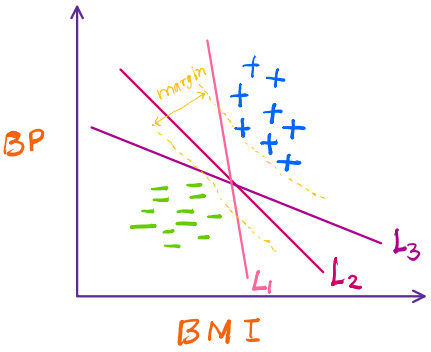
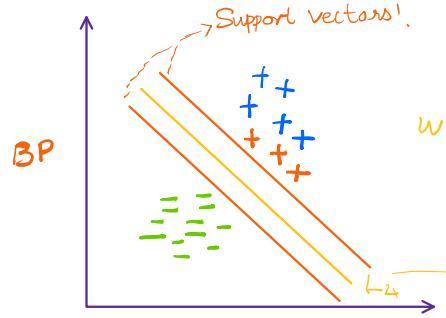
Lecture 15: Support Vector Machines

13 March 2024 11:01

Consider a health data set shown below; where + indicates the existence of a disease and −, the absence of the three classifiers L_1, L_2, L_3. L_2 is better as it allows some boundary (margin)



margin based classifier.



we need to maximize margin!

linear classifier

Let our linear classifier has weight W and bias, b such that all points on the classifier have $W^T x + b = 0$.

we classify points +1, −1 according to $W^T x + b > 0$ or $W^T x + b < 0$.

Dataset, $D = \{(x_1, y_1), \ldots (x_n, y_n)\}$ where $x_i \in \mathbb{R}^d$, d is large i.e., $d >> n$; $y_i \in \{+1, -1\}$

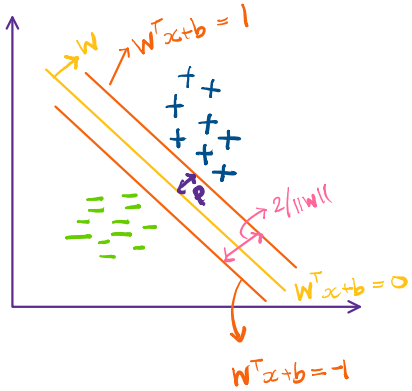
Two Variants of SVM:

1. Hard margin SVMs: allows no misclassification. (does not exist in all cases!)

2. Soft margin SVMs: allows misclassification, but to a limited extend.

HARD MARGIN SVM's:

Formal description:



Pick a point on the decision boundary, say x.

$\therefore W^T x + b = 0$

Move 'p' units in direction on W to reach the margin

$$\therefore W^T \left(x + p\frac{W}{\|W\|}\right) + b = 1$$

$$\Rightarrow W^T x + p\frac{\|W\|^2}{\|W\|} + b = 1$$

$$\Rightarrow p\|W\| = 1 \quad \text{i.e., } p = \frac{1}{\|W\|}$$

Larger the margin larger the

$w \cdot x + b = 1$

$\Rightarrow \quad P \|w\| = 1 \quad$ i.e., $P = \frac{1}{\|w\|}$

Larger the margin larger the robustness / error proofness

Thus, since we need a larger margin, we've

$$\max \frac{2}{\|w\|} \Rightarrow \min \left(\frac{\|w\|^2}{2} \right)$$

We also have following constraints:

$$y_i = +1 \Rightarrow w^T x_i + b \geq 1$$

$$y_j = -1 \Rightarrow w^T x_j + b \leq -1$$

i.e., $y_i (w^T x_i + b) \geq 1 \quad , \forall i = 1, \ldots, n$

Optimization problem:

$$\min \frac{1}{2} \|w\|^2$$

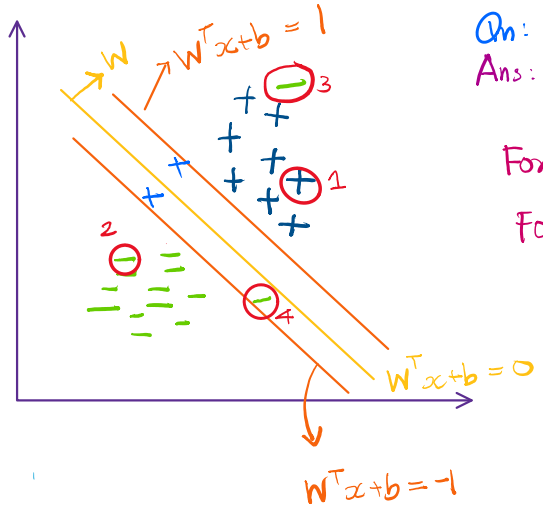
$$s.t. \quad y_i (w^T x_i + b) \geq 0$$

$$\forall i = 1, \ldots, n$$

(Convex optimization problem.)

Remark: for $i \in$ support vectors, $y_i (w^T x_i + b) = 1$

SOFT MARGIN SVM's:



Data points are not linearly separable as shown in fig.

Qn: How to capture the degree of misclassification?

Ans: We use a loss function for this.

