Kernel trick & Kernel functions

$\phi(x) \rightarrow$ represents a transformation

$x \in \mathbb{R}^d$

$\phi(x) : \mathbb{R}^d \longrightarrow \mathbb{R}^n \;|\; n > d$

eg: $x \in \mathbb{R}^3$

$x = [x_1, x_2, x_3]^T$

$$\phi(x) : \mathbb{R}^3 \text{ to } \mathbb{R}^5 = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \\ \sqrt{x_1 x_2 x_3} \\ x_3^2 x_1 x_2 \end{bmatrix}$$

→ We need a method to see similarity of 2 vectors in transformed space, without explicitly transforming both vectors

→ Idea is to develop a kernel function $K(x, x')$ that returns the dot product of transformed vectors $\phi(x)^T \phi(x')$ without the computational complexity of transforming these two vectors into higher dimension

$K(x, x')$ — similarity metric

$= \phi(x)^T \phi(x')$

x, x' are two vectors in original space

$\phi(x) \rightarrow$ Transformation

$K(x, x') \rightarrow$ Kernel function

Dot product is a metric of similarity of two vectors

Calculating dot product in n dimension → n multiplication, ↳ $n-1$ addition

Calculating dot product of vectors in n dimension using Kernel trick } Generally very less

eg:- Original space $\mathbb{R}^2$

$x = (x_1, x_2)$

$z = (z_1, z_2)$

Let x, z be 2 vectors in $\mathbb{R}^2$

Example $K(x, z) = (1 + x^T z)^2$

$= (1 + x_1 z_1 + x_2 z_2)^2$

$= 1 + 2(x_1 z_1 + x_2 z_2) + x_1^2 z_1^2 + x_2^2 z_2^2 + 2 x_1 z_1 x_2 z_2$

$$K(x, z) = \begin{bmatrix} 1 & \sqrt{2} x_1 & \sqrt{2} x_2 & x_1^2 & x_2^2 & \sqrt{2} x_1 x_2 \end{bmatrix}^T \begin{bmatrix} 1 \\ \sqrt{2} z_1 \\ \sqrt{2} z_2 \\ z_1^2 \\ z_2^2 \\ \sqrt{2} z_1 z_2 \end{bmatrix}$$

$K(x, z) = \phi(x)^T \phi(z)$

Total multiplication $= 6$

Total addition $= 5$

$\phi(x) \in \mathbb{R}^6$

#Calculation

→ dot product in 2d = 2 multiplication, 1 addition

→ 1 addition, 1 multiplication $(1+k)^2$

Total 3 multiplication, 2 addition

Validity of Kernel function

1) $K(x,z) = \phi^T(x)\,\phi(z)$ $x, z \in \mathbb{R}^d$

$\phi(x), \phi(z) \rightarrow$ Same transformation applied to both datapoints

2) Mercer's Theorm

2.a) $K(x,z) = K(z,x)$ $\rightarrow$ Symmetric

2.b) Kernel matrix —
$$\begin{bmatrix} K(x_1,x_1) & K(x_1,x_2) \\ K(x_2,x_1) & K(x_2,x_2) \end{bmatrix}$$
→ Has to be positive semidefinite (PSD)
→ Symmetric
→ All eigen values $\lambda_i \geq 0$

→ In exam if Kernel matrix given → then check PSD → eigen value check
→ Trace & determinant check

→ If not given → check if kernel function can be represented as dot product of two symmetric transformations $\phi(x)^T\phi(z)$

Polynomial Kernel function

$$K(x,z) = (x^Tz + c)^d$$

c = constant
d = degree of polynomial

$x \in \mathbb{R}^m$

#Terms in expansion = dim of transformed space = ${}^{d+m}C_m$

$$(x^Tz+c)^d = (\underline{x_1z_1 + x_2z_2 + \dots x_mz_m + c})^d$$

$m+1$ → terms in each product

d products

$$= (x_1z_1 + x_2z_2 + \dots x_mz_m + c)(x_1z_1 + x_2z_2 + \dots x_mz_m + c) \dots\dots (x_1z_1 + x_2z_2 + \dots x_mz_m + c)$$

