Midterm

Nguyen Trung Hai

Nov. 17, 2019

Problem 1

Show that the necessary and sufficient condition for a symmetric matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ *to be positive definite is that all of its eigenvalues* $\{\lambda_i\}$ *are positive.*

Proof. First, let us prove the ($\Rightarrow$) direction: *If* $\mathbf{M}$ *is positive definite then* $\lambda_i > 0$, $\forall i = 1, 2, \ldots, n$.

Let $\mathbf{u}_i$ be an eigenvector of $\mathbf{M}$ with eigenvalue λ_i

$$\mathbf{M}\mathbf{u}_i = \lambda_i \mathbf{u}_i. \tag{1}$$

Because $\mathbf{M}$ is positive definite, we have

$$\mathbf{u}_i^T \mathbf{M}\mathbf{u}_i > 0, \tag{2}$$
$$\mathbf{u}_i^T \lambda_i \mathbf{u}_i > 0, \tag{3}$$
$$\lambda_i \mathbf{u}_i^T \mathbf{u}_i > 0, \tag{4}$$
$$\lambda_i \left\|\mathbf{u}_i\right\|_2^2 > 0. \tag{5}$$

Since $\mathbf{u}_i \neq \mathbf{0}$ by the definition of eigenvector, we have $\left\|\mathbf{u}_i\right\|_2^2 > 0$. So we can divide both sides of (5) by $\left\|\mathbf{u}_i\right\|_2^2$ to arrive at

$$\lambda_i > 0. \tag{6}$$

This proves the ($\Rightarrow$) direction.

Now let us prove the ($\Leftarrow$) direction: *If all eigenvalues of* $\mathbf{M}$ *are positive then* $\mathbf{M}$ *is a positive definite matrix.*

Let $\mathbf{\Lambda}$ be a diagonal matrix whose diagonal elements consist of eigenvalues of $\mathbf{M}$, $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Let $\mathbf{U}$ be a matrix whose i-th column is the i-th eigenvector of $\mathbf{M}$, $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$. By definition, $\mathbf{M}\mathbf{u}_i = \lambda_i \mathbf{u}_i$, $\forall i = 1, 2, \ldots, n$, which can be written in matrix form as follows,

$$\mathbf{M}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}. \tag{7}$$

According to the *spectral theorem*, if $\mathbf{M}$ is symmetric, then there exists an orthonormal basis set for $\mathbb{R}^n$ consisting of eigenvectors of $\mathbf{M}$. In other words, we can choose $\mathbf{U}$ to be an orthonormal matrix, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. Multiplying from the right of Eq. (7) by $\mathbf{U}^T$, we arrive at the following matrix decomposition

$$\mathbf{M} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T, \tag{8}$$

which is called *eigendecomposition* or *spectral decomposition*. By re-arranging Eq. (8) we have

$$\boldsymbol{\Lambda} = \mathbf{U}^T\mathbf{M}\mathbf{U}. \tag{9}$$

Now let us consider the quadratic form $\mathbf{x}^T\mathbf{M}\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x} \neq \mathbf{0}$. Since $\{\mathbf{u}_i\}$ is an orthonormal basis set of $\mathbb{R}^n$, we can write any vector $\mathbf{x} \in \mathbb{R}^n$ as a linear combination of $\{\mathbf{u}_i\}$

$$\begin{aligned}\mathbf{x} &= \sum_{i=1}^{n} a_i\mathbf{u}_i \\ &= \mathbf{U}\mathbf{a},\end{aligned} \tag{10}$$

where $\mathbf{a} = (a_1, a_2, \ldots, a_n)^T$ is also a vector in $\mathbb{R}^n$ and $\mathbf{a} \neq \mathbf{0}$ because $\mathbf{x} \neq \mathbf{0}$.

