Data Mining Lec 1: Kernel (K), random feature map (Φ), MSE

2020年9月12日 1:08

$$K_{Gauss}(x,y) = e^{-\frac{\|x-y\|^2}{2a^2}} = \left(e^{-\frac{\|x-y\|^2}{2}}\right)^{\frac{1}{a^2}}. \quad a\text{-hyperparameter.}$$

$$e^{-\frac{\|x-y\|^2}{2}} = e^{-\frac{\|x\|^2}{2}} \cdot e^{x^T y} \cdot e^{-\frac{\|y\|^2}{2}}$$

$$e^{x^T y} \downarrow \text{ softmax kernel } K_{sm}.$$

$$K_{sm}(x,y) = e^{x^T y}$$
$$= \sum_{i=0}^{+\infty} \frac{(x^T y)^i}{i!} \quad \text{(泰勒展开)}$$
$$= \sum_{i=0}^{\infty} \frac{\left(\sum_j^d x_j y_j\right)^i}{i!}$$
$$= \sum_{i=0}^{\infty} \left([\phi_i(x)]^T \phi_i(y) \right).$$

原因：对称性.

例：$d=2$, $i=2$, $(x_1y_1 + x_2y_2)^2 = x_1^2y_1^2 + x_2^2y_2^2 + 2x_1y_1x_2y_2$

$= \langle (x_1^2, \sqrt{2}x_1x_2, x_2^2), (y_1^2, \sqrt{2}y_1y_2, y_2^2) \rangle$.

$$K_{Gauss}(x,y) = \underset{w \sim iid[0, I_d]}{E} \left(\cos[w^T(x-y)] \right).$$

$$= E_w \left(\cos w^T x \cos w^T y + \sin w^T x \sin w^T y \right)$$

$$\phi(x) = (\cos w_1^T x, \text{--} \cos w_m^T x, \sin w_1^T x, \text{---} \sin w_m^T x)$$

$$\phi(y) = (\cos w_1^T y, \text{———} \cos w_m^T y, \quad \sin w_1^T y, \text{——} \sin w_m^T y).$$

$$w_1, \text{—} w_m \sim \text{i.i.d} (0, Id).$$

General structure of Φ:

$$\phi(x) = \frac{1}{\sqrt{m}} (f_1(w_1^T x) \text{.—} f_1(w_m^T x)$$

$$\text{———} \quad \text{——}$$

$$\text{——} \quad \text{——}$$

$$f_l(w_1^T x) \text{ --- } f_l(w_m^T x))^T. \quad (\text{length}: m \cdot l).$$

$f_1, \text{--} f_l$: $\mathbb{R} \to \mathbb{R}$ (如上述 sin, cos).

$w_1, \text{—} w_m \overset{iid}{\sim} \mathcal{D}$ (usually Gaussian distribution) $N(0, Id)$.

$K(x, y) = x^T y$.

$l = 1$ (只有一种 Identity 变换 $f(x) = x$).

$\mathcal{D} = N(0, Id)$ $\Rightarrow$

$$\Rightarrow \quad \phi(x) = \frac{1}{\sqrt{m}} (w_1^T x, \text{ ·— } w_m^T x)^T. = \frac{1}{\sqrt{m}} G x$$

Proof:

$$\phi(y) = \frac{1}{\sqrt{m}} (w_1^T y \text{ ——— } w_m^T y)^T.$$

$$K(x,y) = E_w (\phi(x)^T \phi(y))$$

$$= E_w \left(\frac{1}{m} (x^T w_1 w_1^T y, \text{.....}) \right)$$

$$= \frac{1}{m} \sum_{i=1}^{m} x^T y \, E(w_i^T w_i) \xrightarrow{} \mathbb{R}.$$

$$= x^T y.$$

$E(w_i^T w_i) = 1$

$$\sum_{i=1} \longrightarrow \mathbb{R}.$$

$$= x^T y.$$

$$G = \begin{bmatrix} w_1^T \\ \vdots \\ w_m^T \end{bmatrix}. \quad (\text{Gaussian Matrix.}).$$

$(m \times d)$.

w_i — wha $\overset{iid}{\sim} N(0, I^d)$

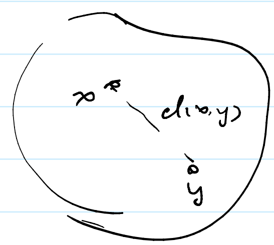
$x: (d \times 1)$.
$G: (m \times d)$

In practice, $m << d$., Dimensionality Reduction

$$\phi(x) = \frac{1}{\sqrt{m}} G \cdot x \quad : \quad (m \times 1).$$

G should be same to all x/y, — 对称

$$d^2(x, y) = \|x-y\|^2$$
$$= \langle x-y, x-y \rangle$$
$$= \langle x, x \rangle + \langle y, y \rangle - 2 \langle x, y \rangle.$$



$$K_{gauss}(x, y) = e^{-\frac{\|x-y\|^2}{2}}$$

$l = 2$, $f_1 = \sin$ $f_2 = \cos$, $\mathcal{D} = N(0, I_d)$.

$$K_{gauss}(x, y) = \Phi(x)^T \Phi(y).$$

$$\Phi(z) = \frac{1}{\sqrt{m}} (\sin w_1^T z, \ldots \sin w_m^T z, \cos w_1^T z, \ldots \cos w_m^T z).$$

$$K_{sm}(x, y) = e^{x^T y}.$$

$$= e^{\frac{\|x\|^2}{2}} \cdot e^{\frac{\|y\|^2}{2}} \cdot K_{gauss}(X, y)$$

$$= e^{\frac{\|x\|^2}{2}} \cdot e^{\frac{\|y\|^2}{2}} \cdot K_{Gauss}(x, y)$$

$$\Phi_{softmax}(z) \overset{def}{=} e^{\frac{\|z\|^2}{2}} \cdot \Phi_{Gauss}(z).$$

