HW2 : Regression using kernels

Presentation paper] Last 2 Wed of "teaching period"]
Project (presentation + 1-page proposal + report (5 pages))

Regression

Linear regression using linear model

$$h_\omega(x) = x \cdot \omega$$

$\{(x_i, y_i)\}_{i\in[m]}$ iid from $\mathcal{D}$

Solve ERM: $\min_{\omega \in \mathbb{R}^d} \frac{1}{m} \sum_{i\in[m]} \| y_i - h_\omega(x_i)\|^2$

$= \min_{\omega \in \mathbb{R}^d} \| X\omega - Y\|^2$

$X \in \mathbb{R}^{m\times d}$ $Y \in \mathbb{R}^m$

$$X = \begin{bmatrix} x_1 \\ \hline x_2 \\ \hline \vdots \\ \hline x_m \end{bmatrix}$$

$(\nabla_\omega \| X\omega - Y\|^2 = 0)$

Solution using optimization:

$$\omega = (X^TX)^{-1}X^TY$$

(assuming that X^TX is invertible)

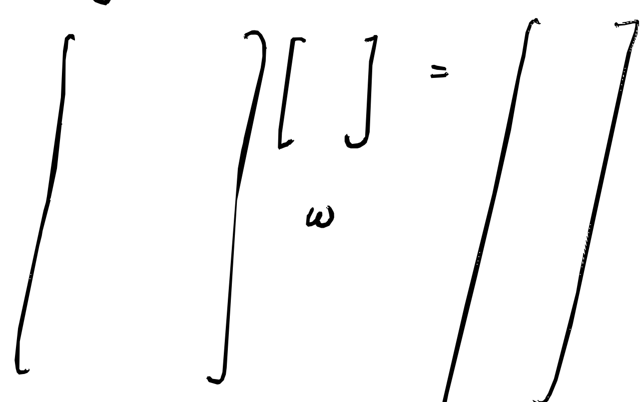
Interpretation of LS regression from linear algebraic viewpoint

- when X is square, $m = d$ and X is invertible,

$$\omega = X^{-1}Y$$

$$X^{-1}Y = (X^TX)^{-1}X^TY$$

- $m > d$ overdetermined case: $X\omega = Y$ may not have a solution!

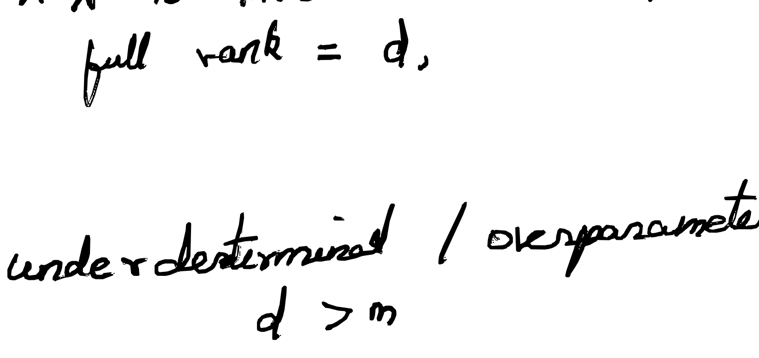


But, $X\omega = Y$ has a solution when Y is in the $\mathrm{Ran}(X)$

$$X^TX\omega = X^TY \Rightarrow \omega = (X^TX)^{-1}X^TY$$

if X^TX is invertible, that is X has full rank $= d$,

- underdetermined / overparameterized $d > m$

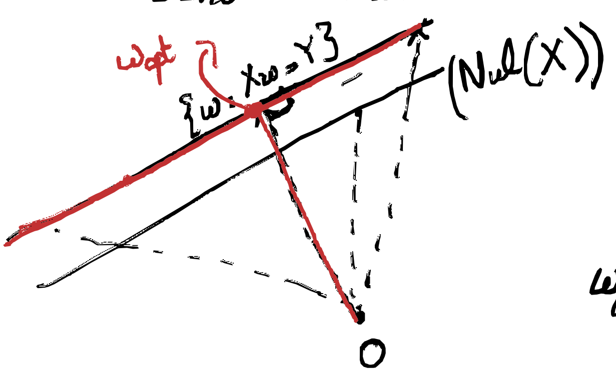


when does $X\omega = Y$ have a solution?
when X^T has full rank $= m$.

