

SVM Objective

He He

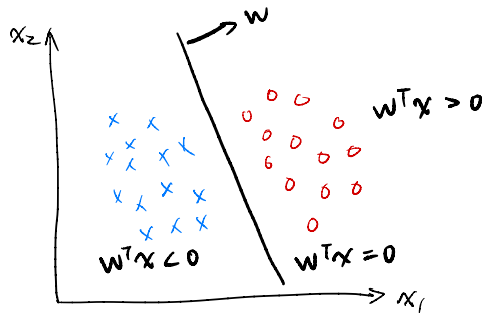
CDS, NYU

Feb 23, 2021

Maximum Margin Classifier

Linearly Separable Data

Consider a linearly separable dataset \mathcal{D} :

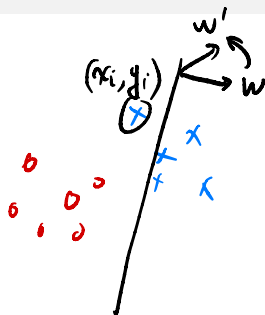


Find a separating hyperplane such that

- $w^T x_i > 0$ for all x_i where $y_i = +1$
- $w^T x_i < 0$ for all x_i where $y_i = -1$

The Perceptron Algorithm

- Initialize $w \leftarrow 0$
- While not converged (exists misclassified examples)
 - For $(x_i, y_i) \in \mathcal{D}$
 - If $y_i w^T x_i < 0$ (wrong prediction)
 - Update $w \leftarrow w + y_i x_i$

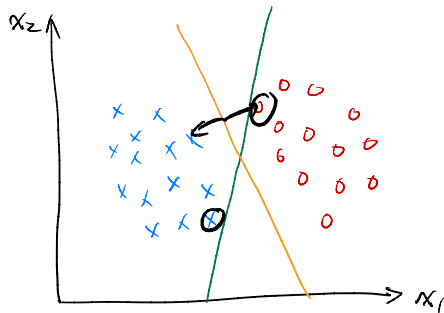


- Intuition: move towards misclassified positive examples and away from negative examples
- Guarantees to find a zero-error classifier (if one exists) in finite steps
- What is the loss function if we consider this as a SGD algorithm?

Maximum-Margin Separating Hyperplane

For separable data, there are infinitely many zero-error classifiers.

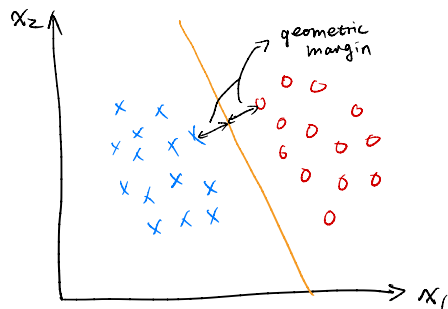
Which one do we pick?



(Perceptron does not return a unique solution.)

Maximum-Margin Separating Hyperplane

We prefer the classifier that is farthest from both classes of points



- Geometric margin: smallest distance between the hyperplane and the points
- Maximum margin: *largest* distance to the closest points

Geometric Margin

We want to maximize the distance between the **separating hyperplane** and the **closest** points.

Let's formalize the problem.

Definition (separating hyperplane)

We say (x_i, y_i) for $i = 1, \dots, n$ are **linearly separable** if there is a $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $y_i(w^T x_i + b) > 0$ for all i . The set $\{v \in \mathbb{R}^d \mid w^T v + b = 0\}$ is called a **separating hyperplane**.

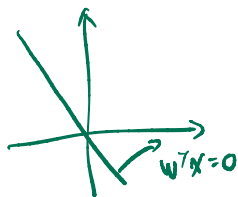
Definition (geometric margin)

Let H be a hyperplane that separates the data (x_i, y_i) for $i = 1, \dots, n$. The **geometric margin** of this hyperplane is

$$\min_i d(x_i, H),$$

the distance from the hyperplane to the closest data point.

