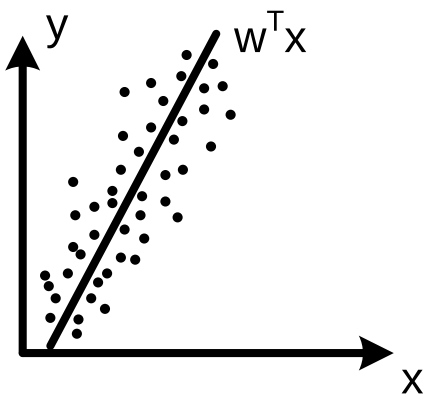
前情提要(LSE)



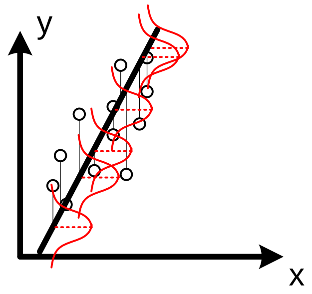
w 是我們想要得到最 fit 直線的參數，在 LSE 時我們也提過，給定許多資料點，我們希望得到 $y = w_0 + w_1 x + w_2 x^2 + \ldots + w_k x^k$ 中每點代入 x 配上適當的 w_i，能得到最接近的 y，這裡我們使用的 design matrix 為$[1\ \ x\ \ x^2\ \ \ldots x^k]$，k 是自己定義的，而在 lession1 時我們使用二維空間來做說明，在這裡 x 不一定只是一維向量，可能是一個多維(假設為 D 維)的向量$\begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_D \end{bmatrix}$，如果用之前的 design matrix，x 為$\begin{bmatrix} 1 & x_1 & x_1^{\,2} & & x_1^{\,k} \\ 1 & x_2 & x_2^{\,2} & \ldots & x_2^{\,k} \\ 1 & \ldots & \ldots & & \ldots \\ 1 & x_D & x_D^{\,2} & & x_D^{\,k} \end{bmatrix}$

之前 LSE 時我們是找$\| A\vec{x} - \vec{b} \|^2$的最小值，這裡的$\vec{x}$就是對應到剛剛提到的 **w**，A 是用我們現有的 input 做成的 design matrix，假設為 X。

data: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

則 $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} = \begin{bmatrix} a + bx_1 + cx_1^{\,2} + \ldots \\ a + bx_2 + cx_2^{\,2} + \ldots \\ \ldots \\ a + bx_3 + cx_3^{\,2} + \ldots \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^{\,2} & \\ 1 & x_2 & x_2^{\,2} & \ldots \\ 1 & \ldots & \ldots & \\ 1 & x_D & x_D^{\,2} & \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ \ldots \end{bmatrix} = X\mathbf{w}$

每個 x 值上都會有一個 Gaussian distribution 對應，就像是上一個 lesson 最後一張圖一樣



但是這裡每個 x 值中 Gaussian distribution 的 mean 和 variance 都是不一樣的，mean 就是我們想要 fit 直線的那一點，variance 則是我們自己給定，每個 x 值可有不同的 variance。則

$$y_i \sim N(X\mathbf{w}, \sigma_i{}^2)$$

直至目前為止，我們都還是在做 LSE，只是我們”距離”的概念變了而已，本來離 mean 越遠，距離越大，但是在這裡，離 mean 越近，機率(距離)越大，故本來的 LSE 是找最小值，若我們用 Gaussian distribution 來表示距離，我們就是找最大值(代表機率最大)

note:這裡和上課用的記號有些不一樣，但我覺得我寫的也 OK，後面的推導才會比較順暢

接下來，我們就要說明這兩條紅線的對應關係

frequentist		bayesian	
LSE	↔	MLE	可能發生overfitting
Regularize LSE	↔	MAP (max posterior)	給定prior
		fully Gaussian ↓ Predictive distribution	

