Sample answers for homework :

(1) Let $\mathbf{w} = [w_1, w_2, w_3]^T$. Represent the following functions in the form of $\mathbf{w}^T\mathbf{A}\mathbf{w}$.

(i) $g(\mathbf{w}) = 5w_1^2 + w_2^2 + 5w_3^2 + 4w_1w_2 - 8w_1w_3 - 4w_2w_3$

(ii) $g(\mathbf{w}) = 3w_1^2 + w_2^2 + 5w_3^2 + 4w_1w_2 - 6w_1w_3 - 4w_2w_3$

Answer: (i) $g(\mathbf{w}) = \mathbf{w}^T \begin{pmatrix} 5 & 2 & -4 \\ 2 & 1 & -2 \\ -4 & -2 & 5 \end{pmatrix} \mathbf{w}$

(ii) $g(\mathbf{w}) = \mathbf{w}^T \begin{pmatrix} 3 & 2 & -3 \\ 2 & 1 & -2 \\ -3 & -2 & 5 \end{pmatrix} \mathbf{w}$

(2) Find the Hessian of $g(\mathbf{w})$ in (1).

Answer: (i) $\begin{pmatrix} 10 & 4 & -8 \\ 4 & 2 & -4 \\ -8 & -4 & 10 \end{pmatrix}$ (ii) $\begin{pmatrix} 6 & 4 & -6 \\ 4 & 2 & -4 \\ -6 & -4 & 10 \end{pmatrix}$

(3) Let $\mathbf{w} = [w_1, w_2]^T$. $J(\mathbf{w}) = 8w_1^2 + 7w_2^2 + 2w_1w_2$. Use Lagrange method to find the minimum of $J(\mathbf{w})$, subject to $h(\mathbf{w}) = 2w_1 + w_2 - 2 = 0$.

Answer:

$$L(\mathbf{w}, \lambda) = 8w_1^2 + 7w_2^2 + 2w_1w_2 + \lambda(2w_1 + w_2 - 2)$$

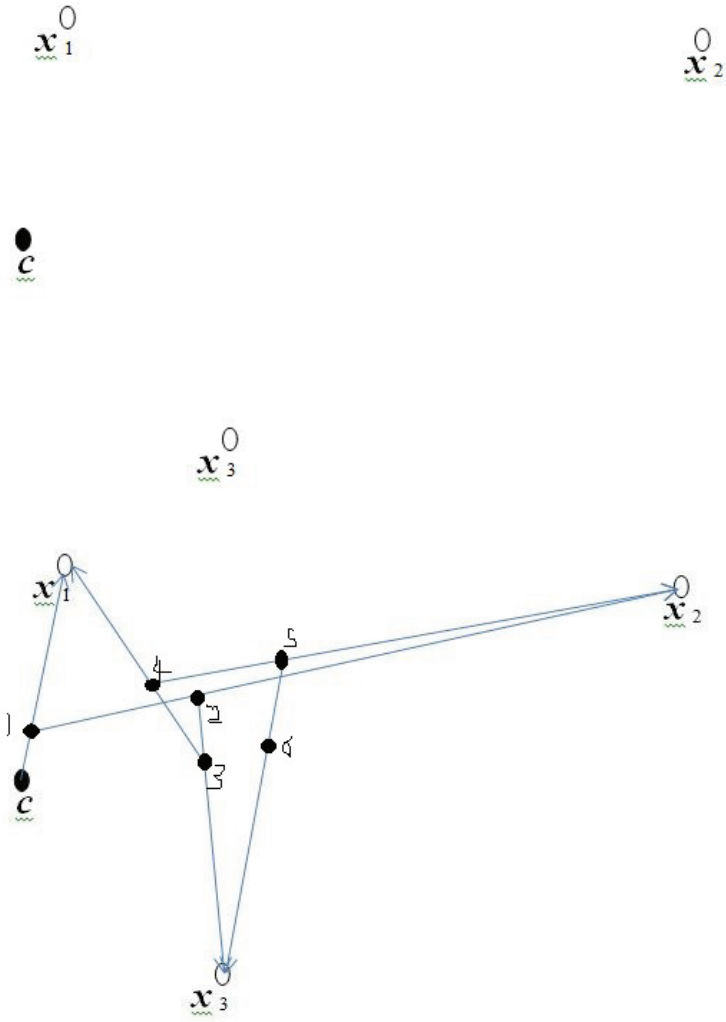
$$\begin{aligned} \frac{\partial L(\mathbf{w})}{\partial w_1} &= 16w_1 + 2w_2 + 2\lambda = 0 \\ \frac{\partial L(\mathbf{w})}{\partial w_2} &= 14w_2 + 2w_1 + \lambda = 0 \\ \frac{\partial L(\mathbf{w})}{\partial \lambda} &= h(\mathbf{w}) = 2w_1 + w_2 - 2 = 0 \end{aligned}$$