Is solution unique? No.

if $z \in \mathrm{Nul}(X)$ and ω solves $X\omega = Y$, then $X(\omega + z) = Y$

Mostly, we want "minimum norm" solution for ω.



$$X\omega = Y$$

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

$$\omega_{opt} = X^Tz$$

$$X\omega = XX^Tz = Y$$

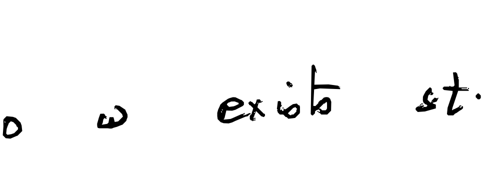
$$z = (XX^T)^{-1}Y$$

(XX^T is invertible)
when X has full row rank $= m$, XX^T is positive def

$$\omega_{opt} = X^Tz = X^T(XX^T)^{-1}Y$$

if no ω exists st. $X\omega = Y$

$$\arg\min_\omega \|X\omega - Y\|^2 = (X^TX)^{-1}X^TY$$



Want: to replace x with some "features" $\Phi(x)$

$$X = \begin{bmatrix} \Phi(x_1) \\ \hline \vdots \\ \hline \Phi(x_m) \end{bmatrix}$$

Solution of LS regression:

$$\omega = (X^TX)^{-1}X^TY$$

o/p fn: $h_\omega(x) = \omega \cdot \Phi(x)$

Questions:
- what are features $\Phi(x)$?
- why/when is lifting linear regression to feature space effective?

$h_\omega(x)$: still a linear function on feature space (of $\Phi(x)$) but can be complicated nonlinear fn on data space ($\mathcal{X}$)

Reproducing kernel Hilbert space

RKHS

$\mathcal{X} \subseteq \mathbb{R}^d$ compact

Theorem: If $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is PD, $\exists!$ RKHS ($\mathcal{H}$) which is defined by the following properties: $\| \ \|^2 = \langle \ , \ \rangle$

- $\exists$ some inner product s.t.
$$\langle \kappa(x, \cdot), \kappa(x', \cdot) \rangle = \kappa(x, x')$$
- $$\mathcal{H} = \overline{\text{span}\{\kappa(x, \cdot) : x \in \mathcal{X}\}}$$
- for any $f \in \mathcal{H}$, $\langle f, \kappa(x, \cdot) \rangle = f(x)$

Recall: κ is PD kernel if Gram matrix $G_{ij} = \kappa(x_i, x_j)$ $i, j \in \mathbb{N}$ is SPSD

Reproducing property

Evalution functional

$$E_x : \mathcal{H} \to \mathbb{R} \qquad E_x \in \mathcal{H}^*$$

$$E_x f = f(x)$$

$$E_x(f + g) = f(x) + g(x) = E_x f + E_x g$$

Linearity ✓

$$|E_x f| = |f(x)| \overset{\text{suppose}}{<} \infty \qquad \text{(e.g. } f \text{ is continuous, } \mathcal{X} \text{ is compact)}$$

if E_x (evaluation functional) at every x is bounded and linear, then, (or $|E_x f| < C \|f\|_{\mathcal{H}}$), Riesz representation theorem holds

So, $\exists!$ $g_x \in \mathcal{H}$ st.
$$\langle f, g_x \rangle_{\mathcal{H}} = E_x f = f(x).$$

$$g_x \equiv \kappa(x, \cdot)$$

$\mathcal{X}$

Proof:

Want: there is an RKHS corresponding to a PD κ.

$$\mathcal{H}_0 = \text{span}\{\kappa(x_i, \cdot)\}_{i \in \mathcal{I}} \qquad \mathcal{I} \text{ is finite}$$

$$f \in \mathcal{H}_0 \Rightarrow f(x) = \sum_{i \in \mathcal{I}} \alpha_i \kappa(x_i, x)$$

Define

$$\langle f, g \rangle := \sum_{i,j \in \mathcal{I}} \alpha_i \beta_j \kappa(x_i, x_j)$$

$$\text{if } f = \sum_i \alpha_i \kappa(x_i, \cdot) \quad g = \sum_i \beta_i \kappa(x_i, \cdot)$$

Want: show $\langle , \rangle$ is an inner product.