MLE

一樣的，我們想找出”給定一組參數 **w**，能得到目前看到的 data 的機率是多少”，想找出這種機率最大時，參數 **w** 為何

$$P(D\,|\,w) \text{ or } P(D_y\,|\,D_x, \mathbf{w})$$

$$= \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{-1}{2\sigma_i^2}(y_i - X_i\mathbf{w})^2} \quad \alpha(\text{正比於}) \;\; e^{\sum_i \frac{-1}{2\sigma_i^2}(y_i - X_i\mathbf{w})^2}$$

若要找最大值，由於需要使用微分，連乘的方式不好使用微分，故我們一樣使用 log 的技巧

$$\log P(D\,|\,\mathbf{w}) = \sum_i \log(\frac{1}{\sqrt{2\pi}\sigma_i}) + \sum_i \frac{-1}{2\sigma_i^2}(y_i - X_i\mathbf{w})^2$$

我們之後會需要對 **w** 做微分，等是右方的第一項會為 0，第二項的係數並不會影響我們找出此多項式最大值，故我們在求的是 $\sum_i (y_i - X_i\mathbf{w})^2$ 的最大值，$\sum_i (y_i - X_i\mathbf{w})^2 = \sum_i (X_i\mathbf{w} - y_i)^2$ ，和 LSE 的形式

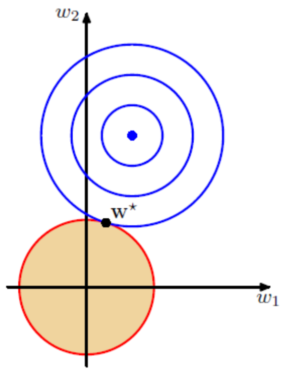
$\| A\vec{x} - \vec{b} \|^2$ 是一樣的

MAP

接下來我們要看是否 regularize LSE(rLSE)其實就是 frequentist 版的 maximum posterior，還記得 rLSE 吧！因為 LSE 很有可能會發生 overfitting 的現象，而發生 overfitting 時通常參數 **w** 都會很大，故我們在 LSE 後面加上一項懲罰項避免參數太大，形式為

$$\min \tilde{E}(\mathbf{w}) = \sum_{n=1}^{N} \{(y(x_n, \mathbf{w}) - t_n\}^2 + \lambda \| \mathbf{w} \|^2$$

用圖來說就是



詳細的就請看 lesson1 了！

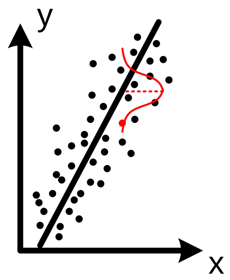
而 MAP 和 MLE 不一樣的就是會給予一個 prior，之後找出哪個參數 **w** 擁有最大的 posterior，我們在這裡就是要說明 rLSE 後面那懲罰項 $\lambda \| \mathbf{w} \|^2$ 其實就是 LSE 的 prior。

data $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

$y_i \sim N(Xw, \sigma_i^2)$ ，還記得做 LSE 或是 rLSE 時，每個 x 值我們計算距離時，使用的距離公式都是一樣的，故為了和 rLSE 的形式一樣，我們假設每個 x 值對應的 y 值得 Gaussian distribution 的所計算的距離公式中的各個維度 variance 都相同且獨立。故

$y_i \sim N(Xw, \sigma^2) = N(Xw, a^{-1})$

這裡的 a^{-1} 只是為了未來計算好看用，且注意，這裡的 y 是 univariate Gaussian distribution，因為這裡的維度只有一維



note: 其實每個 x 值的 variance 可以不同，指是為了對齊 rLSE 才有這種假設

posterior $P(\mathbf{w} \mid D) = \dfrac{P(D \mid \mathbf{w}) P(\mathbf{w})}{P(D)}$，故我們也要給予 prior $P(\mathbf{w})$

