Support Vector Machines

EE219: Large Scale Data Mining

Professor Roychowdhury

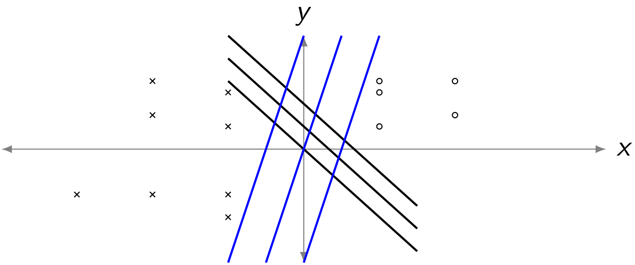
Jan 25, 2017

Summary

- Review
  - SVM basics
  - Calculate the margin
- Hard-margin SVM
  - Dual problem and optimal solution
- Soft-margin SVM
  - Hinge loss
  - Dual problem and optimal solution
- Nonlinear
  - Lifting a vector
  - Gram matrix
  - Kernel

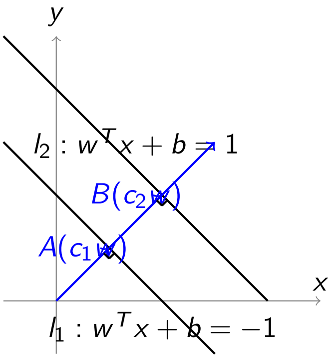
Review SVM: basics

Support Vector Machine is a supervised learning model trained for classification or regression tasks. When it is a binary classifier, it is trained to find a hyperplane such that the distance from it to the nearest datapoint on both side is maximized.



- distance between $w^T x - b = 1$ and $w^T x - b = -1$ is $\frac{2}{w^T w}$
- maximize margin means minimize $\frac{1}{2} w^T w$
- when the slack variable is considered, the objective function to minimize will be $\frac{1}{2} w^T w + \lambda \sum_{i=1}^{n} \epsilon_i$

Review SVM : calculate the margin



- Point A on l_1 and Point B on l_2 satisfy:

$$w^T(c_1 w) + b = -1 \quad (1)$$
$$w^T(c_2 w) + b = 1 \quad (2)$$

- The distance $D(A, B)$ between point A,B is also the distance between line l_1, l_2:

$$\begin{aligned} D(A,B) &= \|c_2 w - c_1 w\|_2 \\ &= (c_2 - c_1)\|w\|_2 \\ &\overset{①}{=} \frac{2}{w^T w}\|w\|_2 \\ &= \frac{2}{\|w\|_2} \end{aligned}$$

- (1) - (2) to get ①

Hard-margin SVM: Dual problem

As stated in previous lecture, for the binary classification problem, when N samples are linear separable, it can be written as N constraints in an optimization problem.

$$y_i = \begin{cases} 1 & \text{if } x_i \in C_1 \\ -1 & \text{if } x_i \in C_2 \end{cases}$$

For max margin classifier, it can be transformed into a minimization problem with cost function: $\frac{1}{2}w^T w$. Then the whole problem can be solved through dual problem.

Primal problem

minimize: $\frac{1}{2}w^T w$

s.t. $y_i(w^T x_i + b) \geq 1$, $i = 1, \ldots, N$

Dual problem

maximize: $-\frac{1}{2}\alpha^T Q\alpha + 1^T\alpha$

s.t. $\alpha \geq 0$ and $y^T\alpha = 0$

Hard-margin SVM: maxizing the margin

- the Lagrange function for the primal problem can be written as $L(w,b,\alpha) = \frac{1}{2}w^T w + \sum_{i=1}^{N} \alpha_i(1 - y_i(w^T x_i + b))$
- $\alpha \in \mathrm{R}^N$ is the Lagrange multiplier($\alpha_i \geq 0$), we hope to $\underset{w,b}{\text{minimize}}\,\underset{\alpha}{\text{maximize}}\; L(w,b,\alpha)$, the optimal value is equal to that in $\underset{\alpha}{\text{maximize}}\,\underset{w,b}{\text{minimize}}\; L(w,b,\alpha)$ when it satisfies Slater's condition, which means strictly feasible in this problem.
- $\frac{\partial L}{\partial w} = 0$, then $w = \sum_{i=1}^{N} \alpha_i y_i x_i$. $\frac{\partial L}{\partial b} = 0$, then $\sum_{i=1}^{N} \alpha_i y_i = 0$.
- substitute w into $L(w,b,\alpha)$, we will get

