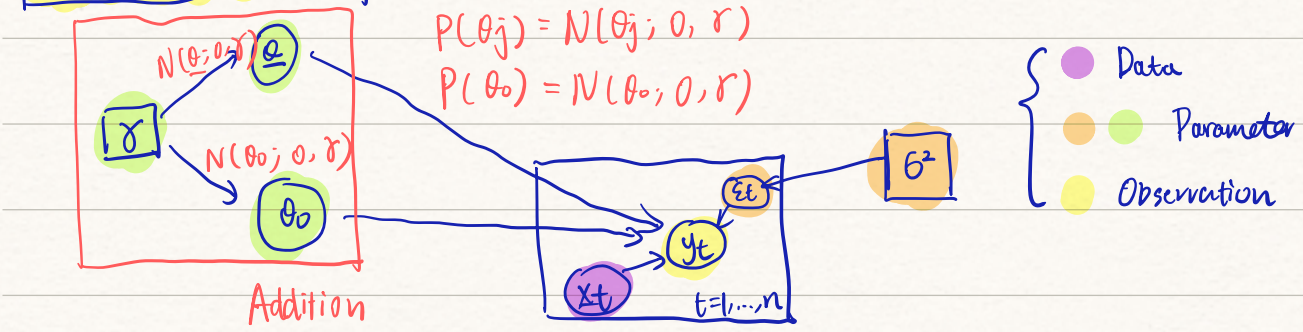
Graphical Model (Bayesian Network)



Generated Way $\Rightarrow$ $y_t = \underline{\theta}^T x_t + \theta_0 + \varepsilon_t \quad \varepsilon_t \sim N(0, \sigma^2)$

γ

Take $(\underline{\theta}, \theta_0)$ as random variables, we have PENALIZED log-likelihood.

$$\ell'(\underline{\theta}, \theta_0, \sigma^2, \gamma) = \underbrace{\text{const} - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t - \underline{\theta}^T x_t - \theta_0)^2}_{\text{Origin log-likelihood}}$$

$$\underbrace{- \frac{1}{2\gamma}(\theta_0^2 + \sum_{j=1}^{n}\theta_j^2) - \frac{d+1}{2}\log(\gamma)}_{\text{Penalty term (on } \theta_0 \text{ \& } \underline{\theta}\text{)}}.$$

let $\frac{\partial \ell'}{\partial \hat{\theta}} = 0 \Rightarrow \hat{\underline{\theta}} = \begin{bmatrix} \hat{\underline{\theta}} \\ \hat{\theta}_0 \end{bmatrix} = (X^T X + \lambda I)^{-1} X^T \underline{y}$.

$\hat{\theta} = [\underline{\theta}^T \ \theta_0]^T$

Solution.

Model: (Ground true model)

$$y_t = (\theta^*)^T x_t + \theta_0^* + \varepsilon_t$$

$$\Leftrightarrow y = X \begin{bmatrix} \underline{\theta}^* \\ \theta_0^* \end{bmatrix} + \varepsilon \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} x_1^T & 1 \\ \vdots & \vdots \\ x_n^T & 1 \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$\hat{\theta}$ $\qquad$ θ^*

{ Bias, Corvariance } of ESTIMATOR & Ground True Parameter

① Bias $= \mathbb{E}\left[\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix} - \begin{bmatrix}\theta^*\\ \theta_0^*\end{bmatrix} \middle| X\right]$

$$\mathbb{E}\left[\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix} \middle| X\right] = (X^TX+\lambda I)^{-1}X^T\,\mathbb{E}[y|X]$$
$$= (X^TX+\lambda I)^{-1}X^T\,\mathbb{E}\left[\left(X\begin{bmatrix}\theta^*\\ \theta_0^*\end{bmatrix}+\varepsilon\right)\middle| X\right]$$
$$= (X^TX+\lambda I)^{-1}X^TX\begin{bmatrix}\theta^*\\ \theta_0^*\end{bmatrix}$$

$$= (X^TX+\lambda I)^{-1}(X^TX+\lambda I-\lambda I)\begin{bmatrix}\theta^*\\ \theta_0^*\end{bmatrix}$$
$$= \begin{bmatrix}\theta^*\\ \theta_0^*\end{bmatrix} - \lambda(X^TX+\lambda I)^{-1}\begin{bmatrix}\theta^*\\ \theta_0^*\end{bmatrix} = \left(I-\lambda(X^TX+\lambda I)^{-1}\right)\begin{bmatrix}\theta^*\\ \theta_0^*\end{bmatrix}$$