The quadratic form becomes

$$\begin{aligned}\mathbf{x}^T\mathbf{M}\mathbf{x} &= (\mathbf{U}\mathbf{a})^T\mathbf{M}\mathbf{U}\mathbf{a} && (11)\\ &= \mathbf{a}^T\mathbf{U}^T\mathbf{M}\mathbf{U}\mathbf{a} && (12)\\ &= \mathbf{a}^T\boldsymbol{\Lambda}\mathbf{a} && (13)\\ &= \sum_{i=1}^{n} \lambda_i a_i^2. && (14)\end{aligned}$$

Since $\lambda_i > 0\ \forall i = 1, 2, \ldots, n$ and $\mathbf{a} \neq \mathbf{0}$, we have $\sum_{i=1}^{n} \lambda_i a_i^2 > 0$ which implies that $\mathbf{x}^T\mathbf{M}\mathbf{x} > 0$. This holds for all $\mathbf{x} \neq \mathbf{0}$. So $\mathbf{M}$ is definite positive, which proves the ($\Leftarrow$) direction. $\square$

Problem 2

To prove the non-negativity of Kullback-Leibler (KL) divergence, we will make use of the *Jensen's inequality* which states that for any convex function f, we have

$$f\left(\sum_{i=1}^{n} a_i x_i\right) \leq \sum_{i=1}^{n} a_i f(x_i), \tag{15}$$

where $a_i \geq 0$ and $\sum_{i=1}^{n} a_i = 1$. The equality holds iff $x_1 = x_2 = \cdots = x_n$ or if f is a linear function. If we interpret a_i as the probability mass function of a discrete random variable X, then the Jensen's inequality can be written as

$$f\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[f(X)\right], \tag{16}$$

where $\mathbb{E}[\cdot]$ denotes the expectation. The equality holds iff X is a constant random variable or if f is a linear function.

(a)

Now let us use the Jensen's inequality to prove that KL divergence is non-negative.

Proof. Let $\mathcal{A} = \{x : p(x) > 0\}$ be the support of $p(x)$.

$$
\begin{aligned}
D(p||q) &= \sum_{x\in\mathcal{A}} p(x) \log \frac{p(x)}{q(x)} && (17)\\
&= \sum_{x\in\mathcal{A}} p(x) \left[-\log \frac{q(x)}{p(x)} \right] && (18)\\
&\geq -\log \sum_{x\in\mathcal{A}} p(x) \frac{q(x)}{p(x)} && (19)\\
&= -\log \sum_{x\in\mathcal{A}} q(x) && (20)\\
&\geq -\log \sum_{x\in\mathcal{X}} q(x) && (21)\\
&= -\log 1 = 0. && (22)
\end{aligned}
$$

Inequality (19) follows from Jensen's inequality where $-\log$ is a strictly convex function. Inequality (21) follows from the fact that $\sum_{x\in\mathcal{A}} q(x) \leq \sum_{x\in\mathcal{X}} q(x)$, where $\mathcal{X}$ is the set of all posible values of x.

Since $-\log$ is a strictly convex function, the equality in (19) holds iff $\frac{q(x)}{p(x)} = c$, where c is a constant. Multiplying both sides by $p(x)$ and taking sum over $\mathcal{A}$ we get $\sum_{x\in\mathcal{A}} q(x) = c\sum_{x\in\mathcal{A}} p(x) = c$. On the other hand, the equality in (21) holds iff $\sum_{x\in\mathcal{A}} q(x) = \sum_{x\in\mathcal{X}} q(x) = 1$ which implies that $c = 1$. Hence $D(p||q) = 0$ iff $p(x) = q(x)$, $\forall x \in \mathcal{X}$. $\square$

(b)

$$
\begin{aligned}
D(p||q) &= p(0) \log \frac{p(0)}{q(0)} + p(1) \log \frac{p(1)}{q(1)} && (23)\\
&= (1-r) \log \frac{1-r}{1-s} + r \log \frac{r}{s} && (24)\\
&= \log \left(\frac{1-r}{1-s} \right)^{(1-r)} + \log \left(\frac{r}{s} \right)^{r} && (25)\\
&= \log \left[\left(\frac{1-r}{1-s} \right)^{(1-r)} \cdot \left(\frac{r}{s} \right)^{r} \right]. && (26)
\end{aligned}
$$

$$D(q||p) = q(0)\log\frac{q(0)}{p(0)} + q(1)\log\frac{q(1)}{p(1)} \tag{27}$$

$$= (1-s)\log\frac{1-s}{1-r} + s\log\frac{s}{r} \tag{28}$$

$$= \log\left(\frac{1-s}{1-r}\right)^{(1-s)} + \log\left(\frac{s}{r}\right)^{s} \tag{29}$$

$$= \log\left[\left(\frac{1-s}{1-r}\right)^{(1-s)} \cdot \left(\frac{s}{r}\right)^{s}\right]. \tag{30}$$

