HW 3

Sherin Karuvallil Saji

q1 + q3 answers are inside a ML_HW3_q1andq3 doc

code for q3 is inside q3.ipynb. (just click on run all)

Answers for q2, q4, q5 are inside this handwritten pdf

Q2)

1. $K(x,z) = K_1(x,z) K_2(x,z)$ kernels over $R^n \times R^n$

To prove that $K(x,z)$ is also a kernel, we need to show that the kernel matrix is positive semi-definite and that $K(x,z)$ is symmetric.

Let's denote kernel matrix corresponding to K_1 as M_1 and kernel matrix corresponding to K_2 as M_2.

kernel matrix M, corresponding to $K(x,z)$ is given by the element-wise multiplication of M_1 & M_2

$$M = M_1 \odot M_2$$

Also for any vector v:

$$v^T M v = v^T (M_1 \odot M_2) v = (v^T M_1 v) * (v^T M_2 v)$$

Since M_1 & M_2 are positive semi-definate,

$v^T M_1 v \geq 0$ and

$v^T M_2 v \geq 0$ for any vector v

$\therefore$ their product

$(v^T M_1 v) * (v^T M_2 v)$

is also non-negative.

$\therefore M$ is positive semi-definate

To show that $K(x,z)$ is symmetric:

We need to show that

$K(x,z) = K(z,x)$

$$K(x,z) = K_1(x,z) K_2(x,z)$$

$$\begin{aligned} K(z,x) &= K_1(z,x) K_2(z,x) \\ &= K_1(x,z) K_2(x,z) \\ &= K(x,z) \end{aligned}$$

because K_1 and K_2 are kernels and definately symmetric meaning

$K_1(x,z) = K_1(z,x)$

$K_2(x,z) = K_2(z,x)$

$\therefore K(x,z) = K(z,x)$

$\therefore K$ is valid kernel

q2)
2. $K(x,z) = aK_1(x,z) + bK_2(x,z)$ where $a,b>0$

for kernel	let the corresponding kernel matrix be
$K(x,z)$	M
$K_1(x,z)$	M_1
$K_2(x,z)$	M_2

To prove that $K(x,z)$ is also a kernel, we need to show that the kernel matrix is positive semi-definite and that $K(x,z)$ is symmetric.

For any vector v:

$$v^T M v = v^T(a*M_1 + b*M_2)v$$
$$= a*(v^T M_1 v) + b*(v^T M_2 v)$$

Since $M1$ & $M2$ are positive semidefinite,
$v^T M_1 v \geq 0$ & $v^T M_2 v \geq 0$ for any vector v

As a & b are positive real numbers,
the combination of
$a*(v^T M_1 v) + b*(v^T M_2 v)$
will also be non-negative.

Proving $K(x,z) = K(z,x)$:

$$K(z,x) = aK_1(z,x) + bK_2(z,x)$$
$$= aK_1(x,z) + bK_2(x,z)$$
$$= K(x,z)$$

K_1 and K_2 are kernels and definately symmetric meaning
$K_1(x,z) = K_1(z,x)$
$K_2(x,z) = K_2(z,x)$

Thus, $K(x,z) = aK_1(x,z) + bK_2(x,z)$ satisfies Mercer's theorem and is a valid kernel.

Q2)
3. $K(x,z) = aK_1(x,z) - bK_2(x,z)$ where $a, b > 0$

To prove that $K(x,z)$ is also a kernel, we need to show that the kernel matrix is positive semi-definite and that $K(x,z)$ is symmetric.

for kernel	let the corresponding kernel matrix be
$K(x,z)$	M
$K_1(x,z)$	M_1
$K_2(x,z)$	M_2

For any vector v:

$v^T M v = v^T(a * M_1 - b * M_2)v$

$= a * (v^T M_1 v) - b * (v^T M_2 v)$

Since $M1$ & $M2$ are positive semidefinite, $v^T M_1 v \geq 0$ & $v^T M_2 v \geq 0$ for any vector v

As a & b are positive real numbers, the combination of $a * (v^T M_1 v) - b * (v^T M_2 v)$ may not always be non-negative.

∴ this case does not satisfy the conditions of Mercer's theorem & is not a valid kernel.

An example of where $K(x,z)$ is not a kernel is:

$$a \underbrace{[0 \;\; 1]}_{1\times 2} \underbrace{\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}}_{2\times 2} \begin{bmatrix} 0 \\ 1 \end{bmatrix} - b [0, 1] \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$= a \underbrace{[3 \;\; 4]}_{1\times 2} \underbrace{\begin{bmatrix} 0 \\ 1 \end{bmatrix}}_{2\times 1} - b [2 \;\; 1] \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$= a \underbrace{(4)}_{1\times 1} - b \underbrace{(1)}_{1\times 1}$$

$$= 4a - b$$

So for the example that $M_1 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$, $M_2 \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix}$

$v^T = [0, 1]$, $v = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

$a = 1$, $b = 8$

These values will give a $K(x,z)$ that is not a valid kernel.

q2)

4. To show that $K(x,z)=f(x)f(z)$ is a valid kernel, we need to demonstrate that for any dataset $X=\{x_1,x_2,\dots,x_n\}$ and any coefficients $c_1,c_2,\dots,c_n$, the Gram matrix G formed using the kernel is positive semi-definite, $c^TGc \geq 0$ for any vector c.