Therefore, $\mathbb{E}\left[\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix} - \begin{bmatrix}\theta^*\\ \theta_0^*\end{bmatrix} \middle| X\right] = -\lambda(X^TX+\lambda I)^{-1}\begin{bmatrix}\theta^*\\ \theta_0^*\end{bmatrix}$

$\Leftrightarrow$ **Bias $= -\lambda(X^TX+\lambda I)^{-1}\begin{bmatrix}\theta^*\\ \theta_0^*\end{bmatrix}$**

Rmk: i) If $\lambda > 0$, there is a NON-ZERO Bias!

ii) $\mathbb{E}\left[\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix} \middle| X\right] := A\begin{bmatrix}\theta^*\\ \theta_0^*\end{bmatrix}$ $\quad A = I-\lambda(X^TX+\lambda I)^{-1}$

① Notice that $0 \preccurlyeq A \preccurlyeq I$.

→ A is positive semi-definite & all eigenvalue ≤ 1

that is to say $0 \leq \lambda_i(A) \leq 1$ $\quad i = 1, 2, \ldots, d+1$

② Notice that $\mathbb{E}\left[\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix} \middle| X\right] = A\begin{bmatrix}\theta^*\\ \theta_0^*\end{bmatrix}$

$\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix}$ is the Regularized solution

⇓

{ tends to shrink the parameters!
bringing them close to 0!

Intuition: Regularization makes $(\hat{\theta}, \hat{\theta}_0)$ smaller!

$\left[\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix} - (X^TX+\lambda I)^{-1}X^Ty = (X^TX+\lambda I)^{-1}X^TX\left[\begin{bmatrix}\theta^*\\ \theta_0^*\end{bmatrix}+\varepsilon\right]\right.$

② Covariance

$$\text{Cov}\left(\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix} \mid X\right) = \mathbb{E}\left[\left(\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix} - \mathbb{E}\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix}\right)\left(\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix} - \mathbb{E}\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix}\right)^T \mid X\right]$$

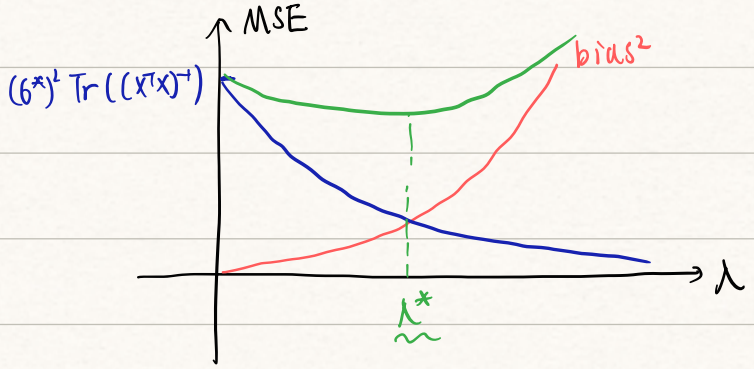
$$= (\sigma^*)^2 \left[(X^TX+\lambda I)^{-1} - \lambda (X^TX+\lambda I)^{-2} \right]$$