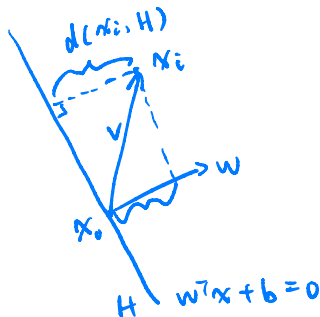
Distance between a Point and a Hyperplane



$$v = \alpha w + \beta w_{\perp}$$

- Projection of $v \in \mathbb{R}^d$ onto $w \in \mathbb{R}^d$: $\frac{v \cdot w}{\|w\|_2} = \alpha$

- Distance between x_i and H :



$$\begin{aligned} v &= x_i - x_0 \\ d(x_i, H) &= \text{proj}(v, w) \\ &= \frac{v \cdot w}{\|w\|_2} \\ &= \frac{(x_i - x_0) \cdot w}{\|w\|_2} \\ &= \frac{x_i^T w + b}{\|w\|_2} \end{aligned}$$

$$d(x_i, H) = \left| \frac{w^T x_i + b}{\|w\|_2} \right| = \frac{y_i(w^T x_i + b)}{\|w\|_2}$$

Maximize the Margin

We want to maximize the geometric margin:

$$\text{maximize } \min_i d(x_i, H).$$

Given separating hyperplane $H = \{v \mid w^T v + b = 0\}$, we have

$$\text{maximize } \min_i \frac{y_i(w^T x_i + b)}{\|w\|_2}.$$

Let's remove the inner minimization problem by

$$\begin{aligned} &\text{maximize } M \\ &\text{subject to } \frac{y_i(w^T x_i + b)}{\|w\|_2} \geq M \quad \text{for all } i \end{aligned}$$

Note that the solution is not unique (why?). $w \leftarrow \alpha w \quad b \leftarrow \alpha b$

Maximize the Margin

Let's fix the norm $\|w\|_2$ to $1/M$ to obtain: $\|w\|_2 = \frac{1}{M}$

$$\begin{array}{ll}\text{maximize} & \frac{1}{\|w\|_2} \\ \text{subject to} & y_i(w^T x_i + b) \geq 1 \quad \text{for all } i\end{array}$$

It's equivalent to solving the minimization problem

$$\begin{array}{ll}\text{minimize} & \frac{1}{2} \|w\|_2^2 \\ \text{subject to} & y_i(w^T x_i + b) \geq 1 \quad \text{for all } i\end{array}$$

Note that $y_i(w^T x_i + b)$ is the (functional) margin.

In words, it finds the minimum norm solution which has a margin of at least 1 on all examples.

Soft Margin SVM

What if the data is *not* linearly separable?

For any w , there will be points with a negative margin.

Introduce **slack variables** to penalize small margin:

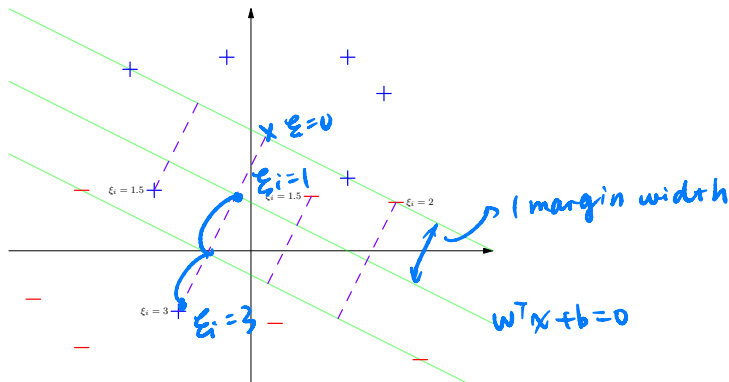
$$\begin{aligned} & \text{minimize} && \frac{1}{2} \|w\|_2^2 + \frac{C}{n} \sum_{i=1}^n \xi_i \\ & \text{subject to} && y_i (w^T x_i + b) \geq 1 - \xi_i \quad \text{for all } i \\ & && \xi_i \geq 0 \quad \text{for all } i \end{aligned}$$

- If $\xi_i = 0 \forall i$, it's reduced to hard SVM.
- What does $\xi_i > 0$ mean?
- What does C control?

