<u>Recap</u>: With (pro. over $S \sim \mathcal{D}^m$) $\geq 1-\delta$

$$R(h_S) \leq \frac{1}{m} \log \frac{|\mathcal{H}|}{\delta} \qquad \mathcal{D}^m(\ )$$

ERM over training S

$m = |S|$

$|\mathcal{H}|$: size of finite Hypothesis class $\mathcal{H}$.

$$R(h_S) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \ell((x,y), h_S)$$

More on generalization bounds after we study regression & kernel regression

Linear <u>regression</u>

$$h(x) = w^T x + b$$

$$\ell((x,y), h) = (w^T x + b - y)^2$$

$$\hat{R}_S(h) = \sum_{(x,y)\in S} \ell(x,y,h)$$

MSE

Solving ERM over linear models (w,b) is called linear regression

Want w, b s.t.

$\hat{R}_S(h)$ is mini

parameterized by w, b.

$$w \equiv [w, b] \in \mathbb{R}^{d+1}$$

$$x \equiv [x, 1] \in \mathbb{R}^{d+1}$$

$$\min_w \hat{R}_S(h) = \min_w \frac{1}{m} \sum_{(x,y)\in S} (y - w\cdot x)^2$$

WKT at minimum w^*,

$$\nabla_w \hat{R}_S(h) = 0$$

$$\Rightarrow \sum_{(x,y)\in S} (y - w\cdot x)\, x = 0 \quad (*)$$

Define $Y = [y_1, \dots, y_m]^T \in \mathbb{R}^m$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{m\times(d+1)}$$

$(*)$ if X^TX is invertible

$$\underline{X^T Y = X^T X w} \Rightarrow w = (X^TX)^{-1} X^T Y$$

$$X^T(Y - Xw) = 0 \in \mathbb{R}^{d+1}$$

Linear <u>algebra</u>

$$\min_w \hat{R}_S(w) = \min_w \frac{1}{m} \sum_{i=1}^{m} (y_i - w^T x_i)^2$$

$$= \min_w \frac{1}{m} \| Y - Xw \|^2$$

Least squares problem

<u>Case 1</u>: $Y = Xw \Rightarrow w = X^{-1} Y$

$d+1 = m$ and X is invertible

<u>Case 2</u>: $d+1 > m$ over parameterized

non-unique! solution for $Xw = Y$.

$$XX^T \in \mathbb{R}^{m\times m}$$

if X has full row rank (rank $= m$)

$$Y = Xw$$

$$\boxed{w = X^T (XX^T)^{-1} Y}$$ is optimal in norm

$$Xw = XX^T (XX^T)^{-1} Y = Y$$

ie min-norm solution. Why?

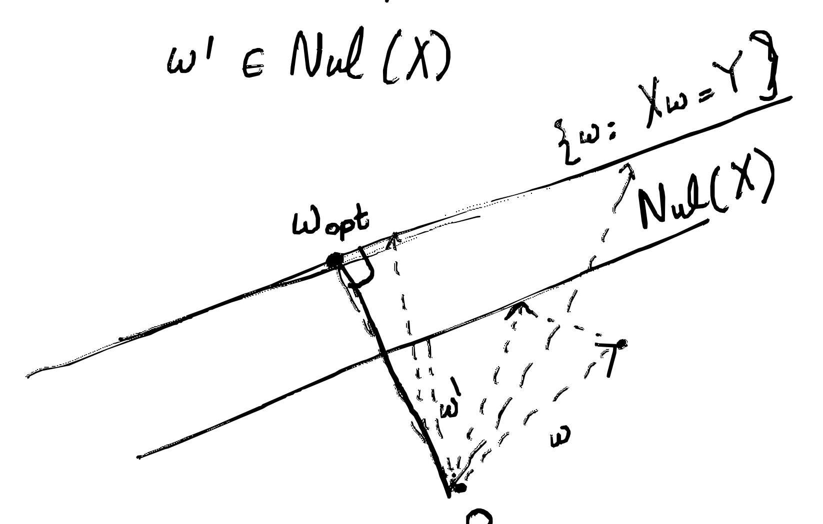
if

$$Xw = Y$$

$$X(w + w') = Y$$

$w' \in \mathrm{Nul}(X)$

$\mathrm{Nul}(X) = \{ w \in \mathbb{R}^{d+1} : Xw = 0 \in \mathbb{R}^m \}$



Ordinary least_squares estimator

$$w_{opt} \perp \mathrm{Nul}(X)$$

$$X w_{opt} = Y$$

$$\mathrm{Nul}(X) \perp \mathrm{Ran}(X^T)$$

Min-norm solution

$$w_{opt} = X^T z$$

$$X w_{opt} = XX^T z = Y$$

$$z = (XX^T)^{-1} Y$$

$$w_{opt} = \underline{X^T (XX^T)^{-1} Y}$$

$$w_{opt} \perp \mathrm{Nul}(X)$$

$$\Rightarrow w_{opt} \in \mathrm{Ran}(X^T)$$

when X is full row rank (rank $= m$)