$(X^TX+\lambda I)^{-1} X^TX (X^TX+\lambda I)^{-1}$

③ $MSE = \|BIAS\|^2 + Variance$

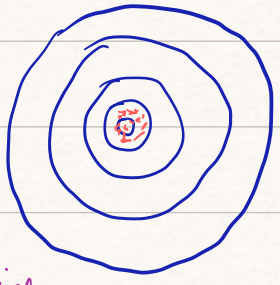
$= \|BIAS\|^2 + Tr(Covariance)$

$$\begin{cases} Bias = -\lambda (X^TX+\lambda I)^{-1} \begin{bmatrix}\theta^*\\ \theta_0^*\end{bmatrix} \\ Cov\left(\begin{bmatrix}\hat{\theta}\\ \hat{\theta}_0\end{bmatrix} \mid X\right) = (\sigma^*)^2 \left[(X^TX+\lambda I)^{-1} - \lambda (X^TX+\lambda I)^{-2} \right] \end{cases}$$

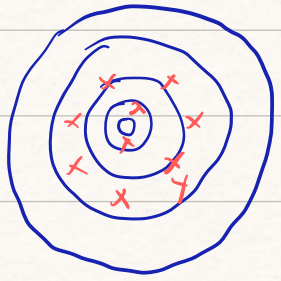


Bias-Variance

Low Bias & Low Variance



Low Bias, High Variance



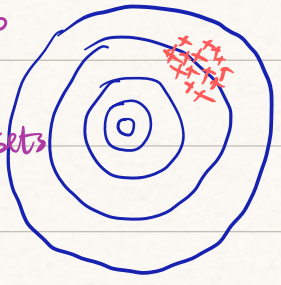
☆ Another example is polynomial curve fitting

- first order ✗
- 3rd order ✓ → n=8, N=20
- 5th order ✗

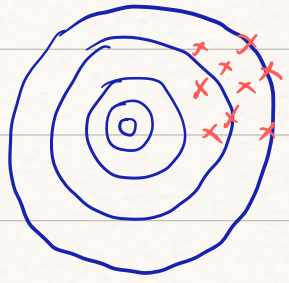
n = # of data points

N = # of datasets

High Bias, Low Variance

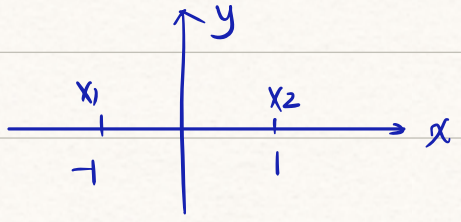


High Bias, High Variance



Example: $x_1 = -1$, $x_2 = +1$ $\quad d=1$

Design matrix $= \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} := X$



$X^T X = 2I = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$

$(\lambda I + X^T X) = (\lambda+2) I = \begin{pmatrix} \lambda+2 & 0 \\ 0 & \lambda+2 \end{pmatrix}$

$\| \text{bias} \|^2 = \left(\frac{\lambda}{2+\lambda}\right)^2 \left[(\theta^*)^2 + (\theta_0^*)^2 \right]$

$\text{Variance} = (\sigma^*)^2 \left(\frac{2}{2+\lambda} - \frac{2\lambda}{(2+\lambda)^2} \right)$

$\Rightarrow \text{MSE} = \frac{4(\sigma^*)^2}{(2+\lambda)^2} + \frac{\lambda^2}{(2+\lambda)^2} \left((\theta^*)^2 + (\theta_0^*)^2 \right)$

① If $\lambda = 0$, $\text{MSE} = (\sigma^*)^2$

② If $\lambda > 0$, and $(\sigma^*)^2 > (\theta^*)^2 + (\theta_0^*)^2$

choose $\lambda = 2$. $\quad \text{MSE}_{\lambda=2} < \frac{(\sigma^*)^2}{2} < \text{MSE}_{\lambda=0}$.

↓

USE Regularization to Reduce MSE!

Active Learning : USE MSE to select future inputs!

① Given $\begin{cases} x_1, x_2, \ldots, x_n \\ \updownarrow \quad \updownarrow \quad \quad \updownarrow \\ y_1, y_2, \ldots, y_n \end{cases}$ choose x_{n+1} so that

MSE on $\{(x_t, y_t)\}_{t=1}^{n+1}$ is small.