(c)

Setting $s = r$ in Eq. (26) gives

$$D(p||q) = \log\left[\left(\frac{1-r}{1-r}\right)^{1-r} \cdot \left(\frac{r}{r}\right)^{r}\right] = \log 1 = 0. \tag{31}$$

Setting $s = r$ in Eq. (30) gives

$$D(q||p) = \log\left[\left(\frac{1-r}{1-r}\right)^{1-r} \cdot \left(\frac{r}{r}\right)^{r}\right] = \log 1 = 0. \tag{32}$$

(d)

Substituting $r = \frac{1}{2}$ and $s = \frac{1}{4}$ in Eq. (26) gives

$$D(p||q) = \log_2\left[\left(\frac{1/2}{3/4}\right)^{1/2} \cdot \left(\frac{1/2}{1/4}\right)^{1/2}\right] \tag{33}$$

$$= \frac{1}{2}\log_2\frac{4}{3} \tag{34}$$

$$\approx 0.2075 \text{ bits.} \tag{35}$$

Substituting $r = \frac{1}{2}$ and $s = \frac{1}{4}$ in Eq. (30) gives

$$D(p||q) = \log_2\left[\left(\frac{3/4}{1/2}\right)^{3/4} \cdot \left(\frac{1/4}{1/2}\right)^{1/4}\right] \tag{36}$$

$$= \log_2\left[\left(\frac{3}{2}\right)^{3/4} \cdot \left(\frac{1}{2}\right)^{1/4}\right] \tag{37}$$

$$\approx 0.18872 \text{ bits.} \tag{38}$$

(e)

Proof.

$$\begin{aligned}
\mathrm{PSI}(\hat{p}, \hat{q}) &= D(\hat{p}||\hat{q}) + D(\hat{q}||\hat{p}) && (39)\\
&= \sum_{i=1}^{B} \hat{p}_i \log \frac{\hat{p}_i}{\hat{q}_i} + \sum_{i=1}^{B} \hat{q}_i \log \frac{\hat{q}_i}{\hat{p}_i} && (40)\\
&= \sum_{i=1}^{B} \left[\hat{p}_i(\log \hat{p}_i - \log \hat{q}_i) + \hat{q}_i(\log \hat{q}_i - \log \hat{p}_i)\right] && (41)\\
&= \sum_{i=1}^{B} \left[\hat{p}_i(\log \hat{p}_i - \log \hat{q}_i) - \hat{q}_i(\log \hat{p}_i - \log \hat{q}_i)\right] && (42)\\
&= \sum_{i=1}^{B} (\hat{p}_i - \hat{q}_i)(\log \hat{p}_i - \log \hat{q}_i) && (43)
\end{aligned}$$

□

Problem 3

The sigmoid function is given by

$$\begin{aligned}
\sigma(a) &= \frac{1}{1 + e^{-1}} && (44)\\
&= \frac{e^a}{e^a + 1}. && (45)
\end{aligned}$$

The tanh function is given by

$$\begin{aligned}
\tanh(a) &= \frac{e^a - e^{-a}}{e^a + e^{-a}} && (46)\\
&= \frac{e^{2a} - 1}{e^{2a} + 1} && (47)\\
&= \frac{e^{2a}}{e^{2a} + 1} - \frac{1}{e^{2a} + 1} + 1 - 1 && (48)\\
&= \frac{e^{2a}}{e^{2a} + 1} + \frac{e^{2a}}{e^{2a} + 1} - 1 && (49)\\
&= 2\sigma(2a) - 1 && (50)
\end{aligned}$$

$$\Rightarrow \sigma(2a) = \frac{1}{2}\left[\tanh(a) + 1\right] \tag{51}$$

Now let us prove that a linear combination of sigmoid functions is equivalent to a linear combination of tanh functions.

