SUPERVISED LEARNING

Input: $\{x_1, \ldots, x_n\}$ $x_i \in \mathbb{R}^d$ ← Features / Attributes

$\{y_1, \ldots, y_n\}$ ← LABELS (supervision)

- → $\{+1, -1\}$, $\{0, 1\}$ → BINARY CLASSIFICATION (spam vs non-spam)
- → $\{0, 1, \ldots, K\}$ → MULTICLASS CLASSIFICATION [digit classification]
- → $\in \mathbb{R}$ → REGRESSION [Rainfall prediction]

REGRESSION

INPUT / TRAINING DATA $\{x_1, \ldots, x_n\}$ $x_i \in \mathbb{R}^d$

$\{y_1, \ldots, y_n\}$ $y_i \in \mathbb{R}$.

Goal: Learn $h : \mathbb{R}^d \to \mathbb{R}$

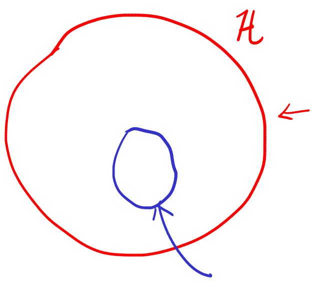
- How do we measure "goodness" of a function $h : \mathbb{R}^d \to \mathbb{R}$
- $\text{error}(h) = \sum_{i=1}^{n} (h(x_i) - y_i)^2$
- How small can this error be? $\boxed{0}$
- which h achieves zero error?

$h(x_i) = y_i \quad \forall i$

What is the problem?

- By "memorizing", we can get zero error on training data
- What we care is about test performance.
- Impose "STRUCTURE" to reduce search space.

SIMPLEST STRUCTURE — LINEAR STRUCTURE.



set of all functions $h : \mathbb{R}^d \to \mathbb{R}$.

$$\mathcal{H}_{linear} = \left\{ h_w : \mathbb{R}^d \to \mathbb{R} \text{ s.t } h_w(x) = w^T x \text{ for all } w \in \mathbb{R}^d \right\}$$

GOAL:

$$\min_{h_w \in \mathcal{H}_{linear}} \sum_{i=1}^{n} \left(h_w(x_i) - y_i\right)^2$$

(or) equivalently.

$$\boxed{\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2} \rightarrow \text{LINEAR REGRESSION}$$

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2 = \underline{\| X^T w - y \|}_2^2$$

$$X^T = \underset{n \times d}{\begin{bmatrix} — x_1 — \\ — x_2 — \\ \vdots \\ — x_n — \end{bmatrix}} \underset{d \times 1}{\begin{bmatrix} | \\ w \\ | \end{bmatrix}} \qquad y = \underset{n \times 1}{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}$$

$$\min_{\underline{w \in \mathbb{R}^d}} (X^T w - y)^T (X^T w - y) \leftarrow \text{unconstrained Quadratic (in } w) \text{ optimisation problem}$$

Solution: Take derivative (gradient) and set to zero.

$$f(w) = (X^T w - y)^T (X^T w - y)$$

$$\nabla f(w) = 2(XX^T)w - 2(Xy)$$

Solution satisfies $\boxed{(\underline{XX^T})w^* = Xy}$

OBSERVATION

Like PCA, w^* depends on a "Covariance" like matrix. But it also involves y.

$$w^* = (XX^T)^{\dagger} X\underline{y}$$ Pseudo-inverse.

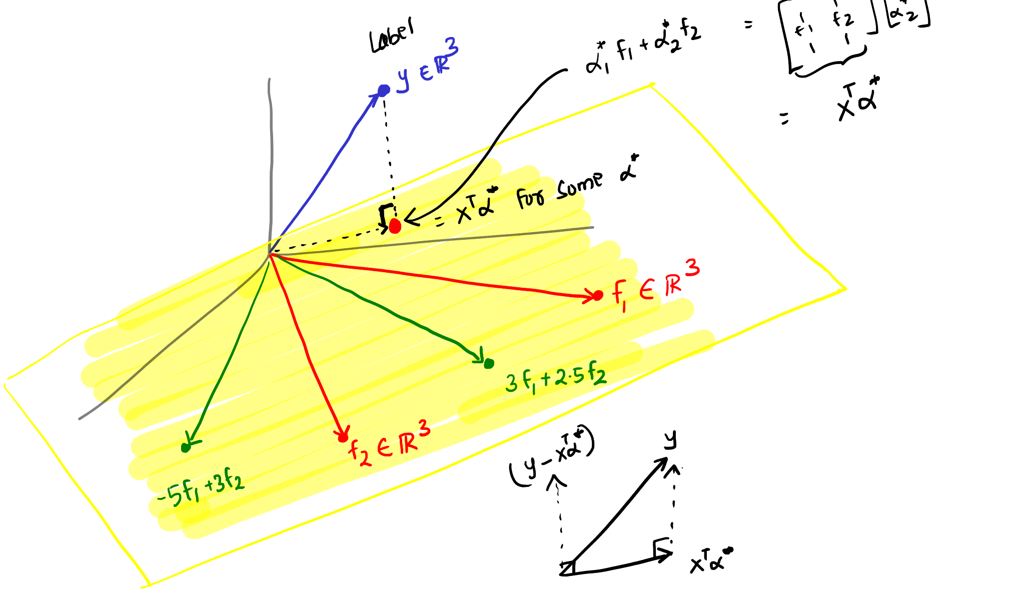
- → Lin reg has closed form solution
- → GEOMETRIC VIEW ?
- → COMPUTATIONAL CONSIDERATIONS ?
- → NON-LINEAR FEATURE → LABEL RELATIONSHIP ?
- → PROBALISTIC VIEW ?

$$w^* = (XX^T)^{\dagger} Xy$$ ← How can we interpret this geometrically?

	f_1	f_2	
	height	weight	
x_1			y_1
x_2			y_2
x_3			y_3

$d = 2$

$n = 3$



$$(y - X^T\alpha^*)^T (X^T\alpha^*) = 0$$

$$\underbrace{y^T X^T \alpha^* - \alpha^{*T}(XX^T)\alpha^*}_{} = 0 \quad \text{—— (1)}$$

Recall, $w^* = (XX^T)^{\dagger} Xy$

Substituting $w^* = \alpha^*$ on L.H.S, we get

$$y^T X^T\left((XX^T)^{\dagger} Xy\right) - \left((XX^T)^{\dagger} Xy\right)^T (XX^T)\left((XX^T)^{\dagger} Xy\right)$$

$$= 0 \quad \left[\text{use } XX^T \text{ is symmetric}\right]$$

CONCLUSION: $X^T w^*$ is the PROJECTION of the Labels onto the subspace spanned by the features.