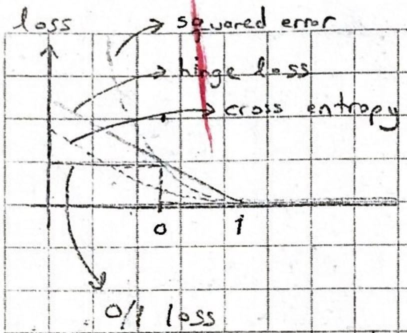

Many loss functions available.

For SVMs, we use hinge loss.

Kernel Trick

If the problem is non-linear, instead of trying to fit a non-linear model, we can map the problem to a new space by doing a non-linear transformation using a suitably chosen basis function and then use a linear model in this new space. The linear model in the new space corresponds to a nonlinear model in the original space.

Let us say we have the new dimensions calculated through the basis functions

$$z = \phi(x) \quad \text{where} \quad z_j = \phi_j(x), \quad j = 1, \dots, k$$

mapping from the d-dimensional x space to the k-dimensional z space.

The discriminant is ;

$$f(z) = w^T z$$

$$f(x) = w^T \phi(x) = \sum_{j=1}^{k} w_j \phi_j(x)$$

$$L_p = \frac{1}{2} \|w\|^2 + C \sum_t \xi^t$$

⇒ nonseparable case because we are not guaranteed that the problem is linearly separable in the new high dimensional space.

The constraints are defined in the new space:

$$y^t w^T \phi(x^t) \geq 1 - \xi^t$$

The Lagrangian is

$$L_p = \frac{1}{2}\|w\|^2 + C\sum_t \xi^t - \sum \alpha_t \left[y^t w^T \phi(x^t) - 1 + \xi^t \right] - \sum_t \mu^t \xi^t$$

$$\frac{\partial L_p}{\partial w} = w = \sum \alpha^t y^t \phi(x^t)$$

$$\frac{\partial L_p}{\partial \xi^t} = C - \alpha^t - \mu^t = 0$$

$$L_d = \sum_t \alpha^t - \frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s y^t y^s \underbrace{\boxed{\phi(x^t)^T \phi(x^s)}}_{K(x^t, x^s)}$$

$\Rightarrow$ kernel function

subject to

$$\sum_t \alpha^t y^t = 0 \quad \text{and} \quad 0 \leq \alpha^t \leq C, \ \forall t$$

Replace the inner product of the basis functions $\phi(x^t)^T \phi(x^s)$ by a Kernel function $K(x^t, x^s)$ between instances in the original input space.
Instead of mapping two instances x^t and x^s to the z space and doing a dot product there, we directly apply the kernel function in the original space

$$L_d = \sum_t \alpha^t - \frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s y^t y^s K(x^t, x^s)$$

The kernel function also shows up in the discriminant

$$f(x) = w^T \phi(x) = \sum_t \alpha^t y^t \phi(x^t)^T \phi(x)$$

$$= \sum \alpha^t y^t K(x^t, x)$$

This implies that if we have the kernel function, we do not need to map it to the new space at all.
For any valid kernel function, there exists a corresponding mapping function, but we do not need to know it.

The most popular general purpose kernel functions:

- Polynomials of degree q: $K(x^t, x) = (x^T x^t + 1)^q$

For example; $q = 2$, $d = 2$ $\quad K(x, y) = (x^T y + 1)^2 = (x_1 y_1 + x_2 y_2 + 1)^2$

$$= 1 + 2x_1 y_1 + 2x_2 y_2 + 2x_1 x_2 y_1 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2$$

corresponds to the inner product of the basis function

$$\phi(x) = [1, \sqrt{2} x_1, \sqrt{2} x_2, \sqrt{2} x_1 x_2, x_1^2, x_2^2]^T$$

- Radial basis functions (RBF)

$$K(x^t, x) = \exp\left[-\frac{\|x^t - x\|^2}{2s^2}\right]$$

defines a spherical kernel as in Parzen windows where x^t is the center and s (supplied by the user) defines the radius.