Proof.

$$
\begin{aligned}
y(x, \mathbf{w}) &= w_0 + \sum_{j=1}^{M} w_j \sigma \left(\frac{x - \mu_j}{s} \right) && (52) \\
&= w_0 + \sum_{j=1}^{M} w_j \sigma \left(2 \frac{x - \mu_j}{2s} \right) && (53) \\
&= w_0 + \sum_{j=1}^{M} \frac{w_j}{2} \left[\tanh \left(\frac{x - \mu_j}{2s} \right) + 1 \right] && (54) \\
&= w_0 + \sum_{j=1}^{M} \frac{w_j}{2} + \sum_{j=1}^{M} \frac{w_j}{2} \tanh \left(\frac{x - \mu_j}{2s} \right) && (55) \\
&\equiv u_0 + \sum_{j=1}^{M} u_j \tanh \left(\frac{x - \mu_j}{2s} \right), && (56)
\end{aligned}
$$

where

$$
u_0 = w_0 + \sum_{j=1}^{M} \frac{w_j}{2}, \tag{57}
$$

and

$$
u_j = \frac{w_j}{2}, \ \forall j = 1, 2, \ldots, M. \tag{58}
$$

□

Problem 4

The cost function for weighted linear regression is given by

$$
J(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{m} w^{(i)} \left(\boldsymbol{\theta}^T \mathbf{x}^{(i)} - y^{(i)} \right)^2, \tag{59}
$$

where $\boldsymbol{\theta} \in \mathbb{R}^{(n+1)}$ is the parameter vector. $\boldsymbol{\theta}^T \mathbf{x}^{(i)}$ is the i-th element of the vector $\mathbf{X}\boldsymbol{\theta}$, $\boldsymbol{\theta}^T \mathbf{x}^{(i)} = [\mathbf{X}\boldsymbol{\theta}]^{(i)}$, where $\mathbf{X} \in \mathbb{R}^{m \times (n+1)}$ is the input data matrix. $y^{(i)}$ is the i-th element of the target vector $\mathbf{y} \in \mathbb{R}^m$.

(a)

To write the cost function above in quadratic form, let's recall that the quadratic form $\mathbf{x}^T \mathbf{M} \mathbf{x}$ can be expanded as a double sum as,

$$
\mathbf{x}^T \mathbf{M} \mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i m_{ij} x_j, \tag{60}
$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$ and $\mathbf{M} = [m_{ij}] \in \mathbb{R}^{n\times n}$ is a symmetric matrix.

Let us introduce the Kronecker delta

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases} \tag{61}$$

Using the Kronecker delta δ_{ij}, we can write $J(\boldsymbol{\theta})$ in Eq. (59) as a double sum

$$\begin{aligned} J(\boldsymbol{\theta}) &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} w^{(i)}\delta_{ij}\left([\mathbf{X}\boldsymbol{\theta}]^{(i)} - y^{(i)}\right)^2 && (62) \\ &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \left([\mathbf{X}\boldsymbol{\theta}]^{(i)} - y^{(i)}\right) w^{(i)}\delta_{ij}\left([\mathbf{X}\boldsymbol{\theta}]^{(j)} - y^{(j)}\right). && (63) \end{aligned}$$

Let $\mathbf{W} \in \mathbb{R}^{m\times m}$ be a matrix whose (i,j) element is $w_{ij} = w^{(i)}\delta_{ij}$. $\mathbf{W}$ is a diagonal matrix whose diagonal elements are the weights, $\mathbf{W} = \mathrm{diag}(w^{(1)}, w^{(2)}, \ldots, w^{(m)})$. With these matrix notations, we can write $J(\boldsymbol{\theta})$ in Eq. (63) as

$$J(\boldsymbol{\theta}) = \frac{1}{2}(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T\mathbf{W}(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}). \tag{64}$$

(b)

To take gradient of the cost function $J(\boldsymbol{\theta})$ in Eq. (64), let us expand it as follows

$$\begin{aligned} J(\boldsymbol{\theta}) &= \frac{1}{2}(\mathbf{X}\boldsymbol{\theta})^T\mathbf{W}\mathbf{X}\boldsymbol{\theta} - \frac{1}{2}(\mathbf{X}\boldsymbol{\theta})^T\mathbf{W}\mathbf{y} - \frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{X}\boldsymbol{\theta} + \frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{y} && (65) \\ &= \frac{1}{2}\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\theta} - \frac{1}{2}\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{W}\mathbf{y} - \frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{X}\boldsymbol{\theta} + \frac{1}{2}\mathbf{y}^T\mathbf{W}\mathbf{y}. && (66) \end{aligned}$$

