Linear Regression

EE219: Large Scale Data Mining

Professor Roychowdhury

Jan 25, 2017

Summary

- Review
  - SVM $y_i = \sum_{j=1}^{d} a_j * x_i(j) + \epsilon_i = x_i^T \theta + \epsilon_i$
  - max margin
- Dual problem and optimal solution
- Nonlinear
  - lifting a vector
  - Gram matrix
  - kernel
- Hinge loss
- Gradient descent

Review SVM :max margin

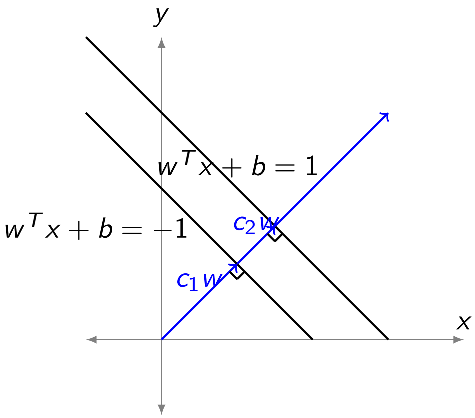


Figure 1: max margin calculation

- for $c_1 w, c_2 w$ on the lines

$$\begin{cases} w^T(c_1 w) + b = 1 \\ w^T(c_2 w) + b = -1 \end{cases}$$

- distance $= (c_2 - c_1)\left\|w\right\|_2 = \frac{2}{w^T w}\left\|w\right\|_2 = \frac{2}{\left\|w\right\|_2}$

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 = 1,2,..n

Dual problem

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

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

Dual problem

- 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

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

Dual problem

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

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

Dual problem,optimal solution

- When w,b is the optimal solution for the primal problem, complementary slackness condition is satisfied: $\alpha_i(1 - y_i(w^T x_i + b)) = 0$ for $i = 1..n$.
- 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 $\alpha_i^*(i = 1..N)$.
- $w^* = \sum_{i=1}^{n} \alpha_i y_i x_i = \sum_{i=1}^{N} \alpha_i^* y_i x_i$
- $b^* = y_j - {w^*}^T x_j = y_j - \sum_{i=1}^{N} y_i \alpha_i^* x_i^T x_j$
- given a new $x \in \mathrm{R}^n$,we classify it based on decision function:c(x) $= sgn({w^*}^T x + b^*) = sgn(\sum_{i=1}^{N} \alpha_i^* y_i x_i^T x + b^*)$

Dual problem – with slack variable

Primal problem

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

s.t. $y_i(w^T x_i + b) \geq 1$,i $=$ 1,2,..n

$\epsilon_i \geq 0$,i $=$ 1,2,..n

Dual problem

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

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

- 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

Nonlinear –lifting a vector

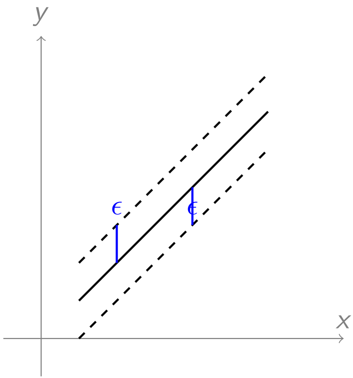
- It's important to use nonlinear classifier because sometimes the data are not linearly separable.
- There are several ways to lift a vector, for example, through polynomial or exponential transformation of the original vector.
- $x_i \in \mathbb{R}^n \rightarrow \phi(x_i) \in \mathbb{R}^m (m > n)$
  - For example, in polynomial transformation, x = $[x_1\, x_2\, ..\, x_n]^T, \phi(x) = [x_1\, x_2\, ..x_1 x_2\, ..\, x_{n-1} x_n]^T, (m = n + \binom{n}{2})$, then the decision function c(x) = 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 c(x) = $sgn({w^*}^T \phi(x) + b) = sgn(\sum_{i=1}^{n} \alpha_i^* y_i \phi(x_i)^T \phi(x) + b^*)$
- 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: $k_{ij} = \exp(-\beta \left\| x_i - x_j \right\|^2)$, then c(x) = $sgn(\sum_{i=1}^{n} \alpha_i^* y_i \exp(-\beta \left\| x_i - x \right\|^2) + b^*)$
    - Gaussian kernel is widely used and you can choose different kernel. Kernel method is computationally efficient.

SVM regression

For $(y_1, x_1), (y_2, x_2), ..(y_n, x_n)$ n observations. The optimization problem can be written as:



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

Gradient descent

- Regularized least squared error:
  $$E(\theta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f_\theta(x_i))^2 + \lambda\theta^T\theta$$
- $\theta^* = \underset{\theta}{\text{argmin}}E(\theta)$. How to learn θ?
- $\theta_{i+1} = \theta_i - \eta(\frac{\partial E}{\partial \theta})$
- step size: η, generally the smaller the step size is, the longer it will take to get optimal choice of θ

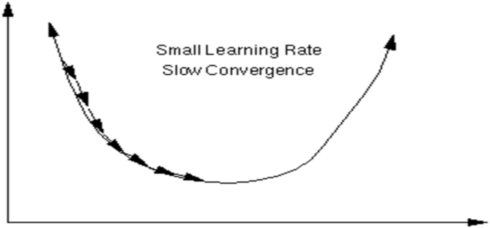


Figure 2: Gradient descent