discussion 9

kernel/SVM

kernel

Now we need a metric to measure such a similarity. Typically, we use inner product, and *kernels function* is therefore defined as

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\top \phi(\mathbf{x}),$$

where $\phi(\cdot)$ is a fixed nonlinear *feature space* mapping.

valid kernel condition

Symmetric and Positive Semi-Definite $\Leftrightarrow$ Kernel Function $\Leftrightarrow < \phi(x), \phi(x') >$ for some $\phi(.)$.

- K(x,z) $= \sum_{i=1}^{m} \alpha_i k_i(x, z)$ (Closed under non-negative linear multiplication)

- K(x,z) $= \sum_{i=1}^{m} \alpha_i k_i(x,z)$ (Closed under non-negative linear multiplication)

Proof: K(x,z) $= \sum_{i=1}^{m} \alpha_i k_i(x,z)$ is a valid kernel

Symmetry: $K(z,x) = \sum_{i=1}^{m} \alpha_i k_i(z,x)$. Because each k_i is a kernel function, it is symmetric, so $k_i(z,x) = k_i(x,z)$. Therefore,
$K(z,x) = \sum_{i=1}^{m} \alpha_i k_i(z,x) = \sum_{i=1}^{m} \alpha_i k_i(x,z) = K(x,z) \Rightarrow K(z,x) = K(x,z)$

Positive Semi-definite: Let $\mathbf{u} \in \mathbb{R}^n$ be arbitrary. The Gram matrix of K, denoted by G has the property, $G_{i,j} = K(x_i, x_j) = \sum_{i=1}^{m} \alpha_i k_i(z,x) \Rightarrow G = \alpha_1 G_1 + ... + \alpha_m G_m$. Now $\mathbf{u}^T G \mathbf{u} = \mathbf{u}^T(\alpha_1 G_1 + ... + \alpha_m G_m)\mathbf{u}$
$= \alpha_1 \mathbf{u}^T G_1 \mathbf{u} + ... + \alpha_m \mathbf{u}^T G_m \mathbf{u} = \sum_{i=1}^{m} \alpha_i \mathbf{u}^T G_i \mathbf{u}$.
$\mathbf{u}^T G_i \mathbf{u} \geq 0$, and $\alpha_i \geq 0$, so $\alpha_i \mathbf{u}^T G_i \mathbf{u} \geq 0$.

- K(x,z) $= x^T A^T A z$ for any matrix $A \in \mathbb{R}^{mXn}$

- K(x,z) $= x^T A^T A z$ for any matrix $A \in \mathbb{R}^{mXn}$

Proof: K(x,z) $= x^T A^T A z$ for any matrix $A \in \mathbb{R}^{mXn}$ is a valid Kernel.

For this proof, we are going to show $K(x, z)$ is an inner product on some Hilbert Space. Let $\phi(x) = Ax$, then $< \phi(x), \phi(z) >= \phi(x)^T \phi(z) = (Ax)^T (Az) = x^T A^T A z = K(x, z) \Rightarrow < \phi(x), \phi(z) >= K(x, z)$.

Therefore, $K(x, z)$ is an inner product on some Hilbert Space.

$K(x, z) = \exp(\gamma \|x - z\|^2)$ for some $\gamma > 0$

$$k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x}) k_1(\mathbf{x}, \mathbf{x}') f(\mathbf{x}')$$