Note that $\mathbf{X}^T\mathbf{W}\mathbf{X}$ is a symmetric matrix because $(\mathbf{X}^T\mathbf{W}\mathbf{X})^T = \mathbf{X}^T\mathbf{W}^T\mathbf{X} = \mathbf{X}^T\mathbf{W}\mathbf{X}$. So the gradient is given by

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\theta} - \frac{1}{2}\mathbf{X}^T\mathbf{W}\mathbf{y} - \frac{1}{2}(\mathbf{y}^T\mathbf{W}\mathbf{X})^T + \mathbf{0} && (67) \\ &= \mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\theta} - \mathbf{X}^T\mathbf{W}\mathbf{y}. && (68) \end{aligned}$$

Assume that $\mathbf{X}$ has full column rank. This implies that $\mathbf{X}^T\mathbf{W}\mathbf{X}$ is invertible. Solving equation $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbf{0}$ for $\boldsymbol{\theta}$ gives

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}. \tag{69}$$

Problem 5

The PDF or PMF of a random variable Y belonging to the exponential family can be written in the form

$$p(y) = b(y)\exp\left[\eta T(y) - a(\eta)\right], \tag{70}$$

for some parameter $\eta \in \mathbb{R}$ and some functions $a(\cdot)$, $b(\cdot)$ and $T(\cdot)$.

(a)

Let us prove that the Gaussian distribution belongs to the exponential family.

Proof. The PDF of a normally distributed random variable Y with mean μ and variance $\sigma = 1$ is given by

$$
\begin{aligned}
p(y;\mu) &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(y-\mu)^2\right] &(71)\\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2 + \mu y - \frac{1}{2}\mu^2\right) &(72)\\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) \exp\left(\mu y - \frac{1}{2}\mu^2\right). &(73)
\end{aligned}
$$

Comparing with Eq. (70) allows us to identify

$$
\begin{aligned}
\eta &= \mu. &(74)\\
a(\eta) &= \frac{1}{2}\eta^2. &(75)\\
T(y) &= y. &(76)\\
b(y) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right). &(77)\\
& &(78)
\end{aligned}
$$

□

(b)

Prove that the Bernoulli PMF belongs to the exponential family.

Proof. The Bernoulli PMF can be written as

$$
\begin{aligned}
p(y;\phi) &= \phi^y (1-\phi)^{1-y} &(79)\\
&= \exp\left[\ln \phi^y + \ln(1-\phi)^{1-y}\right] &(80)\\
&= \exp\left[y \ln \phi + \ln(1-\phi) - y\ln(1-\phi)\right] &(81)\\
&= \exp\left[\left(\ln \frac{\phi}{1-\phi}\right) y + \ln(1-\phi)\right]. &(82)
\end{aligned}
$$

Compare with Eq. (70) we have

$$
\eta = \ln \frac{\phi}{1-\phi}. \tag{83}
$$

$$\Rightarrow \frac{\phi}{1-\phi} = e^{\eta} \tag{84}$$
$$\Rightarrow \frac{1-\phi}{\phi} = e^{-\eta} \tag{85}$$
$$\Rightarrow \frac{1}{\phi} - 1 = e^{-\eta} \tag{86}$$
$$\Rightarrow \frac{1}{\phi} = 1 + e^{-\eta} \tag{87}$$
$$\Rightarrow \phi = \frac{1}{1+e^{-\eta}}. \tag{88}$$

$$\begin{aligned}
a(\eta) &= -\ln(1-\phi) && (89)\\
&= -\ln\left(1 - \frac{1}{1+e^{-\eta}}\right) && (90)\\
&= -\ln\left(\frac{e^{-\eta}}{1+e^{-\eta}}\right) && (91)\\
&= \ln\left(\frac{1+e^{-\eta}}{e^{-\eta}}\right) && (92)\\
&= \ln(e^{\eta}+1). && (93)\\
T(y) &= y. && (94)\\
b(y) &= 1. && (95)
\end{aligned}$$

□

(c)

Prove that the Poisson PMF belongs to the exponential family.