$$\begin{pmatrix} w_1 \\ w_2 \\ \lambda \end{pmatrix} = \begin{pmatrix} 16 & 2 & 2 \\ 2 & 14 & 1 \\ 2 & 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 13/16 \\ 3/8 \\ -55/8 \end{pmatrix}$$

(NB. You can solve the equations any other way.)

The minimum is $J(\mathbf{w}) = 8(13/16)^2 + 7(3/8)^2 + 2(13/16)(3/8) = 6.875$

(4) In Figure below, three data points are shown as circles and a single centre is shown as black dot and initialized. Consider applying the on-line k-means clustering algorithm with learning rate is 0.25. Plot the trajectory of the centre for two training epochs, assuming that the data samples are presented in the order of x_1, x_2 and x_3 and so on.
Answer:



(5) Suppose that we are given a data set consisting of points $x_{i,j}$ from two classes respectively, where $j = 1, 2$, denotes class label, and i denotes the data index. (a) Determine the class label for a new data point $x = 1.5$ using a probabilistic neural network, with the Gaussian function as window function and $\sigma = 1$. (b) How do you find the classification decision boundary of the

probabilistic neural network used in (a)? (b) The data set is as follows:

Class 1: $\{1, 2, 0.5\}$

Class 2: $\{2, 3, 3.5\}$

Answer: (a) The probabilistic neural network is given as

$$y_1(x) = \frac{1}{3}\big(\exp(-\frac{(x-1)^2}{2}) + \exp(\frac{(x-2)^2}{2}) + \exp(-\frac{(x-0.5)^2}{2}))\big)$$
$$y_2(x) = \frac{1}{3}\big(\exp(-\frac{(x-2)^2}{2}) + \exp(\frac{(x-3)^2}{2}) + \exp(-\frac{(x-3.5)^2}{2}))\big)$$

Thus $y_1(1.5) = 0.7905$, and $y_2(1.5) = 0.4475$, $y_1(1.5) > y_2(1.5)$, $x = 1.5$ is classified as class one.

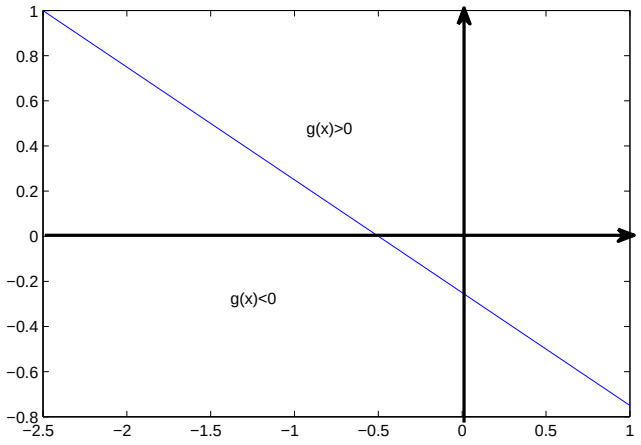
(b) The classification boundary is found by setting $y_1(x) = y_2(x)$, and can be solved numerically.

