpha, r) = 0 \Rightarrow \alpha_i^* + r_i^* = C$, for $\forall i$

- $\alpha_i^*, r_i^*, \xi_i^* \geq 0$, for $\forall i$
- $y^{(i)} \left(\omega^{*T} x^{(i)} + b^* \right) + \xi_i^* - 1 \geq 0$, for $\forall i$
- $\alpha_i^* \left(y^{(i)} \left(\omega^* x^{(i)} + b^* \right) + \xi_i^* - 1 \right) = 0$, for $\forall i$
- $r_i^* \xi_i^* = 0$, for $\forall i$

- Dual problem

$$
\begin{array}{ll}
\max_\alpha & \mathcal{J}(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} > \\
\text{s.t.} & 0 \leq \alpha_i \leq C, \quad \forall i = 1, \cdots, m \\
& \sum_{i=1}^m \alpha_i y^{(i)} = 0
\end{array}
$$

- Solution
  - $\omega^* = \sum_{i=1}^m \alpha_i^* y^{(i)} x^{(i)}$
  - $b^* = \frac{\sum_{i:0<\alpha_i^*<C} (y^{(i)} - \omega^{*T} x^{(i)})}{\sum_{i=1}^m 1(0<\alpha_i^*<C)}$
- Proof.

$$
\begin{array}{l}
\because r_i^* \xi_i^* = 0 \Leftrightarrow (C - \alpha_i^*) \xi_i^* = 0 \\
\therefore \forall i, \alpha_i^* \neq C \Rightarrow \xi_i = 0 \Rightarrow \alpha_i (y^{(i)} (\omega^{*T} x^{(i)} + b^*) - 1) = 0 \\
\therefore \forall i, \alpha_i^* \in (0, C) \Rightarrow y^{(i)} (\omega^{*T} x^{(i)} + b^*) = 1 \Rightarrow \omega^{*T} x^{(i)} + b^* = y^{(i)}
\end{array}
$$

- Corollaries of KKT conditions for soft-margin SVM
  - When $\alpha_i^* = 0, y^{(i)} \left(\omega^{*T} x^{(i)} + b^* \right) \geq 1$, correctly classified.
  - When $\alpha_i^* = C, y^{(i)} \left(\omega^{*T} x^{(i)} + b^* \right) \leq 1$, misclassified.
  - When $0 < \alpha_i^* < C, y^{(i)} \left(\omega^{*T} x^{(i)} + b^* \right) = 1$, support vector.