$m < d+1$

<u>Case 3</u>: $m > d+1$ over determined

X has full column rank (rank $(d+1)$)

then X^TX is invertible and

$$w = (X^TX)^{-1} X^T Y \quad \text{(Normal equation solution)}$$

also solves $Xw = Y$ if $Y \in \mathrm{Ran}(X)$

More generally, $w = (X^TX)^{-1} X^T Y$

is only OLS solution

(see optimization approach above)

Soln to linear regression

$h(x) = \omega^T x$ linear model

$x \to \Phi(x)$

$$h(x) = \omega^T \Phi(x) + b$$

$$= \langle \omega, \Phi(x) \rangle \to \text{inner product in function space}$$

$$\omega = (X^T X)^{-1} X^T Y$$

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix}$$

$$= \begin{bmatrix} \Phi(x_1)^T \\ \vdots \\ \Phi(x_m)^T \end{bmatrix}$$

$$\underset{m \times m}{X^T X_{ij}} = \langle \Phi(x_i), \Phi(x_j) \rangle$$

$$= \mathcal{K}(x_i, x_j) \quad \text{for some PD kernel}$$

$$\underset{m \times 1}{X^T Y} = \langle \Phi(x_i), \Phi(y_i) \rangle$$

To compute ω_{opt}, we only need to invert $(X^T X)$: Gram matrix

Linear regression $\to$ kernel regression

if $X^T X$ is not invertible

$$\omega_{opt} = (X^T X + \lambda I)^{-1} X^T Y$$

regularization parameter

$x \in \mathcal{X}$ $\qquad \Phi(x) \in \mathbb{R}^D$

d-dim. $\qquad$ or $\mathcal{H}$.

Many feature maps can be associated with the same kernel!

Reproducing kernel Hilbert space
RKHS

For any PD kernel $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,
$\exists$ a corresponding Hilbert space $\mathcal{H}$
s.t. for any $f \in \mathcal{H}$

$$f(x) = \langle f, \mathcal{K}(\cdot, x) \rangle_{\mathcal{H}}$$

Reproducing property

- Consider some arbitrary data points $x_1, x_2, \ldots, x_m$
- $\text{span}\{\mathcal{K}(\cdot, x_i)\}$
- $f, g \in \text{span}\{\mathcal{K}(\cdot, x_i)\}$

$$f(x) = \sum_{i=1}^m \alpha_i \mathcal{K}(x, x_i) \leftarrow$$

$$g(x) = \sum_{i=1}^m \beta_i \mathcal{K}(x, x_i)$$

Define

$$\langle f, g \rangle := \sum_{i,j=1}^m \alpha_i \beta_j \mathcal{K}(x_i, x_j)$$

Linearity : Bilinear

$$\langle f, g \rangle = \sum_{j=1}^m f(x_j) \beta_j$$

$$= \sum_{i=1}^m \alpha_i g(x_i)$$

$$\langle f_1 + f_2, g \rangle = \sum_{i=1}^m (\alpha_i^{(1)} + \alpha_i^{(2)}) g(x_i)$$

$$= \sum_{i=1}^m \alpha_i^{(1)} g(x_i) + \sum_{i=1}^m \alpha_i^{(2)} g(x_i)$$

$$= \langle f_1, g \rangle + \langle f_2, g \rangle$$

Positive definiteness

$$\langle f, f \rangle = \sum_{i,j=1}^m \alpha_i \alpha_j \mathcal{K}(x_i, x_j)$$

$$= \alpha^T G \alpha \geq 0$$

(G is Gram matrix, which is PSD for PD kernel $\mathcal{K}$)

if $f = 0$, $\langle f, f \rangle = 0$
if $\langle f, f \rangle = 0$ then $f = 0 \in \mathcal{H}$.

$\mathcal{H}$: $\overline{\text{span}(\{\mathcal{K}(\cdot, x)\}_{x \in \mathcal{X}})}$

To show: completion under $\langle \cdot, \cdot \rangle$

$$\langle \mathcal{K}(\cdot, x_i), \mathcal{K}(\cdot, x_j) \rangle_{\mathcal{H}}$$

$$= \mathcal{K}(x_i, x_j)$$

$$\rightarrow |\langle \mathcal{K}(\cdot, x_i), \mathcal{K}(\cdot, x_j) \rangle|_{\mathcal{H}}^2 \leq \|\mathcal{K}(\cdot, x_i)\|^2 \|\mathcal{K}(\cdot, x_j)\|$$

PD kernel

$$\|\mathcal{K}(\cdot, x_i)\|^2 := \langle \mathcal{K}(\cdot, x_i), \mathcal{K}(\cdot, x_i) \rangle$$

$\langle , \rangle_{\mathcal{H}}$ inner product satisfies CS.

Next : RKHS feature map