CS725:

Quick recap of SVMs

Hard-margin SVM

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

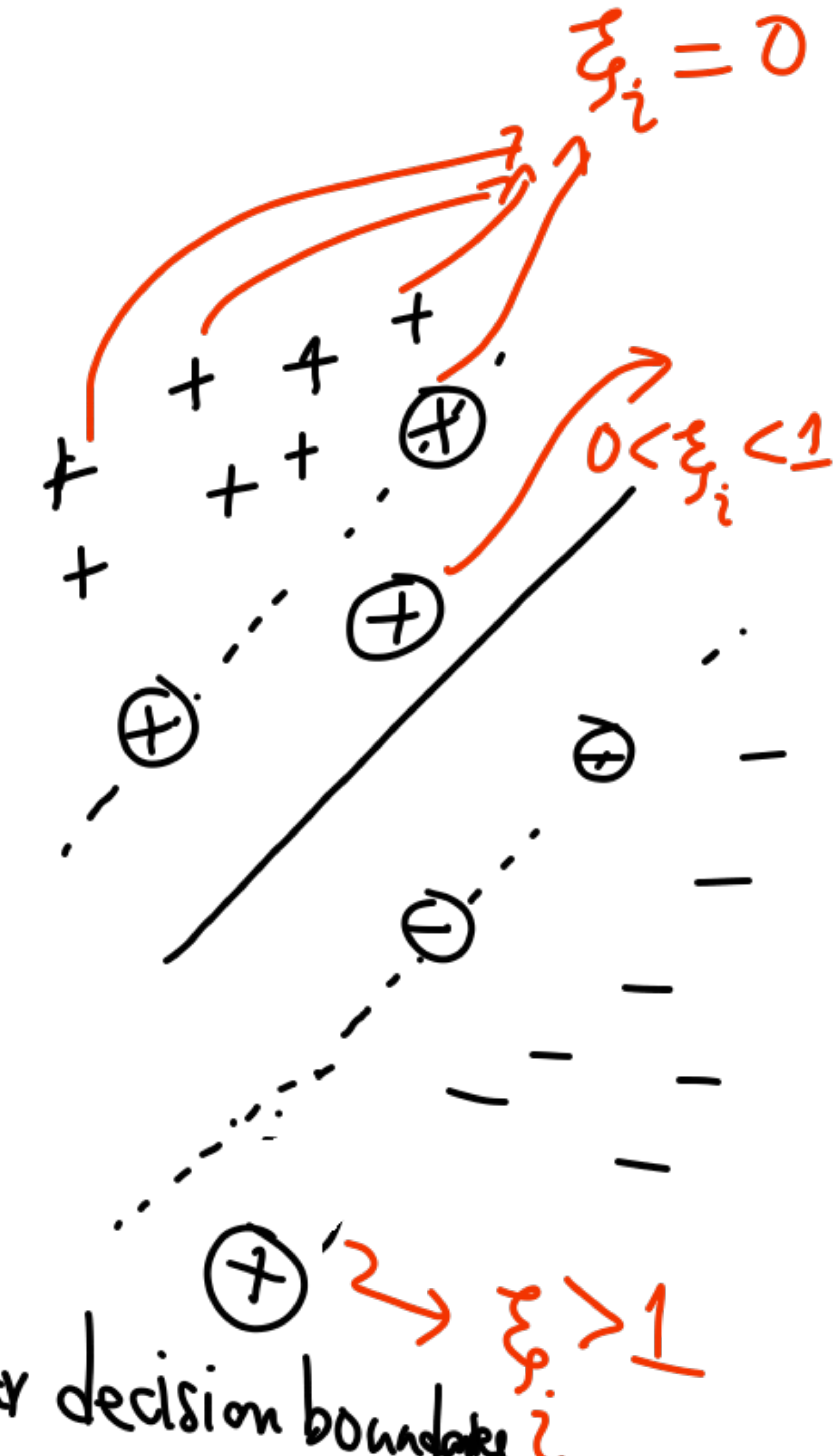
s.t. $y_i(w^T x_i + b) \geq 1$

Soft-margin SVM (Regularized Hinge loss)

$$\min_{w,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

s.t. $y_i(w^T x_i + b) \geq 1 - \xi_i$

$\xi_i \geq 0$



SVMs compared to NNs

- → In its current form as seen so far, SVMs yield linear decision boundaries
- → SVM has a global minimum solution unlike NNs
- → SVMs are theoretically well-motivated

How do we use the SVM at test time?

→ Solve the optimization problem to get w^*, b^*

→ $\hat{y}$ = Predicted label $\in \{-1, 1\}$ for a test instance x $= \text{sign}\Big(\underbrace{{w^*}^T x + b^*}_{\text{LINEAR DECISION BOUNDARY}}\Big)$

PRIMAL and DUAL forms (for optimization)

Given a function $f(w)$ and a set of inequality constraints $g_1(w) \leq 0$, $g_2(w) \leq 0$, $\ldots$, $g_m(w) \leq 0$, then the primal constrained optimization problem is:

$$\min_{w} f(w)$$

$$\text{s.t. } g_i(w) \leq 0 \text{ for } i = 1, \ldots, m$$

An equivalent unconstrained objective by defining a Lagrangian $L(w, \alpha) = f(w) + \sum_i \alpha_i g_i(w)$

and solving for $\min_{w} \max_{\alpha_i \geq 0} L(w, \alpha) = \min_{w} \max_{\alpha_i \geq 0} f(w) + \sum_i \alpha_i g_i(w)$

Why is there equivalence?

