Null Model: a model where $\vec{x}$ does not matter; here, $\mathcal{H} = \Big\{0, 1\Big\}$. Thus

$$g = \mathcal{A}(\mathcal{D} = \vec{y}, \mathcal{H}) = \text{Mode}(\vec{y})$$

Let

$$\mathcal{H} = \Big\{\mathbb{1}_{\vec{w}\cdot\vec{x}>0} : \vec{w} \in \mathbb{R}^{p+1}\Big\}$$

Assume linear separability. Which hyperplane specified by $\vec{w}$ is the best? This would be the max-margin hyperplane, created by a wedge in between 2 lines separating data. This is called the Support Vector Machine (SVM) where SV is essential observations and M is model. Since data point $\vec{x}_i$ are "vectors", they are called the support vectors.
Let

$$H = \Big\{\mathbb{1}_{\vec{w}\cdot\vec{x}+b>0} : w \in \mathbb{R}^p, b \in \mathbb{R}\Big\}$$

Example: Let $x_2 = 2x_1 + 2$. In the general equation of a line form, this is

$$2x_1 - x_2 + 3 = 0$$

To fit this into $\mathcal{H}$, we let $w = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$, $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $b = 3$. Then $\vec{w} \cdot \vec{x} + b = 0$. Here $\vec{w}$ is a normal vector orthogonal to the line.

Euclidean Norm: $\|\vec{w}\| = \sqrt{\sum_{j=1}^{p} w_i^2}$ - this is the length of the vector $\vec{w}$.
Normalized Vector $\vec{w}_0$: $\vec{w}_0 = \frac{\vec{w}}{\|\vec{w}\|}$ - this is the vector $\vec{w}$ in the same direction but with length 1.

$$\|\vec{w}_0\| = \sqrt{\sum_{j=1}^{p} \Big(\frac{w_j}{\|\vec{w}\|}\Big)^2} = \sqrt{\frac{1}{\|\vec{w}\|^2}\sum_{j=1}^{p} w_j^2} = \sqrt{\frac{1}{\|\vec{w}\|^2}\|\vec{w}\|^2} = 1$$

Length of the line: $\vec{l} = \alpha\vec{w}_0$ where $\alpha \in \mathbb{R}$.

$$\|\vec{x}\| = \sqrt{\sum_{j=1}^{p}(\alpha w_{0j})^2} = \sqrt{\alpha^2\sum_{j=1}^{p} w_{0j}^2} = \sqrt{\alpha^2\|\vec{w}_0\|^2} = |\alpha|$$

Let $\vec{l}$ be the vector from the origin to the line $\vec{w} \cdot \vec{x} + b = 0$, perpendicular to it. Solve for α. Note that $\vec{l}$ is on the line $\vec{w} \cdot \vec{x} + b = 0$.

$$\begin{aligned}
\vec{w} \cdot \vec{l} + b &= 0 \\
\vec{w} \cdot \alpha\vec{w}_0 + b &= 0 \\
\vec{w} \cdot \alpha\frac{\vec{w}}{\|\vec{w}\|} + b &= 0 \\
\alpha\frac{\|\vec{w}\|^2}{\|\vec{w}\|} + b &= 0 \\
\alpha &= -\frac{b}{\|\vec{w}\|}
\end{aligned}$$

Therefore $|\alpha| = \frac{b}{\|\vec{w}\|}$.
Going back the graph, we let the upper line be denoted as $\vec{w} \cdot \vec{x} + b + \delta = 0$ and the lower line as $\vec{w} \cdot \vec{x} + b - \delta = 0$. By this logic, the distance between the upper and lower lines is

$$\frac{b+\delta}{\|\vec{w}\|} - \frac{b-\delta}{\|\vec{w}\|} = \frac{2\delta}{\|\vec{w}\|}$$

Since $c(\vec{w} \cdot \vec{x} + b) = 0$, there are infinite solutions because $c \in \mathbb{R}$. If we let $\delta = 1$, then we can find a unique solution to the equation, namely, $\frac{2}{\|\vec{w}\|}$.
Constrain all $y = 1$s to be above the upper line and all $y = 0$s to be below the lower line. Then we find that

$$\begin{aligned}
\vec{w} \cdot \vec{x} + b + 1 &\geq 0 \\
\vec{w} \cdot \vec{x} + b &\geq -1 \\
\forall i \text{ such that } y_i &= 1 \\
\vec{w} \cdot \vec{x}_i + b &\geq -1 \\
(y_i - \frac{1}{2})(\vec{w} \cdot \vec{x}_i + b) &\geq -(y_i - \frac{1}{2}) \\
\text{Since } y_i &= 1 \\
(y_i - \frac{1}{2})(\vec{w} \cdot \vec{x}_i + b) &\geq -(1 - \frac{1}{2}) \\
(y_i - \frac{1}{2})(\vec{w} \cdot \vec{x}_i + b) &\geq -\frac{1}{2}
\end{aligned}$$