(6) Given $\boldsymbol{w} = [1,\ 2]^{\mathrm{T}}$, $w_0 = 0.5$, find $g(x) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0$ for (a) $\boldsymbol{x} = [-1,\ 2]^{\mathrm{T}}$; (b) $\boldsymbol{x} = [-1,\ 1]^{\mathrm{T}}$ and (c) $\boldsymbol{x} = [1,\ -1]^{\mathrm{T}}$ respectively. Plot $g(\boldsymbol{x}) = 0$. Indicate two half planes with $g(\boldsymbol{x}) > 0$, and $g(\boldsymbol{x}) < 0$.

Answer: Only solution for (a) is shown.

$$g(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0 = [1,\ 2][\ \begin{matrix}-1\\2\end{matrix}\ ] + 0.5 = 3.5$$

To find $g(\boldsymbol{x}) = 0$, linking two arbitrary points when $g(\boldsymbol{x}) = 0$, e.g. $[0, -0.5]$ or $[-0.25, 0]$.



(7) A single layer perceptron is given

$$y(\mathbf{x}, \mathbf{w}) = h\Big(\sum_{i=0}^{2} w_i x_i\Big)$$

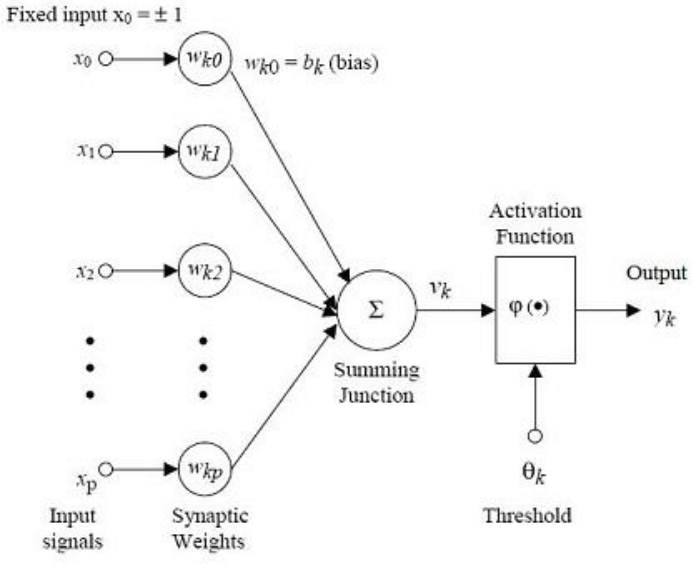
where $\mathbf{x} \in [x_1, x_2]^{\mathrm{T}}$ and $x_0 = 1$, $\mathbf{w} = [w_0, w_1, w_2]^{\mathrm{T}}$. $h(v) = \frac{1}{1+\exp(-50v)}$. Calculate the network output for each of the data input $\mathbf{x} = [1, 2]^{\mathrm{T}}$, $\mathbf{x} = [2, 1]^{\mathrm{T}}$, $\mathbf{x} = [2, 2]^{\mathrm{T}}$, in the cases that the network weights are $\mathbf{w} = [1, 0.5, -0.5]^{\mathrm{T}}$, $\mathbf{w} = [-1, 0.5, -0.5]^{\mathrm{T}}$ and $\mathbf{w} = [0, 0.5, -0.5]^{\mathrm{T}}$ respectively. Sketch the network diagram.

Answer: Only solution when $\mathbf{x} = [1, 2]^{\mathrm{T}}$ and $\mathbf{w} = [1, 0.5, -0.5]^{\mathrm{T}}$ is shown.

$$\sum_{i=0}^{2} w_i x_i = [1, 0.5, -0.5] \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} = 0.5$$

$$y(\sum_{i=0}^{2} w_i x_i) = \frac{1}{1 + \exp(-50 \times 0.5)} = 1$$

The network diagram can be modified based on



(8) The online gradient descent algorithm is used to train a single layer perceptron given by

$$y(\mathbf{x}, \mathbf{w}) = h\Big(\sum_{i=0}^{2} w_i x_i\Big)$$

where $\mathbf{x} \in [x_1, x_2]^{\mathrm{T}}$ and $x_0 = 1$, $\mathbf{w} = [w_0, w_1, w_2]^{\mathrm{T}}$. $h(v) = \tanh(v) = \frac{\exp(v)-\exp(-v)}{\exp(v)+\exp(-v)}$. (Note: $h'(v) = (1 - h(v)^2)$). Assume that the current weight vector is $\mathbf{w} = [1, 0.3, 0.4]^{\mathrm{T}}$, Calculate the new weight updated from a new training datum $[\mathbf{x}, t] = [2, 1, 0.2]^{\mathrm{T}}$, using the learning rate $\eta = 0.02$.