$$k(\mathbf{x}, \mathbf{x}') = \exp(k_1(\mathbf{x}, \mathbf{x}'))$$

SVM

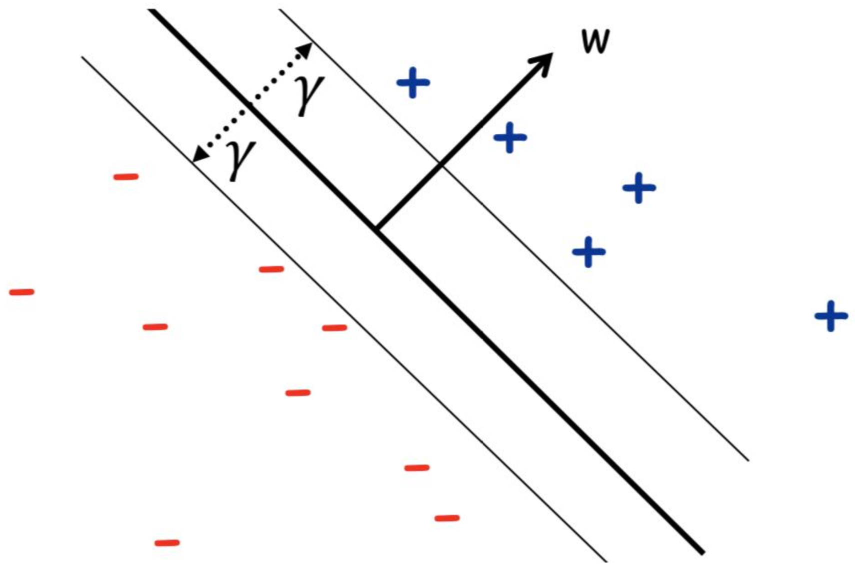
- Consider two-class classification problem using linear model of the form

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b.$$

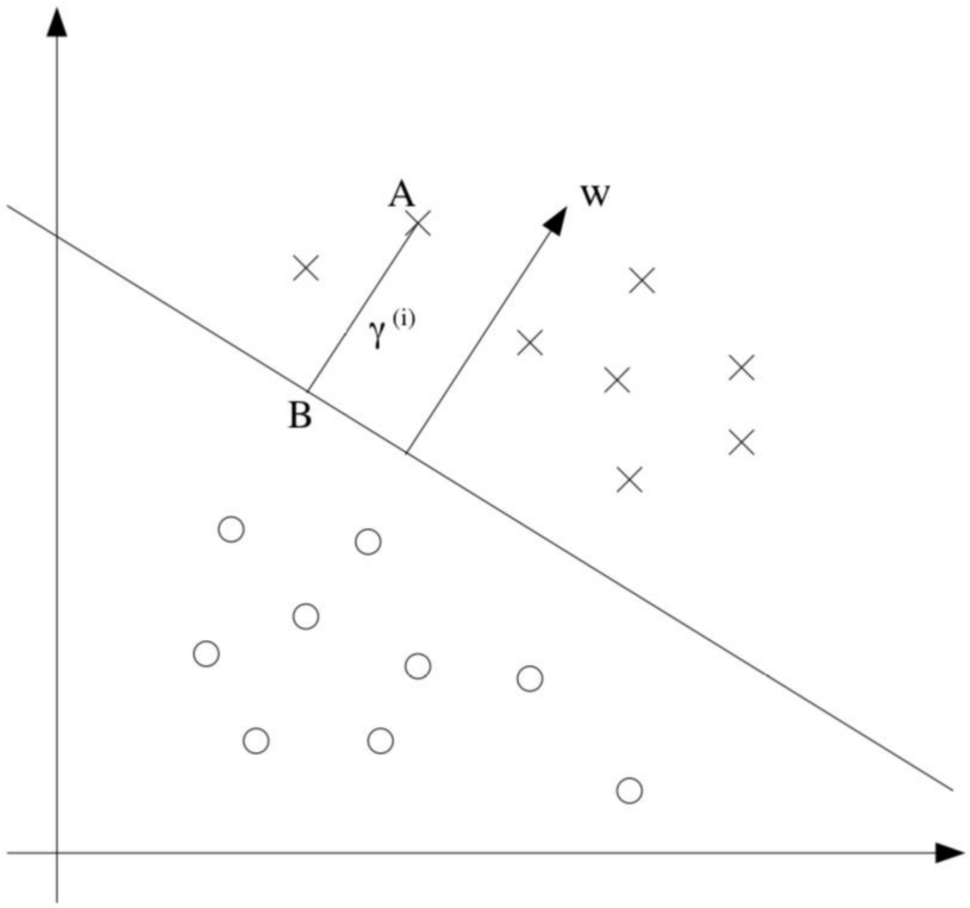
margin

Definition: The margin γ_w of a set of examples S wrt a linear separator w is the smallest margin over points $x \in S$.