(linear model)

→ Assume $y = \langle x, \theta^* \rangle + \theta_0^* + \xi \qquad \xi \sim N(0, (\sigma^*)^2)$

→ Assume X is non-random, so only source of randomness is y (through ε)

$$MSE = (\sigma^*)^2 \, Tr\left((X^T X)^{-1}\right) \qquad X = \begin{bmatrix} x_1^T & 1 \\ \vdots & \vdots \\ x_n^T & 1 \end{bmatrix} \in \mathbb{R}^{n \times (d+1)}$$

Criterion: Try to select x_{n+1}, which is appended to X (in the last row) → $\tilde{X} = \begin{pmatrix} X \\ x_{n+1}^T \; 1 \end{pmatrix} \in \mathbb{R}^{(n+1) \times (d+1)}$

Calculate:

$$\tilde{X}^T \tilde{X} = \begin{bmatrix} X \\ x_{n+1}^T \; 1 \end{bmatrix}^T \begin{bmatrix} X \\ x_{n+1}^T \; 1 \end{bmatrix} = X^T X + \begin{bmatrix} x_{n+1} \\ 1 \end{bmatrix} \begin{bmatrix} x_{n+1} \\ 1 \end{bmatrix}^T$$

$$:= A^{-1} + VV^T$$

(inverse)

Notation: $\begin{cases} A = (X^T X)^{-1} \rightarrow \text{old Gram Matrix} \\ V = \begin{bmatrix} x_{n+1} \\ 1 \end{bmatrix} \rightarrow \text{new data. Vector} \end{cases}$

MSE'

New $MSE = (\sigma^*)^2 \, Tr\left((A^{-1} + VV^T)^{-1}\right)$

Goal: Find a "Valid" $x_{n+1} \in \mathbb{R}^d$ (i.e., $V \in \mathbb{R}^{d+1}$), such that MSE' is as small as possible.

Fact: $(A^{-1} + VV^T) = A - \frac{AVV^TA}{1 + V^TAV}$ (numerator: Matrix; denominator: scalar.)

$MSE' \propto Tr\left((A^{-1} + VV^T)^{-1}\right)$

New MSE $= \mathrm{Tr}(A) - \frac{1}{1+v^TAv}\mathrm{Tr}(Avv^TA)$

$= \mathrm{Tr}(A) - \frac{1}{1+v^TAv}\mathrm{Tr}(v^TAAv)$

$\mathrm{Tr}(A)$: old MSE

$\frac{1}{1+v^TAv}$: positive!

Fact: $v^TAv \geq 0$ for all v

Pf: $A = (X^TX)^{-1}$

X^TX is positive semi-definite

$\Rightarrow$ positive definite (inverse)

$\Rightarrow (X^TX)^{-1}$ positive definite

$\Leftrightarrow$ A positive definite

$\mathrm{Tr}(v^TAAv)$: nonnegative!

Fact: $v^TAAv \geq 0$ for all v

$\Leftrightarrow \|Av\|^2 \geq 0$

Rmk: for $\forall v \in \mathbb{R}^{d+1}$, $MSE' \leq MSE$!

Goal: Find X that maximizes the Reduction (of MSE)

$$\frac{v^TAAv}{1+v^TAv}$$

$$v = \begin{pmatrix} X \\ 1 \end{pmatrix}$$

$MSE_{original} = (\sigma^*)^2 \mathrm{Tr}((X^TX)^{-1}) = (\sigma^*)^2 \mathrm{Tr}(A)$

$MSE_{n+1} = (\sigma^*)^2 \left[\mathrm{Tr}(A) - \frac{v^TAAv}{1+v^TAv} \right]$

$= MSE_{origin} - (\sigma^*)^2 \cdot$ Reduction

Maximize this!

Note:

No matter how smart you are.

Claim: $\frac{v^T A A v}{1+v^T A v} \le \lambda_{max}(A)$ —— (*)

the reduction has an Upper Bound!

and the upper Bound in (*) is attained when $v = t u_1$ as $t \to \infty$.

u_1 is the eigen vector corresponding to $\lambda_{max}(A)$.