Answer: The weight update is given by

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla E_n(\mathbf{w}^{(t)})$$

where $E_n = \frac{(y_n - t_n)^2}{2}$,

$$\nabla E_n = (y_n - t_n) y_n' [1 \;\; \mathbf{x}_n^{\mathrm{T}}]^{\mathrm{T}} = (y_n - t_n)(1 - y_n^2)[1 \;\; \mathbf{x}_n^{\mathrm{T}}]^{\mathrm{T}}$$

Or

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta (y_n - t_n)(1 - y_n^2)[1 \;\; \mathbf{x}_n^{\mathrm{T}}]^{\mathrm{T}}$$

The network output is

$$y_n = tanh(1 + 0.3 \times 2 + 0.4 \times 1) = 0.9640$$

So we have

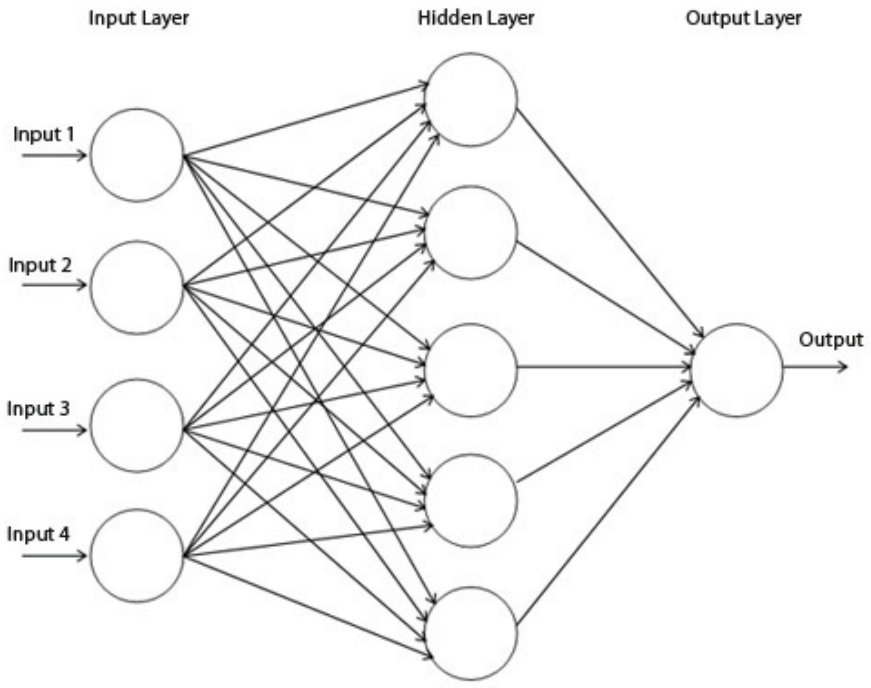
$$\mathbf{w}^{(t+1)} = \begin{pmatrix} 1 \\ 0.3 \\ 0.4 \end{pmatrix} - 0.02(0.9640 - 0.2)(1 - 0.9640^2) \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.9989 \\ 0.2978 \\ 0.3989 \end{pmatrix}$$

(9) The mathematical form for a two layer MLP is

$$y(\mathbf{x}, \mathbf{w}) = \sigma\Big(\sum_{j=0}^{M} w_j^{(2)} h\Big(\sum_{i=0}^{D} w_{ji}^{(1)} x_i\Big)\Big)$$

where $\mathbf{x} \in \Re^D$. $x_0 = 1$. The superscript $^{(1)}$ and $^{(2)}$ indicate the corresponding weights are in *first* or *second* layer. $h(.)$ and $\sigma(.)$ are chosen activation functions. Sketch the diagram of MLP for $M = 2$ and $D = 3$, specifying the weights on the paths of the diagram.