#ways to choose d items from $m+1$ items with replacement & order doesn't matter

$$\boxed{{}^{m+d}C_d = {}^{m+d}C_m}$$

if $c = 0$ $\boxed{{}^{m+d-1}C_d}$

Gaussian Kernel / Radial Basis Function (RBF) Kernel

① Map to ∞ dimension feature space

σ is hyper parameter

σ — Kernel width / Bandwidth
— Influence of individual training samples on the RBF decision boundary
→ Smoothness of DB

$$K(x,z) = \exp\left(-\frac{\|x-z\|_2^2}{2\sigma^2}\right)$$

$\boxed{\text{Sometime } \gamma = \frac{1}{2\sigma^2}}$

$\sigma > 0$

Intuition

Small σ → L2 dist scalles up $\Rightarrow e^{-\text{more}} \Rightarrow K(x,x') \approx 0$ unless x very close to x' → Narrow Influence

Large σ → L2 dist scalles down $\Rightarrow e^{-\text{less}} \Rightarrow K(x,x') \not\approx 0$ even for far away x' → Wide influence

Narrow Kernel
Rough DB

Small ← σ → large

more complexity

Wide Kernel
smooth DB
less complexity

Consider two data points, $x_1 = (1, -1)$ and $x_2 = (2, 2)$, in a binary classification task using an SVM with a custom kernel function $K(x, y)$. The kernel function is applied to these points, resulting in the following matrix, referred to as matrix A:

$$\begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) \\ K(x_2, x_1) & K(x_2, x_2) \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 3 & 6 \end{bmatrix}$$

Which of the following statements is correct regarding matrix A and the kernel function $K(x, y)$?

A) $K(x, y)$ is a valid kernel. ✓
B) $K(x, y)$ is not a valid kernel.
C) Matrix A is positive semi-definite. ✓
D) Matrix A is not positive semi-definite.

$\lambda^2 - 7\lambda - 3 = 0$

$\lambda = \frac{7 \pm \sqrt{49+12}}{2} = \frac{7 \pm \sqrt{61}}{2}$ one (+) one (−)

determinant −ve → one eigen value −ve

kernel matrix not PSD

Given a kernel function $K_1 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and its corresponding feature map $\phi_1 : \mathbb{R}^n \to \mathbb{R}^d$, which feature map $\phi : \mathbb{R}^n \to \mathbb{R}^d$ would correctly produce the kernel $cK_1(x, z)$, where c is a positive constant?

a) $\phi(x) = c\phi_1(x)$
b) $\phi(x) = \sqrt{c}\phi_1(x)$ ✓
c) $\phi(x) = c^2\phi_1(x)$
d) No such feature map exists.

Takes two n-dim vectors & calculates their dot product in higher dimension → scalar

$K(x,z) = \phi(x)^T \phi(z)$

$cK(x,z) = c\,\phi(x^T)\,\phi(z)$

$= \sqrt{c}\,\phi(x^T) \cdot \sqrt{c}\,(\phi z)$

maintain symmetry

Let $K_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be two symmetric, positive definite kernel functions. Which of the following *cannot* be a valid kernel function?

(a) $K(x, x') = 5 \cdot K_1(x, x')$
(b) $K(x, x') = K_1(x, x') + K_2(x, x')$
(c) $K(x, x') = K_1(x, x') + \frac{1}{K_2(x, x')}$ ✓
(d) All three are valid kernels.

a) $\phi(x) \to \sqrt{5}\,\phi(x)$

b) $K(x,x') = \phi_1^T(x)\,\phi_1(x') + \phi_2^T(x)\,\phi_2(x')$

$K(x,x') = [\phi_1(x) \;\; \phi_2(x)]^T \cdot [\phi_1(x') \;\; \phi_2(x')]$

c) Not possible to decompose — can be undefined when $K_2(x,x') = 0$

Result: ① When multiplied by a scalar → Valid kernel function
② When two kernel added → Valid kernel function

Given a kernel function $K(x, x') = f(x)g(x') + f(x')g(x)$, where f and g are real-valued functions ($\mathbb{R}^d \to \mathbb{R}$), the kernel is not valid. What additional terms would you include in K to make it a valid kernel?