(i.e., the top eigen value of A).

Rmk: A is symmetric P.S.D. $(A \in \mathbb{R}^{(d+1)\times(d+1)})$

A { $\lambda_1 \to u_1$; $\lambda_{d+1} \to u_{d+1}$ }

① $\lambda_i \ge 0$. ② $\{u_i\}$ Orthonormal

Rmk: $\lim_{t\to\infty} \frac{v^T A A v}{1+v^T A v}\Big|_{v=tu_1} = \lambda_{max}(A)$

In conclusion, to max the reduction, choose v to be 'aligned' with u_1 and have the largest possible magnitude.

↳ For example, constrain $\|v\| \le C$ for some $0 < C < \infty$.

we choose ① $t = C$ ② $u_1 \to$ the normalized eigenvector (top)

$\Rightarrow v = t u_1$

Pf. $\{u_i\}_{i=1}^{d+1}$ forms as basis of A. (orthogonal eigenvectors)

Every $v \in \mathbb{R}^{d+1}$ can be written as $v = \sum_{i=1}^{d+1} \alpha_i u_i$

$$Av = \sum_i \alpha_i A u_i = \sum_i \alpha_i \lambda_i u_i$$

① $$v^T A v = \left(\sum_i \alpha_i u_i\right)^T \left(\sum_j \alpha_j \lambda_j u_j\right)$$ ← Orthogonal

$$= \sum_i \alpha_i^2 \lambda_i$$ ← $\begin{cases} u_i^T u_j = 0, & i \ne j \\ u_i^T u_j = 1, & i = j \end{cases}$

② $$v^T A A v = \sum_i \alpha_i^2 \lambda_i^2$$

then $\frac{v^TAAv}{1+v^TAv} \le \frac{v^TAAv}{v^TAv} \le \frac{\sum \alpha_i^2 \lambda_i^2}{\sum \alpha_i^2 \lambda_i} \le \frac{\lambda_1 \sum \alpha_i^2 \lambda_i}{\sum \alpha_i^2 \lambda_i} = \lambda_1.$

#

Example: $x_1 = -1 \quad x_2 = +1$

$y = \theta x + \theta_0 + \varepsilon \qquad x \in [-1, 1].$

$v = \begin{bmatrix} x \\ 1 \end{bmatrix} \qquad X = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \qquad X^TX = 2I.$

$$A = (X^TX)^{-1} = \frac{1}{2} I$$

① $v^TAv = \begin{pmatrix} x \\ 1 \end{pmatrix}^T \frac{1}{2} I \begin{pmatrix} x \\ 1 \end{pmatrix} = \frac{x^2+1}{2}$

② $v^TAAv = \frac{x^2+1}{4}$

Want to maximize the Reduction in MSE

$$\text{Reduction} = \frac{v^TAAv}{1+v^TAv} = \frac{\frac{1+x^2}{4}}{1+\frac{1+x^2}{2}} = \frac{1}{2}\frac{z}{1+z} \quad \left(z = \frac{1+x^2}{2}\right)$$