Proof. The Poisson PMF is given by

$$\begin{aligned}
p(y;\lambda) &= \frac{e^{-\lambda}\lambda^{y}}{y!} && (96)\\
&= \frac{1}{y!}\exp(-\lambda)\exp(\ln\lambda^{y}) && (97)\\
&= \frac{1}{y!}\exp\left[(\ln\lambda)y - \lambda\right]. && (98)
\end{aligned}$$

Compare with Eq. (70) we have

$$\begin{aligned}
\eta &= \ln\lambda. && (99)\\
a(\eta) &= \lambda = e^{\eta}. && (100)\\
T(y) &= y. && (101)\\
b(y) &= \frac{1}{y!}. && (102)
\end{aligned}$$

$\square$

Problem 6

(a)

Show that the radial basis function (RBF) is a valid kernel. The RBF is given by

$$K(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|^2}{2}\right). \tag{103}$$

Definition. $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a kernel if there exists a feature map $\phi : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^T \phi(\mathbf{z}). \tag{104}$$

To show that RBF is a valid kernel, we will make use of the following properties:

(i) $K(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T \mathbf{z}$ is a valid kernel corresponding to the feature map $\phi : \mathbb{R}^n \to \mathbb{R}^n$ which is simply the identity function, $\phi(\mathbf{x}) = \mathbf{x}$.

(ii) If $K_a(\mathbf{x}, \mathbf{z}) = \phi_a(\mathbf{x})^T \phi_a(\mathbf{z})$ is a valid kernel then $K(\mathbf{x}, \mathbf{z}) = \alpha K_a(\mathbf{x}, \mathbf{z})$, where α is a positive constant, is also a valid kernel with the new feature map being $\phi(\mathbf{x}) = \sqrt{\alpha}\phi_a(x)$.

(iii) If $K_a(\mathbf{x}, \mathbf{z}) = \phi_a(\mathbf{x})^T \phi_a(\mathbf{z})$ and $K_b(\mathbf{x}, \mathbf{z}) = \phi_b(\mathbf{x})^T \phi_b(\mathbf{z})$, where $\phi_a : \mathbb{R}^n \to \mathbb{R}^{m_a}$ and $\phi_b : \mathbb{R}^n \to \mathbb{R}^{m_b}$, are valid kernels, then $K(\mathbf{x}, \mathbf{z}) = K_a(\mathbf{x}, \mathbf{z}) K_b(\mathbf{x}, \mathbf{z})$ is also a valid kernel. The feature map for K can be obtained by expanding the product of the two kernels

$$K(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{m_a} \phi_a^{(i)}(\mathbf{x}) \phi_a^{(i)}(\mathbf{z}) \sum_{j=1}^{m_b} \phi_b^{(j)}(\mathbf{x}) \phi_b^{(j)}(\mathbf{z}) \tag{105}$$

$$= \sum_{i=1}^{m_a} \sum_{j=1}^{m_b} \left[\phi_a^{(i)}(\mathbf{x}) \phi_b^{(j)}(\mathbf{x})\right] \left[\phi_a^{(i)}(\mathbf{z}) \phi_b^{(j)}(\mathbf{z})\right] \tag{106}$$

$$\equiv \sum_{i=1}^{m_a} \sum_{j=1}^{m_b} \phi^{(i,j)}(\mathbf{x}) \phi^{(i,j)}(\mathbf{z}). \tag{107}$$

So the new feature map corresponding to K is $\phi^{(i,j)}(\mathbf{x}) = \phi_a^{(i)}(\mathbf{x}) \phi_b^{(j)}(\mathbf{x})$ for $i = 1, 2, \ldots, m_a$ and $j = 1, 2, \ldots, m_b$. This property implies that taking a valid kernel to some power, $K_a(\mathbf{x}, \mathbf{z})^p$ where p is some positive integer, also gives a valid kernel.

(iv) If $K_a(\mathbf{x}, \mathbf{z}) = \phi_a(\mathbf{x})^T \phi_a(\mathbf{z})$ is a valid kernel and $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued function, then $K(\mathbf{x}, \mathbf{z}) = f(\mathbf{x}) f(\mathbf{z}) K_a(\mathbf{x}, \mathbf{z})$ is also a valid kernel with the new feature map $\phi(\mathbf{x}) = f(\mathbf{x}) \phi_a(\mathbf{x})$.