Angular Kernel.

$$K_{ang}(x, y) = 1 - \frac{2\theta_{x,y}}{\pi} \in [-1, 1], \quad \theta_{x,y} \in [0, \pi].$$

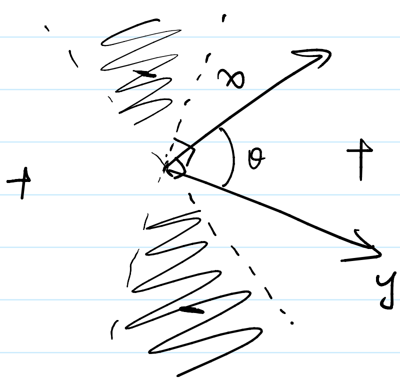
$l = 1, \quad f_1 = \text{sgn}. \quad \mathcal{D} = \mathcal{N}(0, I_d).$

$$\phi(x) = \frac{1}{\sqrt{m}} \left(\text{sgn}(w_1^T x) \;—\; \text{sgn}(w_m^T x) \right).$$

$$K_{ang}(x,y) = E_w \left(\phi(x)^T \phi(y) \right)$$

$$= E_w \left(\frac{1}{m} \cdot \sum_{i=1}^{m} \text{sgn}(w_i^T x)^T \cdot \text{sgn}(w_i^T y) \right).$$

为什么$f_1 = \text{sgn}$能表示 $x . y$ 夹角?



$P(\text{sgn}(w_i^T x) \cdot \text{sgn}(w_i^T y) = 1) \quad \frac{\pi - \theta}{\pi}.$

$P(\quad \cdot \text{——} \quad = -1) \quad \frac{\theta}{\pi}.$

$$\Rightarrow K_{ang}(x,y) = \frac{(\pi - \theta)}{\pi} \cdot 1 + \frac{\theta}{\pi} \cdot -1$$

$$= 1 - \frac{2\theta}{\pi}.$$

Back to dot product model,

$$K(x,y) = x^T y = E_w[\phi(x)^T \phi(y)].$$

$$= E\left[\left(\frac{1}{\sqrt{m}} G x\right)^T \left(\frac{1}{\sqrt{m}} G y\right)\right] \quad \text{is unbiased}$$

$= E[(\frac{1}{\sqrt{m}}Gx)^T(\frac{1}{\sqrt{m}}Gy)]$ is unbiased

Proof: $x' = \frac{1}{\sqrt{m}}Gx$, $y' = \frac{1}{\sqrt{m}}Gy$.

$$E[(\frac{1}{\sqrt{m}}Gx)^T(\frac{1}{\sqrt{m}}Gy)] = E[x'^T y'] = E[\sum_{i=1}^{m} x_i' y_i']$$

$$= \sum_{i=1}^{m} E(\underbrace{x_i' y_i'}_{A_i'}) = mE(A_i) = m\cdot\frac{1}{m}\cdot E[(Gx)_i\cdot(Gy)_i]$$

$$= E[(Gx)_i\cdot(Gy)_i] = E[(g_1x_1 + \dots + g_dx_d)\cdot(g_1y_1 + \dots + g_dy_d)]$$

$$= E[\sum_{i,j} g_ix_ig_jy_j]$$

$$= \sum_{i,j} x_iy_j \, E(g_ig_j)$$

cuz E_w, E wrt w.

Lemma:

$$\begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix}$$

$G(m\times d)$

$x: (d\times 1)$.

$Gx: (m\times 1)$

$(Gx)_i: w_i x$.

denote $w_i = (g_1, \dots, g_d)$

and $w \overset{iid}{\sim} N(0, I_d)$.

$$\therefore E(g_ig_j) = \begin{cases} 0, & i\neq j. \\ E(N^2(0, I_d)) = 1 & i=j. \end{cases}$$

$X \sim N(0,1)$, $DX = 1$

$E(X^2) = (EX)^2 + DX = 1$.

$$\therefore = \sum_i^d x_iy_i$$

$$= x^T y.$$

Mean squared error (MSE)

$$MSE(X) \overset{def}{=} E[(X-\mu)^2].$$ → 估计量.

if X is unbiased, $\mu = EX$.

$$MSE(X) = E[(X-EX)^2]$$
$$= E[X^2 - 2XEX + (EX)^2]$$
$$= E(X^2) - 2(EX)^2 + (EX)^2.$$
$$= E(X^2) - (EX)^2$$
$$= Var(X). \quad \geq 0$$

What is the bound for MSE?

Markov inequality:

if $X \geq 0$, $P(X \geq a) \leq \frac{E(X)}{a}$ for any $a > 0$.

Proof:

$$E(X) = P(X < a) \cdot \underbrace{E(X \mid X < a)}_{\geq 0} + P(X \geq a) \cdot \underbrace{E(X \mid X \geq a)}_{\geq a}.$$

$$\therefore P(X \geq a) \leq \frac{E(X)}{E(X \mid X \geq a)} \leq \frac{E(X)}{a}.$$

$$Y = (X - EX)^2.$$

$$P[(X-EX)^2 \geq a] \leq \frac{E[(X-EX)^2]}{a} = \frac{Var(X)}{a}$$

$$P[|X-EX| \geq t] \leq \frac{Var(X)}{t^2}$$ → Chebyshev's Inequality.