$\frac{z}{1+z} = 1 - \frac{1}{1+z}$ ↑ z

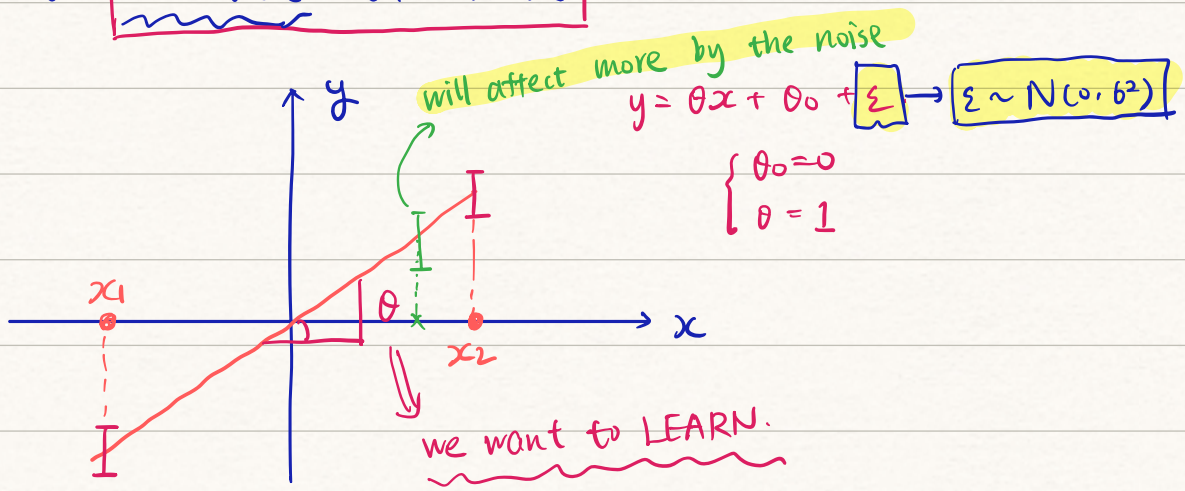
⇒ choose z as big as possible!

⇒ choose $|x|$ as large as possible!

→ Conclusion: choose x (next point) to be ± 1

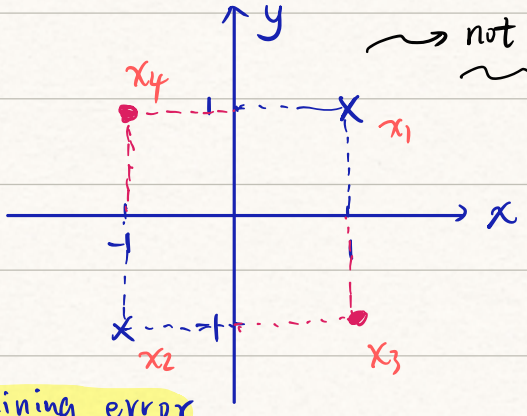
Rmk: For linear models inputs should be as far apart as

possible to **amortize the noise**



Introduction to KERNEL

XOR dataset



(affinely) not linearly separable!

⇓

✡ there exists NO AFFINE classifiers that attain 0 training error

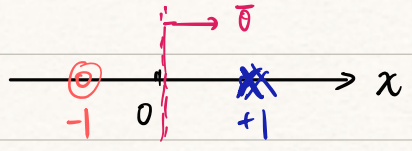
⇓

Construct a New Dataset (by multiplying the 2 coordinations together)

$X_1' = X_{11} X_{12} = 1$ $\quad X_2' = X_{21} X_{22} = 1$ ⇝ New Positive Label

$X_3' = X_{31} X_{32} = -1$ $\quad X_4' = X_{41} X_{42} = -1$ ⇝ New Negative Label

⇓



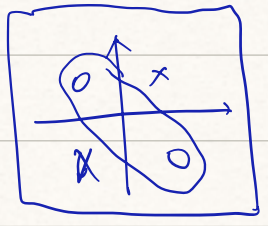
A decision boundary at $x=0$ attains 0 traing error!

1. Consider the linear model : $y = \theta x + \theta_0 + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$

① $\phi : x \in \mathbb{R} \longrightarrow \phi(x) = \begin{pmatrix} 1 \\ \sqrt{2}x \\ x^2 \end{pmatrix}$ ⇝ quadratic kernel

② $\phi : x \in \mathbb{R} \longrightarrow \phi(x) = \begin{pmatrix} 1 \\ \sqrt{3}x \\ \sqrt{3}x^2 \\ x^3 \end{pmatrix}$ ⇝ cubic kernel.