Answer: The network diagram can be extended from



(10) The mathematical form for a two layer MLP is

$$y(\mathbf{x}, \mathbf{w}) = \sigma\Big(\sum_{j=0}^{2} w_j^{(2)} h\Big(\sum_{i=0}^{2} w_{ji}^{(1)} x_i \Big) \Big)$$

where $x_0 = 1$, $\mathbf{x} \in \Re^D$. The superscript $^{(1)}$ and $^{(2)}$ indicate the corresponding weights are in *first* or *second* layer. For

$$\sigma(v) = h(v) = \begin{cases} 1 & \text{if } v \geq 0 \\ 0 & \text{if } v < 0 \end{cases}$$

Calculate the network outputs for each of the data input $\mathbf{x} = [-1, 3]^{\mathrm{T}}$, $\mathbf{x} = [-3, 1]^{\mathrm{T}}$, $\mathbf{x} = [-2, 2]^{\mathrm{T}}$, in the cases that the network weights are $w_{ji}^{(1)} = 1$, $\forall i, j$, $\mathbf{w}^{(2)} = [0, 0.5, -0.5]^{\mathrm{T}}$, $\mathbf{w}^{(2)} = [0, 1.5, -0.5]^{\mathrm{T}}$ and $\mathbf{w}^{(2)} = [0, 0.5, -1.5]^{\mathrm{T}}$ respectively.
Answer: Only solution when $\mathbf{x} = [-1, 3]^{\mathrm{T}}$ and $w_{ji}^{(1)} = 1$, $\forall i, j$, $\mathbf{w}^{(2)} = [0, 0.5, -0.5]^{\mathrm{T}}$ is shown.

$$h\Big(\sum_{i=0}^{2} w_{ji}^{(1)} x_i \Big) = h\Big([1, 1, 1]^T [1, -1, 3] \Big) = h(3) = 1$$

$$y(\mathbf{x}, \mathbf{w}) = \sigma\Big(\sum_{j=0}^{2} w_j^{(2)} h\Big(\sum_{i=0}^{2} w_{ji}^{(1)} x_i \Big) \Big)$$

$$= \sigma\Big([0, 0.5, -0.5]^T [1, 1, 1] \Big) \Big) = \sigma(0) = 1$$

(11) Explain the training step of the error backpropogation algorithm for the two layer perceptron of

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{2} w_j^{(2)} h\Big(\sum_{i=0}^{2} w_{ji}^{(1)} x_i \Big)\Big)$$

$h(v) = \tanh(v) = \frac{\exp(v)-\exp(-v)}{\exp(v)+\exp(-v)}$. (Note: $h'(v) = (1 - h(v)^2)$) where $x_0 = 1$, $\mathbf{x} \in \Re^D$. The superscript $^{(1)}$ and $^{(2)}$ indicate the corresponding weights are in *first* or *second* layer. If the previous network weights are $w_{ji}^{(1)} = 1$, $\forall i, j$, $\mathbf{w}^{(2)} = [0, 0.5, -0.5]^{\mathrm{T}}$, what are the weights after applying one pass of the error backpropogation algorithm with the training sample $[\mathbf{x}, t] = [2, 1, 3]^{\mathrm{T}}$?

Answer:

(a) Apply an input vector $\mathbf{x}_n$ to the network and forward propagate through the network to the find the activations of all hidden and output units.

(b) Evaluate the δ_k for all the outputs units using

$$\delta_k = y_k - t_k$$

(c) Backpropagate the δ to obtain δ_j for each hidden unit in the network, using

$$\delta_j = \frac{\partial E_n}{\partial a_j} = \sum_k \frac{\partial E_n}{\partial a_k} \frac{\partial a_k}{\partial a_j}$$

(d) Evaluate the derivatives using

$$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$$

For example, we perform a forward propagation using

$$a_j = \sum_{i=0}^{D} w_{ji}^{(1)} x_i = 1 \times 1 + 1 \times 2 + 1 \times 1 = 4, \forall j$$

$$z_j = \tanh(a_j) = 0.9993, \forall j$$

$$y_k = \sum_{j=0}^{M} w_{kj}^{(2)} z_j = 0 * 0.9993 + 0.5 * 0.9993 - 0.5 * 0.9993 = 0$$

