What is machine learning

<u>Step 1</u>: Turn $\mathcal{X}$ into a vector space (feature space)

<u>Step 2</u>: Characterize a <u>loss</u>:

loss$(f(x), y)$ measures how well $f(x)$ predicts y.

(usually loss $(y, y) = 0$, loss $(z, y) \geq 0$)

<u>step 3</u>: pick a family $\mathcal{F}$ of functions that might predict y from x

<u>step 4</u>: collect data (examples) $(x_1, y_1), \dots, (x_n, y_n)$

<u>step 5</u>: $$\underset{f \in \mathcal{F}}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^{n} \text{loss}(f(x_i), y_i)$$

<u>step 6</u>: ? What do you do with the optimal f?

How does it perform on new data?

Example:

$$\mathcal{X} = \mathbb{R}^d, \quad \mathcal{Y} = \mathbb{R}$$

$$\text{loss}(f(x), y) = (f(x) - y)^2$$

$$\mathcal{F} = \{ f(x) = w^T x : w \in \mathbb{R}^d \}$$

$$\underset{w}{\text{Minimize}} \ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} (w^T x_i - y_i)^2$$

LEAST SQUARES

We can rewrite this, defining

$$\bar{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Then cost is $\frac{1}{2n} \| X w - y \|^2$

$$\|v\| = \left(\sum_{i=1}^{n} v_i^2 \right)^{1/2} \quad \text{is Euclidean norm}$$

Least-squares solution

I. Orthogonal projection

V an inner product space

$S \subset V$ a subspace

Any $\vec{v} \in V$ can be decomposed as

$\vec{v} = \vec{v}_S + \vec{v}_\perp$ where $\vec{v}_S \in S$

$\langle \vec{v}_\perp, \vec{u} \rangle = 0 \quad \forall \ \vec{u} \in S$

$\bar{P}_S$ is the linear projection onto S

$\bar{P}_S(\vec{v}) = v_S$

FACT: $\| \vec{v} - \bar{P}_S \vec{v} \| \leq \| \vec{v} - \vec{u} \| \quad \forall \ \vec{u} \in S$

$$\Rightarrow \quad \min_{\vec{u} \in S} \| \vec{v} - \vec{u} \| = \| \vec{v} - \bar{P}_S \vec{v} \| \qquad \text{(definition of min)}$$

$$\arg\min_{\vec{u} \in S} \| \vec{v} - \vec{u} \| = \bar{P}_S \vec{v}$$

arg min = set of minimizers

"argument of the minimum"

PROOF OF FACT: $\| \vec{v} - \vec{u} \|^2 = \| \vec{v} - \bar{P}_S \vec{v} + \bar{P}_S \vec{v} - \vec{u} \|^2$

$= \| \vec{v} - \bar{P}_S \vec{v} \|^2 + \| \bar{P}_S \vec{v} - \vec{u} \|^2 \geq \| \vec{v} - \bar{P}_S \vec{v} \|^2$

Projections and Least-Squares

$$\min_{\vec{w}} \|\bar{X}\vec{w} - \vec{y}\| = \min_{\vec{v} \in \text{range}(\bar{X})} \|\vec{v} - \vec{y}\|$$

$$\arg\min_{\vec{v} \in \text{range}(\bar{X})} \|\vec{v} - \vec{y}\| = P_{\text{range}(\bar{X})}\vec{y}$$

Now we also know

$$P_{\text{range}(\vec{X})} y - \vec{y} \in \text{range}(\bar{X})^{\perp} = \text{null}(\bar{X}^T)$$

$$\bar{X}^T(P_{\text{range}(\bar{X})}\vec{y} - \vec{y}) = 0$$

$$\bar{X}^T(\bar{X}\vec{w}_{OLS} - \vec{y}) = 0$$

if $\bar{X}^T\bar{X}$ is invertible:

$$\vec{W}_{OLS} = (\bar{X}^T\bar{X})^{-1}\bar{X}^T y$$

only invertible if $d < n$ ...

$\bar{X}\vec{w}_{ols}$ always unique

$\vec{W}_{ols}$ not always unique.

Note: $\text{null}(\bar{X}) = \text{range}(\bar{X}^T)^{\perp}$. Therefore can assume $\vec{w}_{ols} = \bar{X}^{\perp}\alpha \longrightarrow \vec{w}_{ols} = \sum_{i=1}^{n} \alpha_i \vec{x}_i$

$\vec{w}_{ols}$ is in span of examples.

Optimization approach

$$\vec{w}_\star \in \arg\min \| \bar{X}\vec{w} - \vec{y} \|^2$$

means

$$\| \bar{X}(\vec{w}_\star + \Delta\vec{w}) - \vec{y} \|^2 \geq \| \bar{X}\vec{w}_\star - \vec{y} \|^2 \quad \forall \ \Delta\vec{w}$$

expanding terms, this means

$$\langle \bar{X}\vec{w} - \vec{y}, \bar{X}\Delta\vec{w} \rangle + \| \bar{X}\Delta\vec{w} \|^2 \geq 0$$

$$\Rightarrow \langle \bar{X}^T(\bar{X}\vec{w} - \vec{y}), \Delta\vec{w} \rangle + \| \bar{X}\Delta\vec{w} \|^2 \geq 0$$

$$\Rightarrow \langle \bar{X}^T(\bar{X}\vec{w} - y), \frac{\Delta\vec{w}}{\|\Delta\vec{w}\|} \rangle + \frac{\| \bar{X}\Delta\vec{w} \|^2}{\|\Delta w\|} \geq 0$$

again implying

$$\bar{X}^T(\bar{X}\vec{w} - \vec{y}) = 0$$