Slack Variables

$d(x_i, H) = \frac{y_i(w^T x_i + b)}{\|w\|_2} \geq \frac{1 - \xi_i}{\|w\|_2}$, thus ξ_i measures the violation by multiples of the geometric margin:

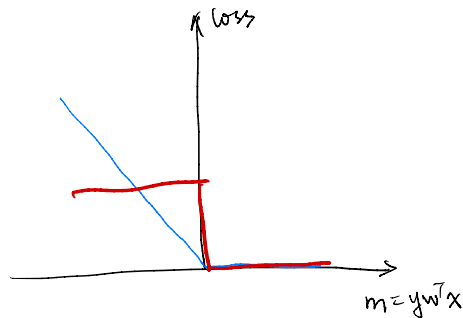
- $\xi_i = 1$: x_i lies on the hyperplane
- $\xi_i = 3$: x_i is past 2 margin width beyond the decision hyperplane



Minimize the Hinge Loss

Perceptron Loss

$$\ell(x, y, w) = \max(0, -yw^T x)$$

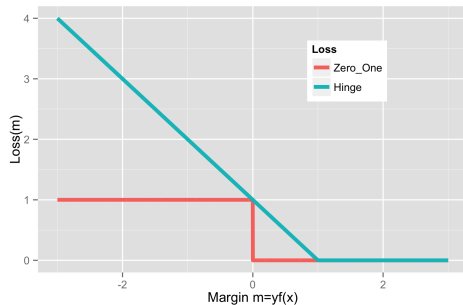


$$w \leftarrow w + y_i x_i$$
$$w \leftarrow w - \eta \nabla f(w)$$

If we do ERM with this loss function, what happens?

Hinge Loss

- SVM/Hinge loss: $\ell_{\text{Hinge}} = \max\{1 - m, 0\} = (1 - m)_+$
- Margin $m = yf(x)$; “Positive part” $(x)_+ = x1(x \geq 0)$.



Hinge is a **convex, upper bound** on 0–1 loss. Not differentiable at $m = 1$. We have a “margin error” when $m < 1$.

Support Vector Machine

Using ERM:

- Hypothesis space $\mathcal{F} = \{f(x) = w^T x + b \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$.

- ℓ_2 regularization (Tikhonov style)

- Hinge loss $\ell(m) = \max\{1 - m, 0\} = (1 - m)_+$ $y w^T x$

- The SVM prediction function is the solution to

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \underbrace{\frac{\lambda}{2} \|w\|^2}_{\text{regularizer}} + \underbrace{\frac{c}{n} \sum_{i=1}^n \max(0, 1 - y_i [w^T x_i + b])}_{\text{ER}}.$$

- Not differentiable because of the max

SVM as a Constrained Optimization Problem

- The SVM optimization problem is equivalent to

$$\begin{aligned} &\text{minimize} && \frac{1}{2} \|w\|^2 + \frac{c}{n} \sum_{i=1}^n \xi_i \\ &\text{subject to} && \xi_i \geq \max(0, 1 - y_i [w^T x_i + b]) . \end{aligned}$$

- Which is equivalent to

$$\begin{aligned} &\text{minimize} && \frac{1}{2} \|w\|^2 + \frac{c}{n} \sum_{i=1}^n \xi_i \\ &\text{subject to} && \xi_i \geq (1 - y_i [w^T x_i + b]) \text{ for } i = 1, \dots, n \\ &&& \xi_i \geq 0 \text{ for } i = 1, \dots, n \end{aligned}$$

Two ways to derive the SVM optimization problem:

- Maximize the (geometric) margin
- Minimize the hinge loss with ℓ_2 regularization

Both leads to the minimum norm solution satisfying certain margin constraints.

- **Hard-margin SVM:** all points must be correctly classified with the margin constraints
- **Soft-margin SVM:** allow for margin constraint violation with some penalty