而為了對齊 rLSE 的 prior 項，可以看一下上面那張圖，所增加的懲罰項的 contour 是由原點向外擴展的完美正圓，故我們假設我們的 prior 為

$P(\mathbf{w}) \sim N(0, b^{-1}I)$

而 $b^{-1} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ \dots & \dots & \dots \;\dots & \dots \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}$，b 為 covariance matrix 的逆矩陣，又稱為 precision matrix

note: 這裡的 prior 是 multivariate Gaussian distribution，因為不同的資料就會有自己的 Gaussian distribution，joint 起來就會是 multivariate Gaussian distribution

b 為 design matrix 對應的 w 的 covariance matrix

e.g

舉例來說，若 design matrix 為 $\begin{bmatrix} 1 & x & x^2 \end{bmatrix}$，方程式為 $y = w_0 + w_1 x + w_2 x^2 = \mathbf{Xw}$，而

$$b^{-1} = \begin{bmatrix} {\sigma_{w_1}}^2 & 0 & 0 \\ 0 & {\sigma_{w_2}}^2 & 0 \\ 0 & 0 & {\sigma_{w_3}}^2 \end{bmatrix}$$

在這裡我們假設 design matrix 各項彼此為獨立，故只有主對角項有值，若其餘項有值，就是 Gaussian process，之後會提到

有了 prior 後，我們就可以將 posterior 的形式寫出來，而分母的 margin 因為和參數 **w** 無關，故 margin 並不會影響我們找最大的 posterior(微分=0，若微分前多項式只有一項，前面的係數並不重要)，故我們就只專注於找 $P(D|\mathbf{w})P(\mathbf{w})$ 的最大值

前面推導過，$P(D|\mathbf{w})\ \alpha\ e^{\frac{-1}{2\sigma_i^2}\sum_i (y_i - X_i\mathbf{w})^2}$

故(因為 prior 的 precision matrix 中的對角項都一樣，我們直接將 b 視為純量，為 variance 的倒數)

$$P(\mathbf{w}|D)\ \alpha\ P(D|\mathbf{w})P(\mathbf{w})\ \alpha\ e^{\frac{-1}{2\sigma_a^2}\sum_i (y_i - X_i\mathbf{w})^2} e^{\frac{-1}{2}\mathbf{w}^T bI\mathbf{w}} = e^{\frac{-a}{2}\sum_i (y_i - X_i\mathbf{w})^2 + \frac{-1}{2}\mathbf{w}^T b\mathbf{w}}$$
$$= e^{\frac{-a}{2}\sum_i (y_i - X_i\mathbf{w})^2 + \frac{-b}{2}\mathbf{w}^T \mathbf{w}}$$

我們要找上式的最大值的 **w**，一樣的我們取 log

$$\log(P(D|\mathbf{w})P(\mathbf{w}))\ \alpha\ \frac{-a}{2}\sum_i (y_i - X_i\mathbf{w})^2 + \frac{-b}{2}\mathbf{w}^T\mathbf{w}$$

寫成 matrix form

$$\frac{-a}{2}\sum_i (y_i - X_i\mathbf{w})^2 + \frac{-b}{2}\mathbf{w}^T\mathbf{w} = \frac{-a}{2}\| X\mathbf{w} - \mathbf{y} \|^2 + \frac{-b}{2}\mathbf{w}^T\mathbf{w}$$
$$= \frac{-a}{2}(\| X\mathbf{w} - \mathbf{y} \|^2 + \frac{b}{a}\mathbf{w}^T\mathbf{w})$$

我們可以觀察一下，一樣的前面的係數不重要，和 rLSE 形式比較，$\frac{b}{a}$ 就是 rLSE 的 λ

接下來，我們要推導，若是 prior 是 multivariate，likelihood 是 univariate，得出來的 posterior 是 multivariate

我們還是使用和前面一樣的方法，我們不重視 exponential 前面的係數項，因為那是最後 normalize 會將係數都處理掉，我們重視的是指數項是不是 quadratic form $(\mathbf{w}-\boldsymbol{\mu})^T\Lambda(\mathbf{w}-\boldsymbol{\mu})$