This RESULT IN a **polynomial regression model !**



$y = \langle \underline{\theta}, \underline{\phi}(x) \rangle + \theta_0 + \varepsilon$ ⇒ Actually, contains all polynomials whose order ≤ 3 (cubic kernel)

2. Polynomial expansion also works in Higher dimenson

↳ i.e., $\underline{X} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ ⇔ e.g. XOR dataset

$\underline{X} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 \xrightarrow[\text{quadratic kernel}]{\phi} \underline{\Phi}(\underline{X}) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \sqrt{2}x_1x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix} \in \mathbb{R}^{\boxed{6}}$ → govern by

- d : input dimension
- p : kernel function

Rmk: $\underline{\Phi}(\underline{X})$ contains all monomials whose orders ≤ 2 (as for quadratic kernel)

↳ defn. monomials order ≤ p.

⇔ $x_1^a x_2^b$ such that $a+b \leq p$.

Note: Dimension of $\underline{\Phi}(\underline{X})$ grows rapidly with (d, p)

Claim: $\dim(\underline{\Phi}(\underline{X})) = C^{p}_{d+p}$

pf: suppose that $\underline{x} \in \mathbb{R}^d$, namely $\begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$

and $x_1^{\lambda_1} \cdots x_d^{\lambda_d}$ find the nonnegative integer

solution $(\lambda_1, \ldots, \lambda_d)$ such that $\sum_{i=1}^{d} \lambda_i \le p$

$\Leftrightarrow \sum_{i=1}^{d} \lambda_i = k + d \longrightarrow (k)$ $\lambda_i \ge 1$, $\lambda_i \in \mathbb{Z}$. $k = 0, 1, 2, \ldots, p$.

the # of feasible solution for (k) is C_{k+d-1}^{d-1}

$$\Leftrightarrow \sum_{k=0}^{p} C_{k+d-1}^{d-1} = C_{d-1}^{d-1} + C_d^{d-1} + \cdots + C_{d+p-1}^{d-1}$$

$$= C_d^{d-1} + C_d^d + \cdots + C_{d+p-1}^{d-1}$$

$$= C_{d+1}^d + C_{d+1}^{d-1} + \cdots + C_{d+p-1}^{d-1}$$

$$= C_{d+2}^d + \cdots + C_{d+p-1}^{d-1}$$

$$= \cdots = C_{d+p}^d \qquad \square$$

Q: Should we first form the LONG FEATURE VECTOR $\underline{\phi}(\underline{x})$ [With dimension $\binom{d}{p+d}$], then take inner product?

$\langle \underline{\phi}(\underline{x}), \underline{\phi}(\underline{x}') \rangle$

Consider: $\underline{x}, \underline{x}' \in \mathbb{R}$, & $\underline{\phi}: x \in \mathbb{R} \longrightarrow (1, \sqrt{3}x, \sqrt{3}x^2, x^3)^T$

then $\langle \underline{\phi}(x), \underline{\phi}(x') \rangle = \begin{pmatrix} 1 \\ \sqrt{3}x \\ \sqrt{3}x^2 \\ x^3 \end{pmatrix}^T \begin{pmatrix} 1 \\ \sqrt{3}x' \\ \sqrt{3}x'^2 \\ x'^3 \end{pmatrix}$

$= 1 + 3xx' + 3(x'x)^2 + (x'x)^3$

$= (1 + x'x)^3$!

Observation : if we denote $x'x = \langle x, x' \rangle_{\mathbb{R}} = z$

$\langle \phi(x), \phi(x') \rangle_{\mathbb{R}^4} = (1+z)^3$

Therefore. Can we evaluate the inner product $\langle x, x' \rangle_{\mathbb{R}}$ first, then passing it to the inner product $\langle \underline{\phi}(x), \underline{\phi}(x') \rangle_{\mathbb{R}^4}$ through a simple function (like $f(z) = (1+z)^3$)

if we : ① $\langle \underline{x}, \underline{x}' \rangle_{\mathbb{R}^d}$

② $\langle \underline{\phi}(\underline{x}), \underline{\phi}(\underline{x}') \rangle_{\mathbb{R}^\infty} = f(\langle x, x' \rangle)$

difficult to calculate directly ! ($O(C_{d+p}^p)$)

Naive Computation

Smart Computation !

easy to calculate ! (only takes time $O(d)$)