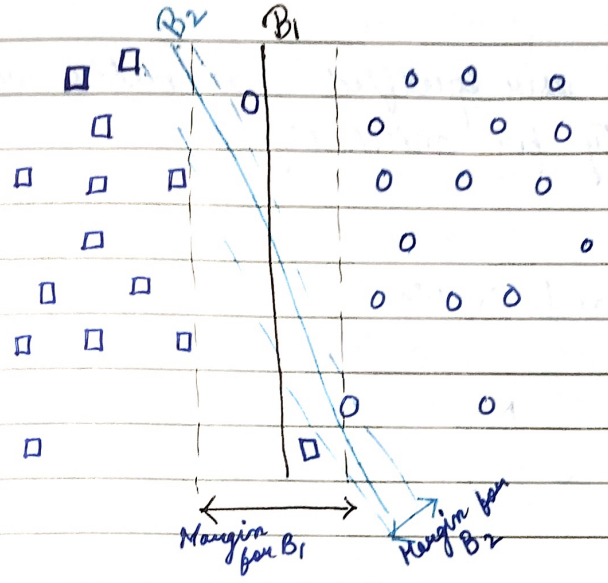
DEC 8

DECISION BOUNDARY OF SVM FOR NONSEPARABLE CASE



In above figure we see that the classifier B_2 successfully classifies all the sample but the decision boundary is small. But the other classifier B_1 makes a slight error while classification but it has a greater decision ~~boundary~~ ~~and~~ margin.

So we add a slack variable to tolerate this small error. Our inequality constraint becomes.

$$w.x_i + b \geq 1 - \xi_i \quad \text{if } y_i = 1$$

$$w.x_i + b \leq -1 + \xi_i \quad \text{if } y_i = -1$$

where $\xi_i > 0$ and ξ_i is the slack variable

Objective function

$$f(w) = \frac{\|w\|^2}{2} + C\left(\sum_{i=1}^{N} \xi_i\right)^k$$

where C ~~and k~~ is an ~~are~~ user specified parameter which represents the penalty for misclassification

Lagrangian dual problem derivation

$$L_p = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\lambda_i\left\{y_i(wx_i+b) - 1 + \xi_i\right\} - \sum_{i=1}^{N}\mu_i\xi_i \quad \text{---(1)}$$

where first two terms are the objective function to be minimized, the third term represents the inequality constraints associated with the slack variable, and the last term is the result of the non-negativity requirements on the values of ξ_i's.

KKT condition

$$\xi_i \geq 0 \;,\; \lambda_i \geq 0,\; \mu_i \geq 0.$$

$$\lambda_i\left\{y_i(w \cdot x_i + b) - 1 + \xi_i\right\} = 0$$

$$\mu_i \xi_i = 0$$

First order derivation of L w.r.t w, b, ξ_i to zero.

$$\frac{\partial L}{\partial w_j} = w_j - \sum_{i=1}^{N} \lambda_i y_i x_{ij} = 0$$

$$\Rightarrow w_j = \sum_{i=1}^{N} \lambda_i y_i x_{ij} \quad \text{— (2)}$$

$$\frac{\partial L}{\partial \xi_i} = C - \lambda_i - \mu_i = 0 \Rightarrow \lambda_i + \mu_i = C \quad \text{— (3)}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N} \lambda_i y_i = 0 \Rightarrow \sum_{i=1}^{N} \lambda_i y_i = 0 \quad \text{— (4)}$$

Putting (5), (4), (3) in (1)

$$L_D = \frac{1}{2} \sum_{ij} \lambda_i \lambda_j y_i y_j x_i \cdot x_j + C \sum_i \xi_i$$

$$- \sum_i \lambda_i \left\{ y_i \left(\sum_j \lambda_j y_j x_i \cdot x_j + b \right) - 1 + \xi_i \right\}$$

$$- \sum_i (C - \lambda) \xi_i$$

$$= \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{ij} \lambda_i \lambda_j y_i y_j x_i \cdot x_j$$

SAMBO DUTTA
UBID:50318667

Pointing out the margins in primal and dual problem:

In the derivation here we have taken a ***soft margin*** since we are tolerating a small error to maximise the margins as oppose to a ***hard margin*** where we tolerate zero error.

In primal problem:

The width of the margins is represented by **2/||W||** which we are trying to maximise in the model.
The objective function $\frac{1}{2}W^2 + C\sum \zeta_i$. Here C is penalty we are incurring where we make kind of a trade off between maximising the margins or reducing the misclassified points.
If we set the value C to be very high, we would then tolerate no misclassification, and on setting the value of C to be very low, we get a wider margin.

In dual problem:

We need to maximise the margin ($\sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i\lambda_j y_i y_j x_i x_j$) with respect to λ which has been derived in the previous section, subject to the constraints,
$\sum_{i=1}^{N} \lambda_i y_i = 0$, $\lambda_i >= 0$.