For a correct classification, $y_i \cdot (w^T x_i + b) \geq 1$

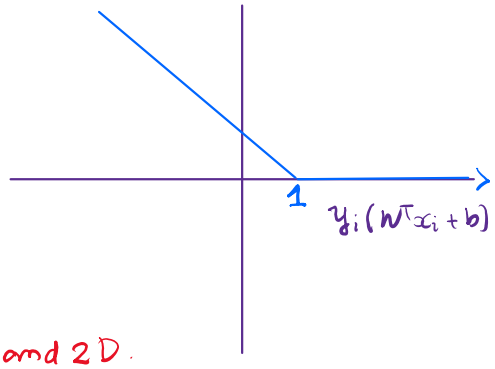
For misclassification: $y_i (w^T x_i + b) < 1$

Observe the quantity $1 - y_i (w^T x_i + b)$ → ≤ 0 for correct, > 0 for misclassified.

This is ≤ 0 for correct classification, and largely positive for great misclassification.

Our loss fn: $L_i (w, b) = \max \{0, 1 - y_i (w^T x_i + b)\}$

(Hinge Loss)



Our new optimization problem:

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} L_i (w, b) + \lambda \|w\|^2$$

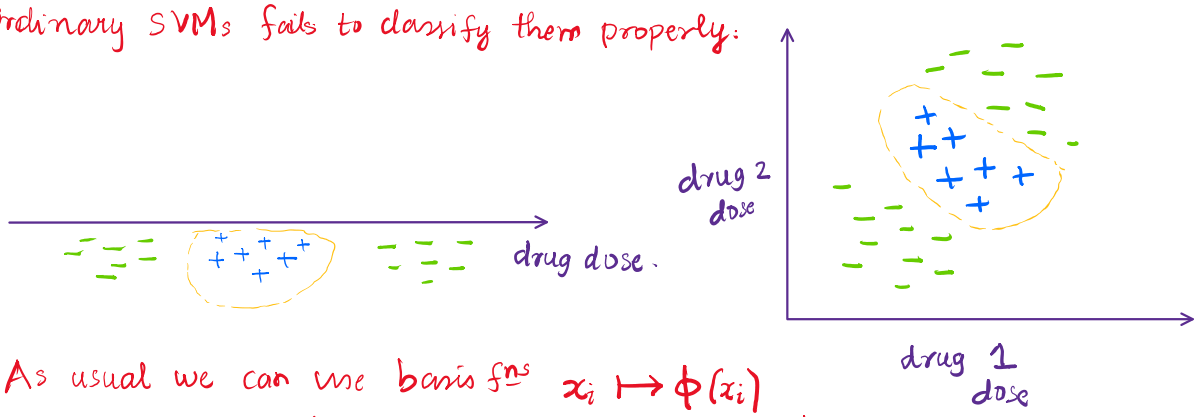
(λ: hyperparameter.)

λ determines tradeoff b/w misclassification and margin maximization.

Consider the following classification problems in 1D and 2D.

Ordinary SVMs fails to classify them properly:

Ordinary SVMs fails to classify them properly:



As usual we can use basis fns $x_i \mapsto \phi(x_i)$

But since $x_i \in \mathbb{R}^d$ and $d \gg n$, computation of ϕ is expensive. Thus we adopt KERNALISATION

KERNALIZATION: A method to calculate the higher dimension transformation in a computationally efficient manner.

Background of Kernalization:

Hard Margin SVM: $\min \frac{1}{2}\|w\|^2$ s.t. $y_i(w^T x_i + b) \geq 1$, $\forall i = 1, \dots n$.

PRIMAL

In general, our objective is:

$\min f_0(x)$ PRIMAL

s.t. $f_i(x) \leq 0$, $i = 1, \dots m$

$h_i(x) = 0$, $i = 1, \dots p$

Domain, D where feasible x's live

$D = \{x : f_i(x) \leq 0, \forall i ; h_j(x) = 0 \ \forall j\}$

Lagrangian:

$\mathcal{L}(x, \lambda, \gamma) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \gamma_j h_j(x)$

Consider the problem, $\lambda_i \geq 0$

$\max_{\lambda, \gamma} \left[\min_{x \in D} \mathcal{L}(x, \lambda, \gamma) \right]$

This forces us to follow constraints of Primal problem, otherwise Lagrange becomes too large.

$g(\lambda, \gamma) = \min_{x \in D} \mathcal{L}(x, \lambda, \gamma) \leq p^*$ ← primal optimal

(Convex opt: dual opt = p^*)

$\max_{\lambda \geq 0, \gamma} g(\lambda, \gamma)$ ← EQUIVALENT TO PRIMAL

called LAGRANGE DUAL

Langrange for HSVM: $\mathcal{L}(w, b, \lambda) = \frac{1}{2}\|w\|^2 - \sum \lambda_i (y_i (w^T x_i + b) - 1)$

$\lambda_i \geq 0$

$g(\lambda) = \min_{w,b} \mathcal{L}(w, b, \lambda)$

$\therefore \frac{\partial \mathcal{L}}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} \lambda_i y_i x_i$ ——①

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \lambda_i y_i = 0 \quad \text{---} \quad (2)$$

Using (1) and (2):

$$g(\lambda) = \sum \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j$$

Dual Problem: $\max_{\substack{\lambda_i \geq 0 \\ \forall i}} g(\lambda)$

Note $d >> n$. But we only need $x_i^T x_j$ inner product information for this problem. $(n \times n)$

$nd >> n^2$

Why we've used dual problem?

- computationally efficient, as $nd >> n^2$
- Kernel friendly.