Definition: The margin γ of a set of examples S is the maximum γ_w over all linear separators w.



margin



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

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

maximum r

Directly optimize for the maximum margin separator: SVMs

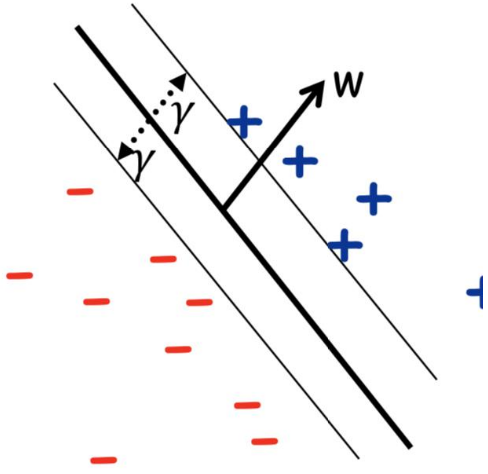
First, assume we know a lower bound on the margin γ

Input: γ, $S=\{(x_1, y_1), \ldots, (x_m, y_m)\}$;

Find: some w where:

- $||w||^2 = 1$
- For all i, $y_i w \cdot x_i \geq \gamma$

Output: w, a separator of margin γ over S



Realizable case, where the data is linearly separable by margin γ

SVM

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

"$||w|| = 1$" constraint is a nasty (non-convex) one,

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

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

minimize w

Directly optimize for the maximum margin separator: SVMs

Input: S={$(x_1, y_1), \ldots, (x_m, y_m)$};

$\text{argmin}_w ||w||^2$ s.t.:

- For all i, $y_i w \cdot x_i \geq 1$

This is a **constrained optimization** problem.

Lagrange duality

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

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

Here, the β_i's are called the **Lagrange multipliers**. We would then find and set $\mathcal{L}$'s partial derivatives to zero:

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0; \quad \frac{\partial \mathcal{L}}{\partial \beta_i} = 0,$$

and solve for w and β.

Consider the following, which we'll call the **primal** optimization problem:

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

To solve it, we start by defining the **generalized Lagrangian**

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

Here, the "$\mathcal{P}$" subscript stands for "primal." Let some w be given. If w violates any of the primal constraints (i.e., if either $g_i(w) > 0$ or $h_i(w) \neq 0$ for some i), then you should be able to verify that

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

$$\begin{aligned} \theta_{\mathcal{P}}(w) &= \max_{\alpha, \beta \,:\, \alpha_i \ge 0} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w) && (1) \\ &= \infty. && (2) \end{aligned}$$

primal/dual problem

- primal

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

- dual

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

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

KKT

Karush-Kuhn-Tucker (KKT) conditions, which are as follows:

$$\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, n \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\le 0, \quad i = 1, \ldots, k \\
\alpha^* &\ge 0, \quad i = 1, \ldots, k
\end{aligned}$$

SVM

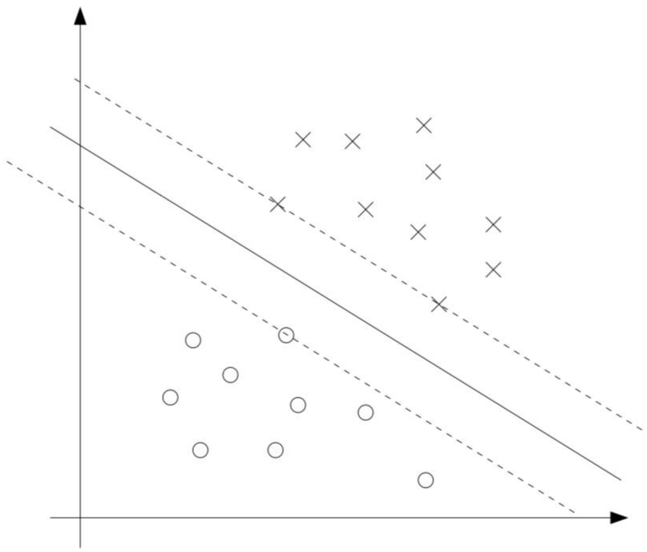
$$\min_{w,b} \quad \frac{1}{2}||w||^2$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m$$

