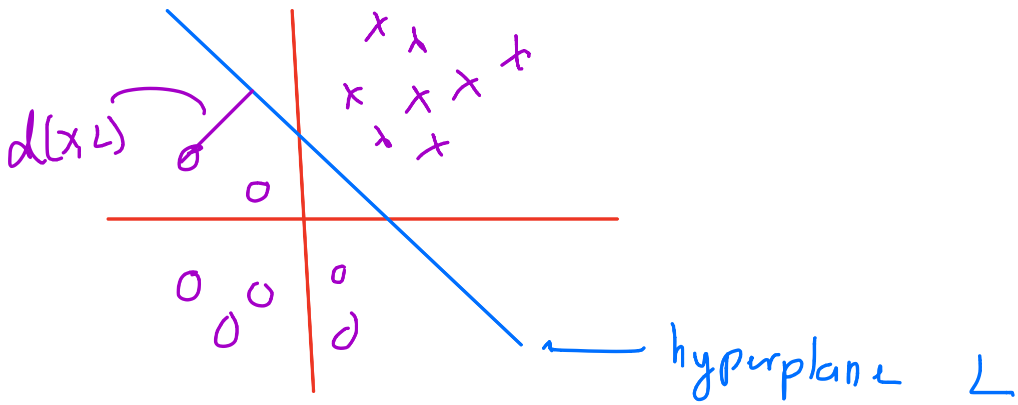
Seperating Hyperplanes



Assumption: The data are perfectly separable.

$$f(x) = \beta_0 + \beta^T x$$

Classifier: $\hat{y} = \begin{cases} +1 & f(x) > 0 \\ -1 & f(x) < 0 \end{cases}$

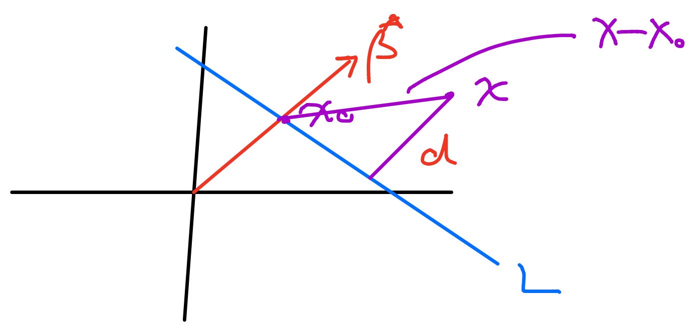
Methods LDA, Perceptron, Maximum margin, SVM,

Loss function:

Define $\mathcal{M} = \{ y_i \text{ missclassified by } f \} \subseteq \mathcal{T}$

$$\text{Loss}(f) = D(\beta_0, \beta) = \sum_{i \in M} d(x_i, L)$$

Goal: Minimize loss. $\nabla f(x) = \beta$ which is perpendicular to L.



$$\beta^* = \frac{\beta}{\|\beta\|} \qquad d = (x - x_0)^T \beta^*$$

Rmk: $(x_1 - x_0) \cdot \beta = 0$

Rmk: $x_0 \in L \Rightarrow f(x_0) = 0$

$$\Rightarrow \beta_0 + \beta^T x_0 = 0$$

$$\Rightarrow \beta^T x_0 = -\beta_0$$

$$d(x, L) = \beta^{*T}(x - x_0) = \frac{1}{\|\beta\|} \beta^T (x - x_0)$$

$$= \frac{1}{\|\beta\|} (\beta^T x - \beta^T x_0) = \frac{1}{\|\beta\|} (\beta^T x + \beta_0)$$

$$= \frac{1}{\|\beta\|} |f(x)|$$

Therefore $f(x) = \|\beta\| \, L(x, L)$

$$\text{Loss}(f(x)) = \frac{1}{|\beta|} \sum_{m} f(x) \equiv L(x).$$

How to minimize this function?

Gradient descent.

$$x_i = x_{i-1} + \gamma \nabla f(x)$$

to derive this solution.

Now,

$$D(\beta_0, \hat{\beta}) = \sum_{i \in M} -Y_i (\beta_0 + \beta^T x) \cdot \frac{1}{\|\beta\|}$$

and applying gradient descent

Solves the problem.

As $\frac{1}{\|\beta\|}$ will eventually converge to a

constant, look to minimize

$$D^*(\beta_0, \beta) \equiv \sum_{i \in M} -Y_i (\beta_0 + \beta_i^T X_i)$$

$$\frac{d D^*(\beta_0, \beta)}{d\beta} = -\sum_{i \in M} Y_i X_i$$

$$\frac{d D^*(\beta_0, \beta)}{d\beta_0} = -\sum_{i \in M} Y_i$$

$$\nabla D^* = \left(-\sum_{M} y_i x_i \, , \, -\sum_{M} y_i\right)$$

Approximate the gradient of D.

Rmk: Perceptron method.

Thm: (Perceptron Convergence)

1. Finishes in finite time if perfectly seperable

2. May run indefinitely if not.

Maximum Margin Classifier

Assume perfectly seperable.

Goal: find $f(x) = \beta_0 + \beta^T x$

$$(\beta_0, \beta_1) = \underset{\{\beta, \beta_0 : \, y_i f(x_i) \geq 1\}}{\operatorname{argmax}} \frac{1}{\|\beta\|}$$

Now notice

$$\{\beta_0,\beta_1 : y_i \cdot f(x_i) \geq 1\} = \{\beta_0,\beta_1 : y_i \cdot \frac{f(x_i)}{\|\beta\|} \geq \frac{1}{\|\beta\|}\}$$

$$= \{\beta_0,\beta_1 : d(x_i, L) \geq \frac{1}{\|\beta\|}\}$$

$= \{\beta_0,\beta_1 :\ Y_i$ correctly classified $\forall i\}$.

So in essense we are maximizing all distances $\{d(x_i, L)\}_{i=1}^{N}$ w.r.t. $L_{(\beta_0,\beta_1)}$ that classify all points correctly.