Next we compute the δ for each output unit using

$$\delta = y - t = 0 - 2 = -2$$

Then we backpropagate to obtain $\delta's$ for the hidden units using

$$\delta_1 = (1 - z_1^2)\sum_{k=1}^{K} w_{k1}^{(2)}\delta_k = (1 - 0.9993^2)(0)(-2) = 0$$

$$\delta_2 = (1 - z_2^2)\sum_{k=1}^{K} w_{k2}^{(2)}\delta_k = (1 - 0.9993^2)(0.5)(-2) = -0.0014$$

$$\delta_3 = (1 - z_3^2)\sum_{k=1}^{K} w_{k3}^{(2)}\delta_k = (1 - 0.9993^2)(-0.5)(-2) = 0.0014$$

Finally the derivatives with respect to the first layer and second layer weights are given by

$$\frac{\partial E_n}{\partial w_{ji}^{(1)}} = \delta_j x_i$$

For above multiply each of δ_j (0, -0.0014, 0.0014) with each of x_i (2, 1) to get six different values.

$$\frac{\partial E_n}{\partial w_j^{(2)}} = \delta z_i = (-2) * 0.9993 = -1.9986, \forall j$$

(12) A radial basis function network has the form of

$$y(\mathbf{x}) = \sum_{i=1}^{3} w_i \exp\left(-\frac{(\|\mathbf{x} - \mathbf{c}_i\|)^2}{2}\right)$$

The centers are $\mathbf{c}_1 = [-1, 3]^{\mathrm{T}}$, $\mathbf{c}_2 = [-3, 1]^{\mathrm{T}}$, $\mathbf{c}_3 = [-2, 2]^{\mathrm{T}}$. Calculate the network outputs for input $\mathbf{x}$ equals to each center $\mathbf{c}_1$, $\mathbf{c}_2$ and $\mathbf{c}_3$, respectively in the cases that the network weights as $\mathbf{w} = [1/3, 1/3, 1/3]^{\mathrm{T}}$,$\mathbf{w} = [2, 1, -1]^{\mathrm{T}}$,$\mathbf{w} = [1, 0, 0]^{\mathrm{T}}$.
Answer: Only solution for $\mathbf{x}$ equals to $\mathbf{c}_1$ and $\mathbf{w} = [1/3, 1/3, 1/3]^{\mathrm{T}}$ is shown.

$$y(\mathbf{x}) = 1/3\exp\left(-\frac{(\|\mathbf{c}_1 - \mathbf{c}_1\|)^2}{2}\right) + 1/3\exp\left(-\frac{(\|\mathbf{c}_1 - \mathbf{c}_2\|)^2}{2}\right)$$

$$+1/3\exp\left(-\frac{(\|\mathbf{c}_1 - \mathbf{c}_3\|)^2}{2}\right) = 0.4621$$

(13) Find the solution of $\min f(x,y) = x^2+y-1$, subject to $x+y \geq 5$.
Answer: Write objective

$$x^2 + y - 1 - \mu(x + y - 5)$$

KKT condition means:

$$\begin{array}{l} 1.\ \mu \geq 0 \\ 2.\ \begin{cases} 2x - \mu = 0 \\ 1 - \mu = 0 \end{cases} \\ 3.\ \mu(x + y - 5) = 0 \end{array}$$

So we have $\mu = 1$, $x + y - 5 = 0$, $x = \mu/2 = 0.5$, $y = 4.5$.

(14) Supposing $\mathbf{x} = [x_1, x_2]^{\mathrm{T}}$, $\mathbf{y} = [y_1, y_2]^{\mathrm{T}}$, determine the feature map for $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^{\mathrm{T}}\mathbf{y} + 1)^2$.

$$\begin{aligned} \textbf{Answer}: (\mathbf{x}^{\mathrm{T}}\mathbf{y} + 1)^2 &= (x_1y_1 + x_2y_2 + 1)^2 \\ &= x_1^2y_1^2 + x_2^2y_2^2 + 2x_1y_1x_2y_2 + 2x_1y_1 + 2x_2y_2 + 1 \\ &= [x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1]^{\mathrm{T}}[y_1^2, y_2^2, \sqrt{2}y_1y_2, \sqrt{2}y_1, \sqrt{2}y_2, 1] \\ &= \phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{y}) \end{aligned}$$

So the feature map $\phi(\mathbf{x}) = [x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1]$

(15) Determine if $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^{\mathrm{T}}\mathbf{y}+1)^2+\exp(-\|\mathbf{x}-\mathbf{y}\|^2)$ is a kernel.