$$
\begin{aligned}
L(w,b,\alpha) &= \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i y_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j {x_i}^T x_j \\
&= \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j {x_i}^T x_j
\end{aligned}
$$

Hard-margin SVM: dual problem for maxizing the margin

Let $Q_{ij} = y_i y_j {x_i}^T x_j$, then $L(w, b, \alpha) = 1^T\alpha - \frac{1}{2}\alpha^T Q\alpha$. So the dual problem can be formulated as

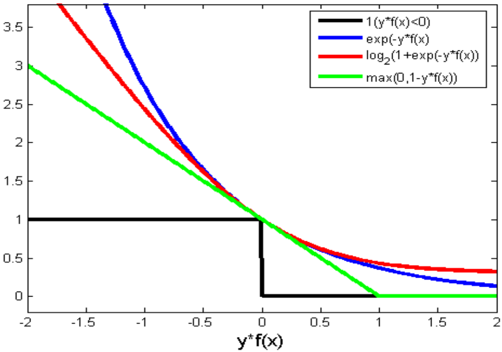
$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & 1^T\alpha - \frac{1}{2}\alpha^T Q\alpha \\ \text{subject to} \quad & \alpha_i \geq 0,\ i = 1, \ldots, N \\ & y^T\alpha = 0 \end{aligned}$$

Hard-margin SVM: optimal solution

- When w, b is the optimal solution for the primal problem, complementary slackness condition is satisfied: $\alpha_i(1 - y_i(w^T x_i + b)) = 0$, $i = 1, \ldots, N$.
- **Complementary slackness condition** can be satisfied in two ways:
  - $\alpha_i = 0$
  - $y_i(w^T x_i + b) = 1$
- Vectors x_i for which $y_i(w^T x_i + b) = 1$ are called **support vectors**. Support vectors lie on the margin. For each x_i, there is a corresponding $\alpha_i > 0$, let it be $\alpha_i^*(i = 1, \cdots, n)$.
- $w^* = \sum_{i=1}^{n} \alpha_i y_i x_i = \sum_{i=1}^{N} \alpha_i^* y_i x_i$
- $b^* = y_j - {w^*}^T x_j = y_j - \sum_{i=1}^{N} y_i \alpha_i^* x_i^T x_j$
- given a new $x \in \mathrm{R}^n$,we classify it based on decision function:c(x) $= sgn({w^*}^T x + b^*) = sgn(\sum_{i=1}^{N} \alpha_i^* y_i x_i^T x + b^*)$

Soft-margin SVM: Hinge Loss

SVM are extended with hinge loss to handle cases where training data are not linearly seperable. For a training data's label $y = \pm 1$, the hinge loss of the prediction $f(x)$ is defined as $\max(0, 1 - yf(x))$. It is also called soft-margin SVM since it allows some datapoints on the wrong size with penalty $1 - yf(x)$ in the objective function. Note here the penalty for datapoint on the right side is still 0.