Now for all i such that $y_i = 0$,

$$\begin{aligned}
\vec{w} \cdot \vec{x}_i + b - 1 &\geq 0 \\
\vec{w} \cdot \vec{x}_i + b &= 1 \\
(y_i - \frac{1}{2})(\vec{w} \cdot \vec{x}_i + b) &\geq (y_i - \frac{1}{2}) \\
\text{Since } y_1 &= 0 \\
(y_i - \frac{1}{2})(\vec{w} \cdot \vec{x}_i + b) &\geq -\frac{1}{2}
\end{aligned}$$

This is the condition of perfect separability.
To maximize $\frac{2}{\|\vec{w}\|}$, we want to minimize $\|\vec{w}\|$ subject to for all i, $(y_i - \frac{1}{2})(\vec{w} \cdot \vec{x}_i + b) \geq -\frac{1}{2}$ over $\vec{w} \in \mathbb{R}^p$ and $b \in \mathbb{R}$. The $\left\{\vec{w}, b\right\}$ solution is the SVM solution. This approach assumes perfect linear separability. In the real world, we do not always get that. We need a better $\mathcal{A}$.

Let $SAE = \sum_{j=1}^{n} \mathbb{1}_{\hat{y}_i \neq y_i}$. Let's make a loss function dependent on how bad the error is. Consider the following:

$$H_i = \max\left\{0, -\frac{1}{2} - (y_i - \frac{1}{2})(\vec{w} \cdot \vec{x}_i + b)\right\}$$

This is called the hinge loss.
Imagine the point is correctly classified and that it obeys the inequality. Consider it is above

the inequality by $d \geq 0$. Then

$$(y_i - \frac{1}{2})(\vec{w} \cdot \vec{x}_i + b) = -\frac{1}{2} + d \geq -\frac{1}{2}$$

This is correct and so

$$H_i = \max\left\{0, -\frac{1}{2} - (-\frac{1}{2} + d)\right\} = \max\left\{0, -d\right\} = 0$$

This makes sense because if it is correctly classified, then there should be zero error.
Imagine the point is incorrectly classified and hence does not obey the inequality. Consider it is below the inequality by $d > 0$. Then

$$(y_i - \frac{1}{2})(\vec{w} \cdot \vec{x}_i + b) = -\frac{1}{2} - d < -\frac{1}{2}$$

This is incorrect but

$$H_i = \max\left\{0, -\frac{1}{2} - (-\frac{1}{2} - d)\right\} = \max\left\{0, d\right\} = d > 0$$

This makes sense; a mistake was made.

Minimize

$$\frac{1}{n}\sum_{i=1}^{n} H_i + \alpha + \|\vec{w}\|^2$$

This is the same as

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} \max\left\{0, -\frac{1}{2} - (y_i - \frac{1}{2})(\vec{w} \cdot \vec{x}_i + b)\right\}}_{\text{min. avg. hinge loss}} + \underbrace{\lambda\|\vec{w}\|^2}_{\text{max margin}}$$

This is our objective/loss function and it is a tradeoff between two goals. The parameters are $w \in \mathbb{R}^p$ and $b \in \mathbb{R}$. Therefore we still have $p + 1$ parameters. What is λ? It is a predefined constant called a "hyperparameter." I is considered a tuning knob on the $\mathcal{A}$ itself. Recall $g(\vec{x}) = \mathbb{1}_{\vec{w}\cdot\vec{x}+b}$. There is no λ here. In fact, λ only affects which $g \in \mathcal{H}$ will be selected.
Note: Perceptron algorithm does not work if not linearly separable!

If $\lambda \approx 0$, we only care about errors and not a max margin. One error far away can ruin the nice separation line. If $\lambda \approx \infty$, we only care about the best line of separation and not about errors. This makes no sense.
Considering λ is picked reasonably, how do we solve for $\left\{\vec{w}, b\right\}$? Use modern numerical optimization methods: quadratic programming, sub-gradient descent, coordinate descent, etc.