- bilinearity:
  - symmetric $\langle f, g \rangle = \langle g, f \rangle$

$$\langle f+g, h \rangle = \langle f, h \rangle + \langle g, h \rangle$$

$$\alpha \langle f, h \rangle = \langle \alpha f, h \rangle$$

- positive definite

$$\langle f, f \rangle \geq 0$$

and $\langle f, f \rangle = 0$ iff $f = 0$

$$\langle f, f \rangle \geq 0$$

$$\Rightarrow \sum_{i,j \in \mathcal{I}} \alpha_i \alpha_j \kappa(x_i, x_j) = \alpha^T \underbrace{G}_{\text{Gram matrix}} \alpha$$

≥ 0

because G is SPSD. ($\because \kappa$ is PD)

- if $f = 0$, then, $\langle f, f \rangle = 0$ ($\because \alpha_i = 0$).
- if $\langle f, f \rangle = 0$, then $f = 0 \in \mathcal{H}_0$.

$$\langle f, f \rangle = \sum \alpha_i \alpha_j \kappa(x_i, x_j)$$

Need to use Cauchy Schwartz for PD kernels!

Lemma: if κ is a PD kernel, then, $|\kappa(x, x')|^2 \leq |\kappa(x, x) \kappa(x', x')|$

Proof:

$$G = \begin{bmatrix} \kappa(x, x) & \kappa(x, x') \\ \kappa(x', x) & \kappa(x', x') \end{bmatrix}$$

$$\kappa(x, x') \cdot \kappa(x', x) = (\kappa(x, x'))^2 \quad (\because \kappa \text{ is sym})$$

$$\kappa(x, x) \kappa(x', x') - \kappa(x, x')^2 \geq 0.$$

$$\kappa(x, x') = \langle \phi(x), \phi(x') \rangle$$

$$|\langle \phi(x), \phi(x') \rangle|^2 \leq \|\phi(x)\|^2 \|\phi(x')\|^2$$

- Define $\tau(f, f) = \langle f, f \rangle$ is PD kernel on $\mathcal{H}_0$

$$G_{ij} = \tau(f_i, f_j) \qquad f_i, f_j \in \mathcal{H}_0.$$

$$\sum_{i,j} c_i c_j G_{ij} = \sum_{i,j} c_i c_j \tau(f_i, f_j)$$

$$= \sum_{i,j} c_i c_j \langle f_i, f_j \rangle$$

$$= \langle \sum_i c_i f_i, \sum_i c_i f_i \rangle$$

(Linearity of inner product) ≥ 0

So, G is SPSD.

- CS for τ.

$$|\tau(f, \kappa(x, \cdot))| \leq \tau(f, f)\, \tau(\kappa(x, \cdot), \kappa(x, \cdot))$$

$$|\langle f, \kappa(x, \cdot) \rangle| \leq \langle f, f \rangle \kappa(x, x)$$

Use reproducing property.

$$|f(x)| = |\langle f, \kappa(x, \cdot) \rangle|$$

Use CS for τ

$$|f(x)| = |\langle f, \kappa(x, \cdot) \rangle|$$

$$\leq \langle f, f \rangle \underbrace{\langle \kappa(x, \cdot), \kappa(x, \cdot) \rangle}_{\text{repr. prop.}}$$

$$= \langle f, f \rangle \kappa(x, x)$$

$\therefore$ if $\langle f, f \rangle = 0$, then, $|f(x)| = 0$ at all x

$$\Rightarrow f = 0 \in \mathcal{H}_0.$$

We have shown that

$$\langle f, g \rangle := \sum_{i,j} \alpha_i \beta_j \kappa(x_i, x_j)$$

where $f = \sum \alpha_i \kappa(x_i, \cdot)$

$g = \sum \beta_i \kappa(x_i, \cdot)$

is a valid inner product.

$\Rightarrow$ H_0 is a pre-Hilbert space

$Cl(H_0) \equiv \mathcal{H}$ is a Hilbert space

$H_0 \cup \{\text{limit points}\}$

H_0 is dense in $\mathcal{H}$

by Hahn-Banach theorem, reproducing prop. holds in $\mathcal{H}$.

Heine-Borel: closed + bounded sets of $\mathbb{R}^n \equiv$ compact

If there are 2 kernels κ, τ associated with $\mathcal{H}$,

$$\langle \kappa(x, \cdot), \tau(y, \cdot) \rangle = \tau(y, x)$$

$$= \kappa(y, x)$$

$\Rightarrow$ κ is unique.

Next time: Mercer map.

Reproducing property

$$\langle f, g \rangle = \sum_{i,j} \alpha_i \beta_j \, k(x_i, x_j)$$

if $g = k(x_j, \cdot)$ $\qquad f(x) = \sum_{i \in I} \alpha_i \, k(x_i, x)$

then,

$$\langle f, k(x_j, \cdot) \rangle = \sum_i \alpha_i \, k(x_i, x_j)$$

$$= f(x_j)$$

satisfied!