Soft-margin SVM: dual problem with slack variables

Primal problem

minimize: $\frac{1}{2}w^T w + \gamma \sum_{i=1}^{N} \epsilon_i$

$$s.t.\ y_i(w^T x_i + b) \geq 1 - \epsilon_i,\ i = 1, \ldots, N$$
$$\epsilon_i \geq 0,\ i = 1, \ldots, N$$

Dual problem

maximize: $-\frac{1}{2}\alpha^T Q\alpha + 1^T\alpha$

$$s.t.\ 0 \leq \alpha \leq \gamma\mathbf{1},$$
$$y^T\alpha = 0$$

- Similarly, the Lagrange function for the primal problem can be written as $L(w, b, \alpha, \lambda) =$
  $\frac{1}{2}w^T w + \sum_{i=1}^{N} \alpha_i(1 - y_i(w^T x_i + b)) + \gamma 1^T\epsilon - \sum_{i=1}^{N} \lambda_i \epsilon_i$
- $\frac{\partial L}{\partial w} = 0$, then $w = \sum_{i=1}^{N} \alpha_i y_i x_i$. $\frac{\partial L}{\partial b} = 0$, then $\sum_{i=1}^{N} \alpha_i y_i = 0$
- for $\epsilon_i \geq 0$, $\frac{\partial L}{\partial \epsilon} = 0$, then $\gamma - \alpha_i - \lambda_i = 0$ and since $\lambda_i \geq 0$, it can be simplified as $\gamma - \alpha_i \geq 0$ to remove variable λ_i.

Nonlinear –lifting a vector

- It's important to use nonlinear classifier because sometimes the data are not linearly separable.
- There are several ways to lift a vector, for example, through polynomial or exponential transformation of the original vector.
- $x_i \in \mathbb{R}^n \rightarrow \phi(x_i) \in \mathbb{R}^m (m > n)$
  - For example, in polynomial transformation
    - $x = [x_1\, x_2\, ..\, x_n]^T, \phi(x) = [x_1\, x_2\, ..x_1 x_2\, ..\, x_{n-1} x_n]^T$, here the dimension m of the new feature will equal $n + \binom{n}{2})$).
    - The decision function $c(x)$ can be written as $sgn(w^T \phi(x) + b)$.

Gram matrix and kernel

- Q is called Gram matrix
- In the linear case, $Q_{ij} = y_i y_j {x_i}^T x_j$
- After lifting the vector, $Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j)$
- Decision function:

$$\begin{aligned} c(x) &= sgn({w^*}^T \phi(x) + b) \\ &= sgn(\sum_{i=1}^{n} \alpha_i^* y_i \phi(x_i)^T \phi(x) + b^*) \end{aligned}$$

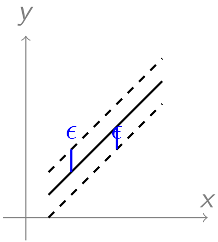
- Let $k_{ij} = \phi(x_i)^T \phi(x_j)$,then $k : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ is called kernel.
  - For example, Gaussian Kernel: $k_{ij} = \exp(-\beta \|x_i - x_j\|^2)$, then c(x) $= sgn(\sum_{i=1}^{n} \alpha_i^* y_i \exp(-\beta \|x_i - x\|^2) + b^*)$
  - Gaussian kernel is widely used and you can choose different kernels. Kernel method is computationally efficient.

SVM regression

- SVM regression uses ϵ-insensitive loss function proposed by Vapnik(1995):

$$L_\epsilon(y, f(x,w)) = \begin{cases} 0 & if\,|y - f(x,w)| \le \epsilon \\ |y - f(x,w)| - \epsilon & otherwise. \end{cases}$$

- Training samples inside ϵ region will have no penalty and deviation of training samples outside the region will be measured by slack variables ϵ_i, ϵ_i^*
- For N observations $(y_1, x_1) \cdots (y_N, x_N)$, the optimization problem can be written as:



$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} w^T w + \gamma \sum_{i=1}^{N} (\xi_i + \xi_i^*) \\ \text{subject to} \quad & y_i - (w^T x_i + b) \le \epsilon + \xi_i \\ & (w^T x_i + b) - y_i \le \epsilon + \xi_i^* \\ & \xi_i^* \ge 0, \xi_i \ge 0 \end{aligned}$$