(v) If $K_a(\mathbf{x}, \mathbf{z}) = \phi_a(\mathbf{x})^T \phi_a(\mathbf{z})$ and $K_b(\mathbf{x}, \mathbf{z}) = \phi_b(\mathbf{x})^T \phi_b(\mathbf{z})$ are valid kernels, then $K(\mathbf{x}, \mathbf{z}) = K_a(\mathbf{x}, \mathbf{z}) + K_b(\mathbf{x}, \mathbf{z})$ is also a valid kernel with the new feature map $\phi(\mathbf{x}) = [\phi_a(\mathbf{x}), \phi_b(\mathbf{x})]^T$.

(vi) If $K_a(\mathbf{x}, \mathbf{z})$ is a valid kernel then $K(\mathbf{x}, \mathbf{z}) = \exp[K_a(\mathbf{x}, \mathbf{z})]$ is also a valid kernel. To prove this let's use Taylor series expansion of the exponential function

$$\exp[K_a(\mathbf{x}, \mathbf{z})] = \sum_{i=1}^{\infty} \frac{[K_a(\mathbf{x}, \mathbf{z})]^i}{i!}. \tag{108}$$

By applying property (iii) for power, property (ii) for multiplication by a positive constant and property (v) for sum, we can easily prove that $\exp[K_a(\mathbf{x}, \mathbf{z})]$ is, indeed, a valid kernel.

Proof. Now let's apply these properties to prove that RBF is a valid kernel. Expanding the norm gives

$$\begin{aligned} \|\mathbf{x} - \mathbf{z}\|^2 &= (\mathbf{x} - \mathbf{z})^T (\mathbf{x} - \mathbf{z}) && (109) \\ &= \mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{z} - \mathbf{z}^T\mathbf{x} + \mathbf{z}^T\mathbf{z} && (110) \\ &= \mathbf{x}^T\mathbf{x} + \mathbf{z}^T\mathbf{z} - 2\mathbf{x}^T\mathbf{z}. && (111) \end{aligned}$$

$$\begin{aligned} K(\mathbf{x}, \mathbf{z}) &= \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|^2}{2}\right) && (112) \\ &= \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2}\right) \exp\left(-\frac{\mathbf{z}^T\mathbf{z}}{2}\right) \exp\left(\mathbf{x}^T\mathbf{z}\right) && (113) \end{aligned}$$

$\mathbf{x}^T\mathbf{z}$ is a valid kernel by property (i). Then $\exp\left(\mathbf{x}^T\mathbf{z}\right)$ is a valid kernel by property (vi). Finally, $\exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2}\right) \exp\left(-\frac{\mathbf{z}^T\mathbf{z}}{2}\right) \exp\left(\mathbf{x}^T\mathbf{z}\right)$ is a valid kernel by property (iv), where $f(\mathbf{x}) = \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2}\right)$. Therefore, RBF is a valid kernel. $\square$

(b)

To show that a function is not a valid kernel we need to prove that the *Gram matrix* is, in general, NOT positive semi-definite. It is, therefore, sufficient to indicate that Gram matrix is not positive semi-definite for at least one set of input vectors.

Proof. Consider a set consisting of two vectors, $\mathcal{X} = \{\mathbf{x}, \mathbf{z} \in \mathbb{R}^2\}$ chosen as follows

$$\mathbf{x} = \left(-\frac{r}{a}, 0\right)^T, \quad \mathbf{z} = (a, a)^T, \quad \forall a \in \mathbb{R}, a \neq 0. \tag{114}$$

The elements of Gram matrix $K \in \mathbb{R}^{2\times2}$ for the set $\mathcal{X}$ are