Answer: Both $(\mathbf{x}^{\mathrm{T}}\mathbf{y} + 1)^2$ and $\exp(-\|\mathbf{x} - \mathbf{y}\|^2)$ are kernels, $\alpha\mathbf{x}^{\mathrm{T}}\mathbf{y}+1)^2+\beta\exp(-\|\mathbf{x}-\mathbf{y}\|^2)$ are kernels for $\alpha > 0, \beta > 0$. So $(\mathbf{x}^{\mathrm{T}}\mathbf{y}+1)^2+\exp(-\|\mathbf{x}-\mathbf{y}\|^2)$ is a kernel (it's a case of $\alpha = \beta = 1$)

(16) Consider two class data set

x_1	x_2	t
-1	-1	1
-1	1	-1
1	-1	-1
1	1	1

What is the kernel matrix if $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^{\mathrm{T}}\mathbf{y}+1)^2+\exp(-\|\mathbf{x}-\mathbf{y}\|^2)$?

Answer: $K = \begin{pmatrix} 10 & 1.0183 & 1.0183 & 1.0003 \\ 1.0183 & 10 & 1.0003 & 1.0183 \\ 1.0183 & 1.0003 & 10 & 1.0183 \\ 1.0003 & 1.0183 & 1.0183 & 10 \end{pmatrix}$

(17) For the data set in (16), what is the objective function of support vector machine is dual form in the cases that the kernel is $k(\mathbf{x},\mathbf{y}) = (\mathbf{x}^{\mathrm{T}}\mathbf{y}+1)^2$ or $k(\mathbf{x},\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2)$ respectively?
Answer: For $k(\mathbf{x},\mathbf{y}) = (\mathbf{x}^{\mathrm{T}}\mathbf{y}+1)^2$, We have

$$K = \begin{pmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{pmatrix}$$

and the objective function in dual form is

$$\begin{aligned}\tilde{L}(\mathbf{a}) &= \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n,\mathbf{x}_m) \\ &= a_1 + a_2 + a_3 + a_4 - \frac{1}{2}\Big(9a_1^2 - 2a_1a_2 - 2a_1a_3 \\ &\quad + 2a_1a_4 + 9a_2^2 + 2a_2a_3 - 2a_2a_4 + 9a_3^2 \\ &\quad - 2a_3a_4 + 9a_4^2\Big)\end{aligned}$$

For $k(\mathbf{x},\mathbf{y}) = \exp(-\|\mathbf{x}-\mathbf{y}\|^2)$, we have

$$K = \begin{pmatrix} 1 & 0.0183 & 0.0183 & 0.0003 \\ 0.0183 & 1 & 0.0003 & 0.0183 \\ 0.0183 & 0.0003 & 1 & 0.0183 \\ 0.0003 & 0.0183 & 0.0183 & 1 \end{pmatrix}$$

and the objective function in dual form is

$$\begin{aligned}\tilde{L}(\mathbf{a}) &= \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n,\mathbf{x}_m) \\ &= a_1 + a_2 + a_3 + a_4 - \frac{1}{2}\Big(a_1^2 - 0.0366a_1a_2 - 0.0366a_1a_3 \\ &\quad + 0.0006a_1a_4 + a_2^2 + 0.0006a_2a_3 - 0.0366a_2a_4 + a_3^2 \\ &\quad - 0.0366a_3a_4 + a_4^2\Big)\end{aligned}$$

(18) What variables are solved in the support vector regression in dual form, and how are they used in the model prediction for new data samples?
Answer: It solves Lagrange parameters $\mathbf{a}$.