The Gram matrix G is defined as follows:

$$G_{ij} = K(x_i,x_j) = f(x_i)f(x_j)$$

Now, let's compute c^TGc:

$$c^TGc = \sum_{i=1}^{n}\sum_{j=1}^{n} G_{ij}c_ic_j$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} f(x_i)f(x_j)c_ic_j$$

$$= \sum_{i=1}^{n}\left(f(x_i)\,c_i \sum_{j=1}^{n} f(x_j)c_j\right)$$

$$= \sum_{i=1}^{n}\left(f(x_i)\,c_i\right)^2$$

Since the square of any real number is non-negative,

$$c^TGc \geq 0.$$

$\therefore$ $K(x,z)=f(x)f(z)$ is positive semi-definite.

Also, $K(x,z)$ is symmetric because:

$$K(x,z) = f(x)\,f(z)$$

$$K(z,x) = f(z)\,f(x)$$

$$= f(x)\,f(z)$$

$$= K(x,z)$$

$\therefore$ By Mercer's theorem, $K(x,z)$ is a kernel

3. $e^{(t)}(\theta) = \log\left(1 + e^{-y^{(t)}(\theta \cdot x^{(t)}) + \theta_0}\right)$

$$u = -y^{(t)}(\theta \cdot x^{(t)} + \theta_0)$$

$$e^{(t)}(\theta) = \log(1 + e^{u})$$

$$u = -y^{(t)}(\theta \cdot x^{(t)} + \theta_0)$$

$$\frac{du}{d\theta} = -y^{(t)} x^{(t)}$$

$$\frac{\nabla e^{t}(\theta, \theta_0)}{\nabla \theta} = \frac{\frac{du}{d\theta} e^{u}}{1 + e^{u}}$$

$$= \frac{\left(\frac{du}{d\theta} e^{u}\right)(e^{-u})}{(1 + e^{u})(e^{-u})}$$

$$= \frac{\frac{du}{d\theta}}{e^{-u} + 1}$$

$$= \frac{-y^{(t)} x^{(t)}}{1 + e^{y^{(t)}(\theta \cdot x^{(t)})}}$$

$$= \frac{-y^{(t)} x^{(t)}}{1 + \exp(y^{(t)}(\theta \cdot x^{(t)} + \theta_0))}$$

$$u = -y^{(t)}(\theta \cdot x^{(t)} + \theta_0)$$

$$\frac{du}{d\theta_0} = -y^{(t)}$$

$$\frac{\nabla e^{t}(\theta, \theta_0)}{\nabla \theta_0} = \frac{\frac{du}{d\theta_0} e^{u}}{1 + e^{u}}$$

$$= \frac{\frac{du}{d\theta_0}}{e^{-u} + 1}$$

$$= \frac{-y^{(t)}}{1 + \exp(y^{(t)}(\theta \cdot x^{(t)} + \theta_0))}$$

4.

$$y = \underline{\underline{w_0}} + \sum_{i=1}^{n} w_i x_i + w_i x_i^2$$

$$E = \frac{1}{2}\sum_j (y_j - \overset{\text{constant}}{y_j^*})^2$$

$$\frac{\partial E}{\partial w_0} = \sum_j (y_j - y_j^*)\Big(\underbrace{\frac{\partial y_j}{\partial w_0}}_{\to 1}\Big)$$

$$= \sum_j (y_j - y_j^*)$$

$$\frac{\partial E}{\partial w_i} = \sum_j (y_j - y_j^*)\left(\frac{\partial y_j}{\partial w_i}\right)$$

differentiated w_0 w.r.t. w_i

$$\frac{\partial y_j}{\partial w_i} = 0 + \sum_{i=1}^{n} x_i + x_i^2$$

$$\frac{\partial E}{\partial w_i} = \sum_j (y_j - y_j^*)\left(\sum_{i=1}^{n} x_i + x_i^2\right)$$

$\therefore$ The update rule :

$$w0\_new = w0\_old - \text{learning rate} \times \sum_j (y_j - y_j^*)$$

$$wi\_new = wi\_old - \text{learning rate} \times \sum_j (y_j - y_j^*)\left(\sum_{i=1}^{n} x_i + x_i^2\right)$$

5) For 2 features:

Naive Bayes needs to estimate the following probabilities:

1. $P(X1 \mid y=0)$
2. $P(X1 \mid y=1)$
3. $P(X2 \mid y=0)$
4. $P(X2 \mid y=1)$
5. $P(Y=0)$
6. $P(Y=1) = 1 - P(Y=0)$ ↳ so not needed

The goal is to predict the class label (0 or 1) based on the values of X1 and X2.

∴ Total parameters = 5

For 3 features:

Naive Bayes needs to estimate the following probabilities:

1. $P(X1 \mid y=0)$
2. $P(X1 \mid y=1)$
3. $P(X2 \mid y=0)$
4. $P(X2 \mid y=1)$
5. $P(X3 \mid y=0)$
6. $P(X3 \mid y=1)$
7. $P(Y=0)$
8. $P(Y=1) = 1 - P(Y=0)$ ↳ so not needed, just find from $P(Y=0)$

∴ Total parameters = 7

For n features:

For every feature i in 1, ..., n :

Need to find:

$P(X_i \mid y=0)$

$P(X_i \mid y=1)$

on top of this, need to find

$P(Y=0)$

& $P(Y=1) = 1 - P(Y=0)$ ↳ not needed, just find from $P(Y=0)$

∴ Total parameters = 2n + 1