ML Fundamentals: Generative Learning Algorithms

February 2, 2021

1 Support Vector Machines

In the below figure 1, we can see a decision boundary $\theta^T x = 0$ as the separating decision boundary. Logistic regression, $h_\theta(x) = g\left(\theta^T x\right)$ would assign higher probability, if the point is farther from the decision boundary for the respective label. For point A, $\theta^T x \gg 0$, so the probability $y = 1$ would be very high.

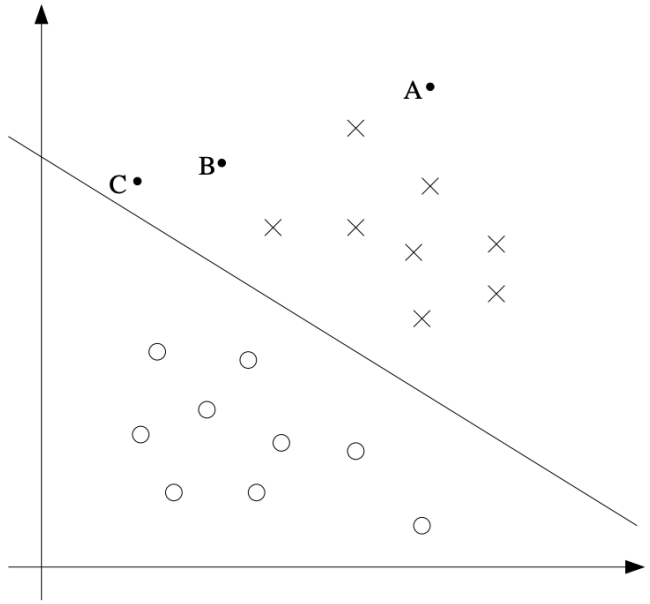

Figure 1: Illustration of separating decision boundary

Now for ease of formulating SVM, lets change formulation,

1. $h_{w,b}(x) = g\left(w^T x + b\right)$ instead of $h = g\left(\theta T x\right)$
2. $g(z) = 1$ if $z \geq 0$, and $g(z) = -1$ otherwise.
3. Functional margin: $\hat{\gamma}^{(i)} = y^{(i)}\left(w^T x + b\right)$. It can be scaled higher if we scale w and b.
4. Geometric margin: Its the distance of a point from decision boundary, since its a distance it has to be in the direction of unit normal vector $w/||w||$. Its projection on decision boundary is $x^{(i)} - \gamma^{(i)} \cdot w/\|w\|$. And, it can be substituted to $w^T x + b = 0$, so the equation would be,

$$w^T\left(x^{(i)} - \gamma^{(i)}\frac{w}{\|w\|}\right) + b = 0$$

Solving for geometric margin, $\gamma^{(i)}$ yields

$$\gamma^{(i)} = y^{(i)}\left(\left(\frac{w}{\|w\|}\right)^T x^{(i)} + \frac{b}{\|w\|}\right)$$

Hence, $\gamma^{(i)} = y^{(i)}\left(\left(\frac{w}{\|w\|}\right)^T x^{(i)} + \frac{b}{\|w\|}\right)$

5. Minimum functional margin and geometric margin can be computed as, $\hat{\gamma} = \min_{i=1,\ldots,m} \hat{\gamma}^{(i)}$, and, $\gamma =$

$\min_{i=1,\ldots,m} \gamma^{(i)}$. We can note that γ and $\hat{\gamma}$ are equal when $||w|| = 1$
6. We can also conclude, $\gamma = \hat{\gamma}/||w||$
7. Now we need to maximize geometric margin, γ, which is the minimum distance from decision boundary. Hence,

$$\begin{aligned} &\max_{\gamma,w,b} \gamma \\ &\text{s.t. } y^{(i)} \left(\frac{w}{\|w\|} x^{(i)} + \frac{b}{\|w\|} \right) \geq \gamma, \quad i = 1, \ldots, m \end{aligned}$$

When $||w|| = 1$, geometric mean would be equal to functional mean $\hat{\lambda}$. Hence

$$\begin{aligned} &\max_{\gamma,w,b} \gamma \\ &\text{s.t. } y^{(i)} \left(w^T x^{(i)} + b \right) \geq \gamma, \quad i = 1, \ldots, m \\ &\quad \|w\| = 1 \end{aligned}$$

We can get rid of the constraint $||w|| = 1$, by replacing functional margin instead of geometric margin in objective function and constraint. Hence,

$$\begin{aligned} &\max_{\gamma,w,b} \frac{\hat{\gamma}}{\|w\|} \\ &\text{s.t. } y^{(i)} \left(w^T x^{(i)} + b \right) \geq \hat{\gamma}, \quad i = 1, \ldots, m \end{aligned}$$

Lets set functional margin $\hat{\lambda} = 1$, as weights w and b would scale accordingly. Hence,

$$\begin{aligned} &\min_{\gamma,w,b} \frac{1}{2} \|w\|^2 \\ &\text{s.t. } y^{(i)} \left(w^T x^{(i)} + b \right) \geq 1, \quad i = 1, \ldots, m \end{aligned}$$