Options:

(A) $f(x) + g(x)$
(B) $f(x)g(x) + f(x')g(x')$
(C) $f(x)f(x') + g(x)g(x')$ ✓
(D) $f(x') + g(x')$

☆☆☆☆☆

$f(x)\,g(x') + f(x')\,g(x)$

$f(x)\,g(x) + f(x')\,g(x')$

☆ options A, D can be eliminated using symmetric check.

Naive Bayes Risk (Misclassification probability)

1) Idea find joint $P(X_1, X_2 \dots X_n, Y) = P(x_1|y) \cdot P(x_2|y) \dots P(x_n|y) \cdot P(y)$

2) Calculate prediction for each input combination $\hat{y} = \underset{y}{\text{argmax}}\ P(X_1, X_2, X_3, \dots X_n, Y=y)$

$$\text{Risk} = 1 - \sum_{\substack{\text{all possible}\\ \text{inputs}}} P(X_1, X_2, X_3 \dots X_n, Y=\hat{y})$$

$$\text{Risk} = \sum_{\substack{\text{all possible}\\ \text{inputs}}} P(X_1, X_2, X_3, \dots X_n, Y=\neg\hat{y})$$ for Binary classification

eg:-

$$P(Y=1)=0.4, \quad P(Y=0)=0.6.$$

The class-conditional probabilities are:

$$P(X_1=1 \mid Y=1)=0.8, \quad P(X_1=1 \mid Y=0)=0.3$$
$$P(X_2=1 \mid Y=1)=0.7, \quad P(X_2=1 \mid Y=0)=0.2.$$

Which of the following statements is/are CORRECT?
(Select all that apply.)

A. For an observation $(X_1, X_2)=(1,1)$, the Naive Bayes classifier predicts $Y=1$.

B. For an observation $(X_1, X_2)=(0,1)$, the Naive Bayes classifier predicts $Y=0$.

C. The posterior probability $P(Y=1 \mid X_1=1, X_2=1)=\frac{28}{37}$.

D. The overall probability of misclassification under this Naïve Bayes model is less than 0.25.

$$Y=1: 0.4(0.2)(0.7)=0.056, \quad Y=0: 0.6(0.7)(0.2)=0.084$$

$\Rightarrow$ predict $Y=0 \rightarrow$ B✓

Posterior check:

$$P(Y=1 \mid 1,1)=\frac{0.224}{0.26}=\frac{56}{65} \neq \frac{28}{37}$$

$\rightarrow CX$

Overall correct prob:

$$0.224+0.084+0.144+0.336=0.788$$
$$P(\text{ misclassification })=1-0.788=0.212<0.25$$

$\rightarrow$ D✓

MLE argument for Least Square error & Absolute Error

$$y = f(x) + \epsilon$$

$$\epsilon = f(x) - y \qquad \epsilon \sim N(0, \sigma^2)$$

Trying to predict $\hat{y} = f(x)$ → Assumption → Gaussian error assumption. → True function $f(x)$ is linear

y – label

$\hat{y}$ – prediction

$$\epsilon = (\hat{y} - y) \longrightarrow \epsilon_i = (\hat{y}_i - y_i) \; : \; \hat{y}_i = x_i\theta$$

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} \left[P(\hat{y}_i - y_i / x_i, \theta) \right] = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\hat{y}_i - y_i}{\sigma} \right)^2 \right) \right]$$

Taking log likelihood

$$\theta^* = \underset{\theta}{\operatorname{argmax}} = -\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

→ Minimise least square

→ $\theta^* =$ mean y_i if only one feature $X = 1 \; \forall i$

Double Exponential Distribution (Laplace distribution assumption)

Assumption → $\epsilon \sim Laplace(0, b)$

$$f_\epsilon(\epsilon) = \frac{1}{2b} \exp\left(\frac{-|\epsilon|}{b} \right)$$

$$y = f(x) + \epsilon$$

$$\epsilon_i = (\hat{y}_i - y_i)$$

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} P(\hat{y}_i - y_i) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} \left[\frac{1}{2b} \exp\left(-\frac{|\hat{y}_i - y_i|}{b} \right) \right] = \underset{\theta}{\operatorname{argmax}} -\sum_{i=1}^{n} |\hat{y}_i - y_i|$$

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} |\hat{y}_i - y_i|$$

→ Minimise absolute error

→ $\theta^* =$ median of y_i if only one feature $X = 1 \; \forall i$
