1. Why we need the basis function $\phi(x)$ for linear regression? And what is the benefit for applying basis function over linear regression? (5%)

$y(x, w) = w_0 + w_1 x_1 + \cdots + w_D x_D$

線性迴歸因為其函數是線性，這對model帶來極大的局限性.

因此我們將 x 做非線性函數轉換，i.e.，

$y(x, w) = \sum_{j=0}^{M-1} w_j \phi_j(x) = w^T \phi(x)$，$\phi_0(x) = 1$.

其中 $\phi(x)$ 就是將 x 做非線函數轉換的 basis function，目的是讓 input variable x_i 透過 $\phi(x)$ 使 model 限制更少，x 從只能是線性函數變成可以是非線性函數.

benifit：

有了 basis function，$y(x,w)$ 的 x 可以有更多選擇，因為 x 不必需要線性，只要系數項 w 維持線性即可，根據不同的 basis func.，我們的 model 可以更多元，

e.g.：多項式、gaussion、sigmoid …

藉由 $\phi(x)$ 的非線性轉換為 model 帶來更廣的應用.

以上是 basis function 在 linear regression 的作用. #

2. Prove that the predictive distribution just mentioned is the same with the form

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}(t|m(x), s^2(x))$$

where

$$m(x) = \beta \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \boldsymbol{\phi}(x).$$

Here, the matrix $\mathbf{S}^{-1}$ is given by $\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^{\mathrm{T}}$ (15%)

(hint: $p(\mathbf{w}|\mathbf{x}, \mathbf{t}) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w})p(\mathbf{w})$ and you may use the formulas shown in page 93.)

$$p(t|x, X, t) = \int_{-\infty}^{\infty} p(t|x, w, \beta) P(w|X, t)\, dw$$

$$p(w|X, t) \propto p(t|X, w)\, p(w|\alpha)$$

$$p(t|X, w) = \mathcal{N}(t|w^T \Phi(x), \beta^{-1} I)$$

$$= \mathcal{N}(t|w^T A + b, L^{-1})$$

知 $\underline{A = \Phi(x)^T}$, $\underline{b = 0}$, $\underline{L = \beta I}$

$$P(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1} I) = \mathcal{N}(w|\mu, \Lambda^{-1})$$

知 $\underline{\mu = 0}$, $\underline{\Lambda = \alpha I}$

$$p(w|X, t) = \underline{\mathcal{N}(w|\Sigma\{A^T L(w - b) + \Lambda\mu\}, \Sigma)}$$

其中 $\Sigma = (\alpha I + A^T L A)^{-1}$

已知的 $A = \Phi(x)^T$, $b = 0$, $L = \beta I$, $\mu = 0$, $\Lambda = \alpha I$

代入 $p(w|X, t)$ 中，

$\Rightarrow \mathcal{N}(w \mid S(\Phi^T(x)\beta t, S)$,

其中 $\underline{S = (\alpha I + \Phi(x)\beta\Phi(x)^T)^{-1}}$

至目前為止我們推出Posterior的分配，

接下來將Posterior視為prior為新資料更新

我們直接用課本P.93的結果.

$$p(t \mid w, x) = \mathcal{N}(t \mid w^T\Phi(x), \beta^{-1})$$

$$= \mathcal{N}(t \mid w^T A + b, L^{-1})$$

$\Rightarrow \underline{A = \Phi(x)}$, $\underline{b = 0}$, $\underline{L = \beta I}$

$$p(w \mid x, t) = \mathcal{N}(w \mid S(\beta\Phi(x)t), S)$$

$$= p(w \mid \mu, \Lambda^{-1})$$

$\Rightarrow \underline{\mu = S(\beta\Phi(x)t)}$, $\underline{\Lambda^{-1} = S}$

取代 $A = \Phi(x)^T$, $b = 0$, $L = \beta I$, $\mu = 0$, $\Lambda = \alpha I$

故 $p(t \mid x, \mathbf{x}, \mathbf{t})$

$$= \mathcal{N}(t \mid A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$$

$$= \mathcal{N}(t \mid \beta\Phi(x)^T S\Phi(x)t, \beta^{-1} + \Phi(x)^T S\Phi(x))$$

#

3. Could we use linear regression function for classification? Why or why not? Explain it! (10%)

NO. 線性迴歸用來分類主要有以下2個缺點:

1. 存在 hetroscedasticity (異質變異數) 問題.

e.g.: $y_i = \begin{cases} 1 \\ 0 \end{cases}$,

$$E(Y_i) = \beta_0 + \beta_1 X_i = P_i = P(Y_i = 1)$$

機率	Y_i	$u_i = Y_i - E(Y_i) = Y_i - P_i$
$P_i = E(Y_i)$	1	$1 - P_i$
$1-P_i = 1-E(Y_i)$	0	$-P_i$
$E(u_i)$	$(1-P_i)P_i + (-P_i)(1-P_i) = 0$	
$E(u_i^2)$	$(1-P_i)^2 P_i + (-P_i)^2(1-P_i) = P_i(1-P_i)$	

$$\therefore Var(u_i) = E[u_i^2] - [E(u_i)]^2$$

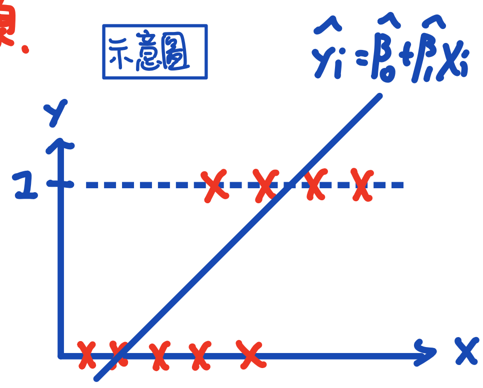
$$= P_i(1-P_i)$$

$$= E(Y_i)[1 - E(Y_i)]$$

$$= (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i)$$

$\because$ 誤差項 depends on X_i

$\therefore$ 存在異質變異數問題.



2. 機率有可能超過1或出現負數