1.1 Lagrange Duality

Lagrange establish relationship between primal and dual problem.
A linear primal optimisation problem, objective function is linear combination of n variables, and there are m constraints, each of which put upper bound on a linear combination of n problem.
In linear dual problem, lower bound linear constraints are applied over primal objective.
Consider the following primal optimization problem,

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \ldots, k \\ & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

Lets define generalised Lagrangian form to solve it,

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Which can be written as the dual problem as follow,

$$\min_w \theta_{\mathcal{P}}(w) = \min_w \max_{\alpha,\beta:\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

Where maximize over α, β is a primal problem, which would ensure that constraints are satisfied, and minimization over primal problem will give the optimal solution p^*
We can now pose a dual optimization problem, let say solution to this dual problem is d^*

$$\max_{\alpha,\beta:\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha,\beta:\alpha_i \geq 0} \min_w \mathcal{L}(w, \alpha, \beta)$$

Now $p^* == d^*$ under Karush-Kuhn-Tucker conditions (KKT condition).
Where the few constraints which takes value zero, i.e. $g_i(w) = 0$ (among $g_i(w) <= 0$) would allow α_i to choose non-zero value, and these are the only active support vectors, which are at the geometric margin distance away from the decision boundary.

1.2 Optimal Margin Classifier

We had following primal optimization problem to solve for SVM (optimal margin classifier case).

$$\begin{aligned}
&\min_{\gamma,w,b} \frac{1}{2}\|w\|^2 \\
&\text{s.t. } y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1, \quad i = 1, \ldots, m
\end{aligned}$$

Where, we can write constraints as follow to formulate it as Lagrangian.

$$g_i(w) = -y^{(i)}\left(w^T x^{(i)} + b\right) + 1 \leq 0$$

Hence Lagrangian for our optimisation problem would be,

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

And, it can be solved as a dual problem

$$p^* = d^* = \max_{\alpha:\alpha_i \geq 0} \min_{w,b} \mathcal{L}(w, b, \alpha)$$

Lets minimize $\mathcal{L}(w, b, \alpha)$ to get solution θ_D, i.e. the value of w and b. Lets take derivative against w and b, and set it to zero.

$$\nabla_w \mathcal{L}(w, b, \alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0$$

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

Similarly, derivative with respect to b, would give,

$$\frac{\partial}{\partial b}\mathcal{L}(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

By substituting value of w when partial derivative w.r.t. w is zero, and equation when partial derivative w.r.t. b is zero, the dual max min problem becomes

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

1.3 Prediction

Predicting the label for input x would require to solve following equation

$$\begin{aligned}
w^T x + b &= \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}\right)^T x + b \\
&= \sum_{i=1}^{m} \alpha_i y^{(i)} \left\langle x^{(i)}, x \right\rangle + b
\end{aligned}$$

Note that, α_i would be non-zero only for support vectors.

1.4 Kernels

We can map x into features like $[x, x^2, x^3]$ or other such projections. Lets call this feature mapping as $\phi(x)$. Kernel directly gives the dot product of feature mappings, without explicitly deriving feature mappings for inputs. So,

$$K(x, y) = \phi(x)^T \phi(y)$$

1. If kernel is $K(x, y) = (x^T z)^2$, then the feature mapping is

$$\phi(x) = [x_1x_1, x_1x_2, x_1x_3, x_2x_1, x_2x_2, x_2x_3, x_3x_1, x_3x_2, x_3x_3]^T$$

So, computing feature mappings and then applying dot product would be $o(n^2)$ operation, whereas kernels directly gives $o(n)$ operation, where n is the dimension of input vector x.

2. For Gaussian kernel, $K(x, z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$, the feature mapping $\phi(x)$ would be infinite features.

For samples $(x_1, x_2, ..., x_m)$, An m x m matrix is kernel matrix if it is symmetric and positive definite.

1.5 Regularization and the non-separable case

Data is not always linearly separable (likelihood of separating data linearly is much higher by projecting it in higher dimension), and also hard constraint of separating can significantly alter decision boundary when there are few outliers.

To handle this, we will allow few samples are permitted to have functional margin less than 1, of-course at a cost C. So, objective function would look like,

$$\begin{aligned}
&\min_{\gamma,w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \\
&\text{s.t. } y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\
&\qquad \xi_i \geq 0, \quad i = 1, \ldots, m
\end{aligned}$$

And, corresponding Lagrangian form would be as below:

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

Solving this as a dual problem would give following equation:

$$\begin{aligned}
&\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \left\langle x^{(i)}, x^{(j)} \right\rangle \\
&\text{s.t. } 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m \\
&\qquad \sum_{i=1}^{m} \alpha_i y^{(i)} = 0
\end{aligned}$$

Note that, somewhat surprisingly, in adding (1 regularization, the only change to the dual problem is that what was originally a constraint that $0 \leq \alpha_i$ has now become $0 \leq \alpha_i \leq C$