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

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

support vectors

- These three points are called the support vectors in this problem.



Lagrangian for our optimization problem

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

process

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

This implies that

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

As for the derivative with respect to b, we obtain

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

plug that back into the Lagrangian

$$\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} - b \sum_{i=1}^{m} \alpha_i y^{(i)}.$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0.$$

$$\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}$$

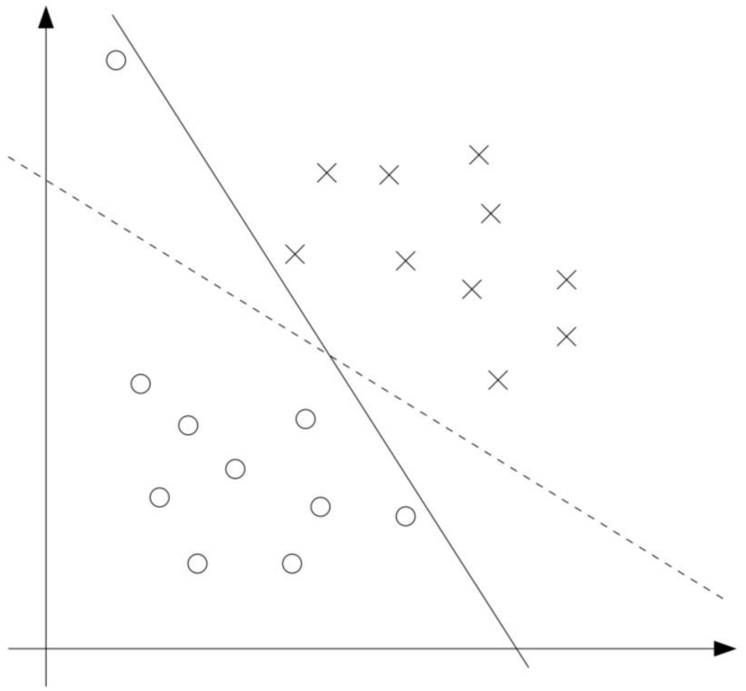
the following dual optimization problem:

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

Regularization and the non-separable case
Soft-Margin SVM

- make the algorithm work for non-linearly separable datasets as well as be less sensitive to outliers

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



As before, we can form the Lagrangian:

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

Here, the α_i's and r_i's are our Lagrange multipliers (constrained to be ≥ 0). We won't go through the derivation of the dual again in detail, but after setting the derivatives with respect to w and b to zero as before, substituting them back in, and simplifying, we obtain the following dual form of the problem:

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

svm -- hinge loss

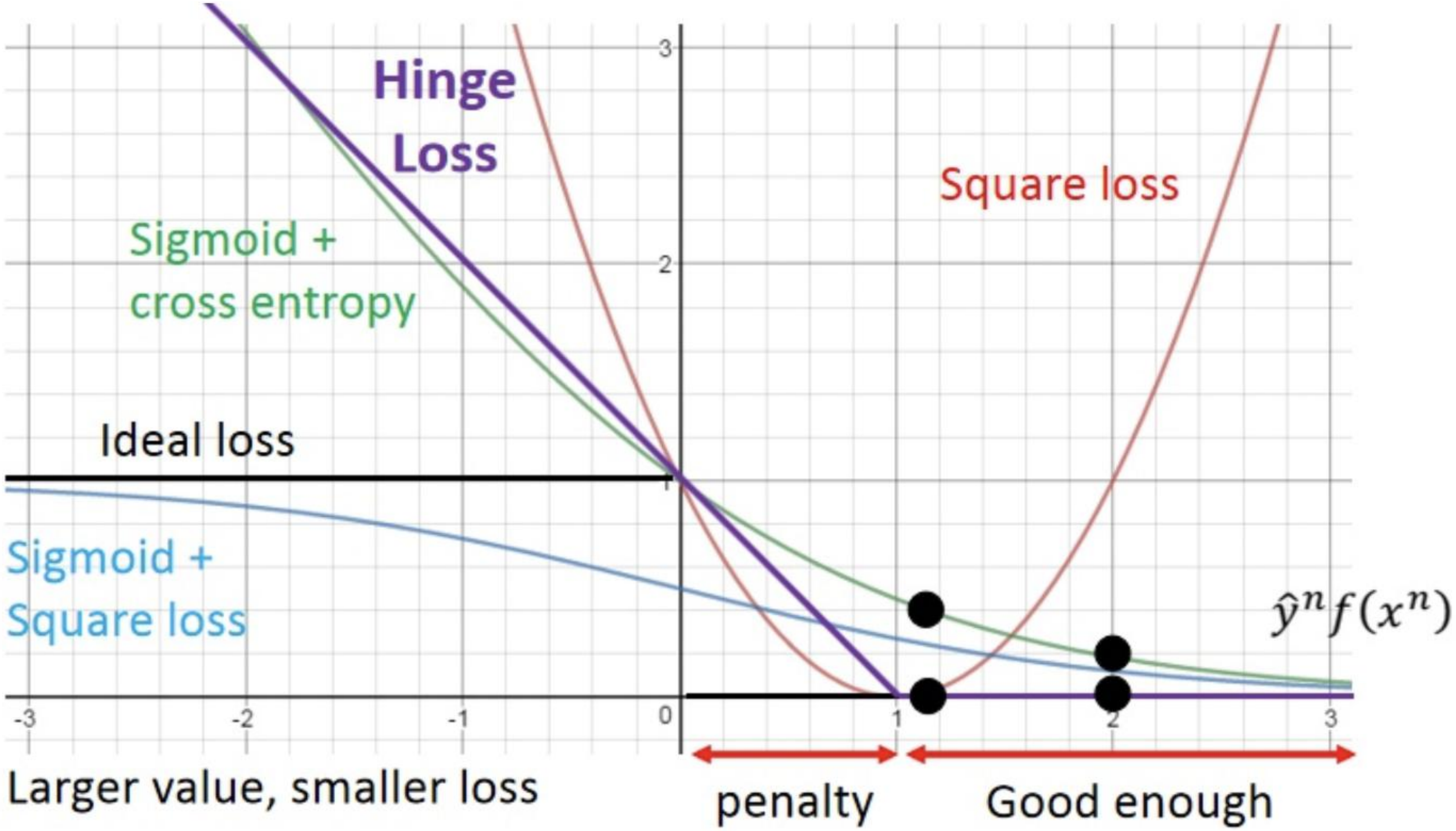
$$\frac{\partial L}{\partial \zeta_i} = c - r_i = 0 \Rightarrow r_i = c \, , \forall i$$

$$\min_{\vec{w},b} \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(\vec{w}\cdot\vec{x}_i + b))$$

$$\boxed{\ell(y_i, f(\vec{x}_i; \vec{w}, b)) = \max(0, 1 - y_i f(\vec{x}_i; \vec{w}, b))}$$

$$\min_{\vec{w},b} C\sum_{i=1}^{n} \ell(y_i, f(\vec{x}_i; \vec{w}, b)) + \frac{1}{2}\|\vec{w}\|^2$$

hinge loss



Hinge Loss
Square loss
Sigmoid + cross entropy
Ideal loss
Sigmoid + Square loss
$\hat{y}^n f(x^n)$
-3
-2
-1
0
1
2
3
Larger value, smaller loss
penalty
Good enough