The Benefits for **maximising the margins** are:

I. A larger margin will help in generalising the model when we use sample outside the training data.
II. It prevents overfitting of the model as discussed in the bias-variance tradeoff. Thus instead of hard margin where training error is zero we prefer soft margin which generalises more. It is a method of regularisation.

Charcterizing support vectors:

(λi = 0 and ξi = 0) referring that the datapoint xi has been correctly classified

(0<λi <C): ,ξi ≡0, $y_i(w^T x_i + b) = 1$which concludes that xi is a support vector. All the support vectors which satisfy the condition 0<λi <C are ***free or unbounded support vectors*.**

(λi =C): $y_i(w^T x_i + b) = 1 - \xi_i$, ξi ≥0, implies xi is a support vector. When λi =C the support vectors are ***bounded support vectors***; that is, they lie inside the margin. Furthermore, for 0 ≤ ξi < 1, xi is correctly classified, but if ξi ≥ 1, xi is misclassified.

Benefits of Solving Dual rather than Primal Problem:

- According to the KKT conditions described in the derivation one of them states $\lambda_i(y_i(w \,.\, x_i + b) - 1 + \zeta_i) = 0$. Thus for all points not lying on support vectors must have λ to be zero. This reduces parameters in dual problems.

- Kernels map points to high dimensional plane. As we know the dual problem relies on the product $x_i^T x_j$. This saves a lot of computation when kernels are applied as they need not compute mapping of each point but just their dot product. Thus dual helps in classifying non linearly separable points with less computational effort.

- Microsoft Research also introduced Sequential Minimal Optimisation for solving dual Lagrangian SVM problem.

References:

https://www.stat.berkeley.edu/~arturof/Teaching/EE127/Notes/support_vector_machines.pdf

Introduction to Data Mining:
Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2005. Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc., USA.

https://homepage.cs.uri.edu/faculty/hamel/courses/2014/spring2014/csc581/lecture-notes/12-dual-maximum.pdf

3. Formulate the primal problem and derive the dual if there are multiple classes.

Crammer and Singer proposed an approach for multiclass problems.

Assuming there are K-classes and ℓ-training examples

Solving the primal problem

$$\min_{w_m, \xi_i} \frac{1}{2} \sum_{m=1}^{K} w_m^T w_m + C \sum_{i=1}^{\ell} \xi_i$$

$$w_{y_i}^T \phi(x_i) - w_m^T \phi(x_i) \geq e_i^m - \xi_i \quad , i = 1 \dots \ell$$

where $e_i^m \equiv 1 - \delta_{y_i, m}$ and

$$\delta_{y_i, m} \equiv \begin{cases} 1 & \text{if } y_i = m \\ 0 & \text{if } y_i \neq m \end{cases}$$

$\therefore$ Therefore the decision func. is

$$\text{arg max}_{m = 1, \dots, K} w_m^T \phi(x)$$

The dual problem is

$$\min_{\alpha} f(\alpha) = \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} K_{i,j} \bar{\alpha}_i^T \bar{\alpha}_j + \sum_{i=1}^{\ell} \bar{\alpha}_i^T \bar{e}_i$$

$$\sum_{m=1}^{K} \alpha_i^m = 0, \quad i = 1, \dots, \ell,$$

$$\alpha_i^m \leq 0 \text{ if } y_i \neq m$$

$$\alpha_i^m \leq C \text{ if } y_i = m$$

$$i = 1, \dots, \ell, \quad m = 1, \dots, k$$

where $K_{i,j} \equiv \phi(x_i)^T \phi(x_j)$

$$\bar{\alpha}_i \equiv \left[\alpha_i^1, \dots, \alpha_i^K\right]^T \text{ and } \bar{e}_i \equiv \left[e_i^1, \dots, e_i^K\right]^T$$

Then

$$w_m = \sum_{i=1}^{l} \alpha_i^m \phi(x_i)$$

If we write $\alpha \equiv \left[\alpha_1^1, \alpha_2^m \ldots, \alpha_1^K \ldots, \alpha_i^1 \ldots, \alpha_l^K\right]^T$

and $e \equiv \left[e_1^1, \ldots, e_1^K \ldots, e_l^1 \ldots, e_l^K\right]^T$

then the dual objective function can be written as.

$$\frac{1}{2}\alpha^T (K \otimes I)\alpha + e^T \alpha$$

where I is an K by K identity matrix and $\otimes$ is Kronecker product. Since K is positive semi-definite, $K \otimes I$, is the Hessian of the dual objective function is also positive semi definite

Thus the decision function is

$$\underset{m=1,\ldots,K}{\operatorname{argmax}} \sum_{i=1}^{l} \alpha_i^m K(x_i, x)$$

References

- A comparison of Methods for Multiclass Support Vector Machines
  Chih-Wei Hsu and Chih-Jen Lin
- On the algorithmic implementation of multiclass Kernel based vector machines
  Koby Crammer, Yoram Singer