See $\max_{\alpha_i \geq 0} f(w) + \sum_i \alpha_i g_i(w) = \begin{cases} f(w) & \text{if } w \text{ satisfies all the constraints} \\ \infty & \text{otherwise} \end{cases}$

The primal solution : $p^* = \min_{w} \max_{\alpha_i \geq 0} L(w, \alpha)$

The dual solution ; $d^* = \max_{\alpha_i \geq 0} \min_{w} L(w, \alpha)$

Dual is typically less than or equal to the primal solution

i.e. $d^* \leq p^*$

Fortunately for SVMs, $d^* = p^*$!

Dual formulation for SVMs

Hard-margin SVM dual

$$\max_{\alpha_i} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

s.t.

$$\alpha_i \geq 0$$

$$\sum_i \alpha_i y_i = 0$$

Training instances appear as dot products in the training objective

Soft-margin SVM dual

$$\max_{\alpha_i} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

s.t.

$$0 \leq \alpha_i \leq C$$

$$\sum_i \alpha_i y_i = 0$$

From solving the dual formulations, we have:

① $$w^* = \sum_i \alpha_i y_i x_i, \quad i = 1, \ldots, n$$

[w^* is determined by those training pts with non-zero α_i's]

② $$\forall_i \quad \alpha_i g_i(w) = 0$$ [Comes out from technical (KKT) condition that holds for the SVM optimization problem]

for non-zero α_i's, $g_i(w) = 0 \Rightarrow y_i(w^T x_i + b) = 1$ // SUPPORT VECTORS

③ The training instances x_i, x_j appear as dot or inner products in the training objective

Predict at test time using an SVM

$\hat{y} = \text{sign}(w^{*T} x + b^*)$ for a test instance x → Ⓐ

$w^* = \sum_i \alpha_i y_i x_i$ → Ⓑ

Substitute Ⓑ in Ⓐ, $\hat{y} = \text{sign}\left(\sum_i \alpha_i y_i x_i^T x + b\right)$

Dot product between x and every training pt x_i

So far, SVMs yield linear decision boundaries

How do we use SVMs for non-linearly separable data?

original decision rule $\hat{y} = \text{sign}\left(\sum_i \alpha_i y_i x_i^T x + b\right)$

$\Downarrow$ Transform x using some basis function $\phi(x)$.

$$\hat{y} = \text{sign}\left(\sum_i \alpha_i y_i \phi(x_i)^T \phi(x) + b\right)$$

How do we compute $\phi(x)$? Also, for high-dimensional spaces, computing $\phi(x_i)^T \phi(x)$ is very expensive

Adopt the <u>KERNEL TRICK</u> using a kernel $K(x,y)$ defined on two inputs x, y in the original feature space.

A kernel takes two inputs in the original feature space, and produces an output which is the dot product of these two inputs projected to an implicit, high-dimensional space!

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

Example of a simple polynomial kernel (of order 2)

Consider a 2D point $x = (x_1, x_2)$

Define $\phi(x) = \begin{bmatrix} x_1^2 \\ \sqrt{2}\, x_1 x_2 \\ x_2^2 \end{bmatrix}$ $\quad \phi : \mathbb{R}^2 \to \mathbb{R}^3$

$$K(x, y) = (x^T y)^2$$
$$= \phi(x)^T \phi(y)$$

When is a kernel valid?

① If we can construct some $\phi(x)$, s.t. $K(x,y) = \phi(x)^T \phi(y)$

② Mercer's Theorem: Let $K: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Then, for K to be a valid kernel, it is necessary and sufficient that a subset of points $x_1, \dots, x_n$ $(n<\infty)$ results in a kernel matrix which is symmetric and positive semi-definite.

[$n \times n$ matrix whose $(i,j)^{th}$ entry is $K(x_i, x_j)$]

③ Given existing valid kernels, some properties hold over them to result in a valid kernel. Below, $K(x,y)$ is valid;

ⓐ $K(x,y) = K_1(x,y) + K_2(x,y)$, K_1, K_2 are both valid

ⓑ $K(x,y) = K_1(x,y)\,K_2(x,y)$, K_1, K_2 are both valid

ⓒ $K(x,y) = c\,K_1(x,y)$, $c \neq 0$

Example of using the properties from 3 in the last slide

Is $K(x,y) = (x^T y)^d$ for any integer $d > 0$ a valid kernel?

YES! If $K_0(x,y) = x^T y$ [Linear kernel]

then by the product rule, $K(x,y) = (x^T y)^d$ is also valid

How about $K(x,y) = \left(c + x^T y\right)^d$?

YES! c is obtained via a constant kernel ($\phi(x) = \sqrt{c}$, which is valid)

By sum and prod rules, $K(x,y)$ is valid

$K(x,y) = (x^T y)^d$ POLYNOMIAL KERNEL of order d

$K(x,y) = (c + x^T y)^d$ POLYNOMIAL KERNEL of order up to d

<u>HW</u> Q. Is $K(x,y) = ||x+y||^2$ a valid kernel?