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}) &= \tanh(\mathbf{x}^T\mathbf{x} + r) = \tanh\left(\frac{r^2}{a^2} + r\right), && (115) \\ K(\mathbf{x}, \mathbf{z}) &= K(\mathbf{z}, \mathbf{x}) = \tanh(\mathbf{x}^T\mathbf{z} + r) = \tanh(-r + r) = 0, && (116) \\ K(\mathbf{z}, \mathbf{z}) &= \tanh(\mathbf{z}^T\mathbf{z} + r) = \tanh(2a^2 + r). && (117) \end{aligned}$$

$$K = \begin{pmatrix} \tanh\left(\frac{r^2}{a^2} + r\right) & 0 \\ \\ 0 & \tanh(2a^2 + r) \end{pmatrix}. \tag{118}$$

Since K is already a diagonal matrix, we can easily read out its eigenvalues as

$$\lambda_1 = \tanh\left(\frac{r^2}{a^2} + r\right), \tag{119}$$

$$\lambda_2 = \tanh(2a^2 + r), \tag{120}$$

and eigenvectors as

$$\mathbf{v}_1 = (1, 0)^T, \tag{121}$$

$$\mathbf{v}_2 = (0, 1)^T. \tag{122}$$

If either of both eigenvalues of K are negative, then K is not positive semi-definite. Note that $\tanh(x) < 0$ if $x < 0$. Since $r < 0$, we can always choose a to make either eigenvalue λ_1 or λ_2 negative. This implies that when $r < 0$, we can always choose a set of vectors that makes Gram matrix not positive semi-definite. So the sigmoid kernel is not a valid kernel for all $r < 0$. □

Problem 7

The XOR dataset is given by

$$\mathbf{x}^{(1)} = (0, 0), \quad y^{(1)} = \mathrm{XOR}(0, 0) = 0, \tag{123}$$

$$\mathbf{x}^{(2)} = (0, 1), \quad y^{(2)} = \mathrm{XOR}(0, 1) = 1, \tag{124}$$

$$\mathbf{x}^{(3)} = (1, 0), \quad y^{(3)} = \mathrm{XOR}(1, 0) = 1, \tag{125}$$

$$\mathbf{x}^{(4)} = (1, 1), \quad y^{(4)} = \mathrm{XOR}(1, 1) = 0. \tag{126}$$

(a)

Now we will prove that the XOR dataset is not linearly separable.

Proof. Suppose that the dataset is linearly separable. This means that the parameters $\mathbf{w} = (w_1, w_2)^T$ and b must satisfy the following four inequalities

$$\begin{cases} w_1 0 + w_2 0 + b < 0 \\ w_1 0 + w_2 1 + b > 0 \\ w_1 1 + w_2 0 + b > 0 \\ w_1 1 + w_2 1 + b < 0 \end{cases} \tag{127}$$

$$\Rightarrow \begin{cases} b < 0 \\ w_2 + b > 0 \\ w_1 + b > 0 \\ w_1 + w_2 + b < 0 \end{cases} \tag{128}$$

Adding the first and the fourth inequalities in (128) gives

$$w_1 + w_2 + 2b < 0. \tag{129}$$

Adding the second and the third inequalities in (128) gives

$$w_1 + w_2 + 2b > 0. \tag{130}$$

The parameters, w_1, w_2 and b must simultaneously satisfy both (129) and (130), which is impossible. Therefore the dataset is not linearly separable. $\square$

Please see the jupyter notebook for answers to Problems **(7b)**, **(7c)**, **(8)**, and **(9)**.

Problem 10

(a)

Feed-forward equations

First let us make some notational conventions.

Let $\mathbf{a}^{(l)} \in \mathbb{R}^{d^{(l)}}$ be the output vector at layer l and also the input vector into layer $l+1$. $d^{(l)}$ is the number of units in layer l. $d^{(1)}$ is the number of input features. $\mathbf{a}^{(1)} \equiv \mathbf{x}$.

Let $\mathbf{W}^{(l)} \in \mathbb{R}^{d^{(l-1)} \times d^{(l)}}$ be the weight matrix connecting layer $(l-1)$ and layer l.

The net input $\mathbf{z}^{(l)} \in \mathbb{R}^{(l)}$ at layer l is given by

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l)T}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}, \tag{131}$$

where $\mathbf{b}^{(l)}$ is the bias vector at layer l. The activated output of layer l is given by

$$\mathbf{a}^{(l)} = \sigma(\mathbf{z}^{(l)}), \tag{132}$$

where $\sigma(\cdot)$ is the sigmoid function.