指數項：
$$\begin{aligned} & a\| X\mathbf{w} - \mathbf{y} \|^2 + b\mathbf{w}^T\mathbf{w} = a(X\mathbf{w}-\mathbf{y})^T(X\mathbf{w}-\mathbf{y}) + b\mathbf{w}^T\mathbf{w} \\ & = a(\mathbf{w}^T X^T X\mathbf{w} - 2\mathbf{w}^T X^T\mathbf{y} + \mathbf{y}^T\mathbf{y}) + b\mathbf{w}^T\mathbf{w} \\ & = \mathbf{w}^T(aX^T X + bI)\mathbf{w} - 2a\mathbf{w}^T X^T\mathbf{y} + a\mathbf{y}^T\mathbf{y} \end{aligned}$$

對應的 quadratic form：
$$\begin{aligned} & (\mathbf{w}-\boldsymbol{\mu})^T\Lambda(\mathbf{w}-\boldsymbol{\mu}) \\ & = \mathbf{w}^T\Lambda\mathbf{w} - 2\mathbf{w}^T\Lambda\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\mu} \end{aligned}$$

後面的常數項其實不重要，因為常數項可以放到 exponential 前面做為係數，最後的 normalize (marginalize)會幫我們修正，所以常數項不重要

用數學式子敘述剛剛的概念一次：

$$e^{\mathbf{w}^T(aX^TX+bI)\mathbf{w}-2a\mathbf{w}^TX^T\mathbf{y}+a\mathbf{y}^T\mathbf{y}}=e^{\mathbf{w}^T(aX^TX+bI)\mathbf{w}-2a\mathbf{w}^TX^T\mathbf{y}+a\mathbf{y}^T\mathbf{y}}=\mathrm{e}^{(\mathbf{w}^T\Lambda\mathbf{w}-2\mathbf{w}^T\Lambda\mu+\mu^T\mu)-\mu^T\mu+a\mathbf{y}^T\mathbf{y}}$$

$$=e^{-\mu^T\mu+a\mathbf{y}^T\mathbf{y}}e^{\mathbf{w}^T\Lambda\mathbf{w}-2\mathbf{w}^T\Lambda\mu+\mu^T\mu}=Ae^{(\mathbf{w}-\boldsymbol{\mu})^T\Lambda(\mathbf{w}-\boldsymbol{\mu})}$$

注意，這裡我們的變數只有 **w**，其餘都視為常數

則只要令 $\begin{array}{l}\Lambda=aX^TX+bI\\ \boldsymbol{\mu}=a\Lambda^{-1}X^T\mathbf{y}\end{array}$ ，指數項就可以變成 quadratic form，我們就可以得到 posterior 為

multivariate Gaussian distribution

$$P(\mathbf{w}\,|\,D)\sim N(a\Lambda^{-1}X^T\mathbf{y},(aX^TX+bI)^{-1})$$

而 posterior 的 mean 就是我們想要求的最佳回歸直線的其中一點

$$\text{mean}=a\Lambda^{-1}X^TY=a(aX^TX+bI)^{-1}X^T\mathbf{y}=(\frac{a}{a}X^TX+\frac{b}{a}I)X^T\mathbf{y}=(X^TX+\lambda I)X^T\mathbf{y}$$

和 rLSE 微分得出來的形式是一樣的，故 rLSE 也就是 frequentist 版的 MAP

note(作業第三題需要):

我們可以用這樣的概念去做 online learning，因為我們的 prior 和 posterior 的形式都是 multivariate Gaussion distribution，所以我們只需要找出拿到新 data 後，前一次的 posterior 的 mean, covariance 和下一次的 mean, covariance 是甚麼關係即可。

我們在做第一次的 iteration 時，我們的 prior 的 mean 是 0，covariance 是任意給定的值(課本假設為無限大，反正做越多次 iteration 會越來越小，所以無論選甚麼 covariance 值都沒關係)，我們會得到 posterior 的 mean 和 covariance matrix，也就是前面推導過的 $\begin{array}{l}\Lambda=aX^TX+bI\\ \boldsymbol{\mu}=a\Lambda^{-1}X^T\mathbf{y}\end{array}$

若我們要做第二次 iteration，前一個，也就是第一次的 posterior 的 mean 不為零向量，covariance matrix 也不一定是對角矩陣，故公式需要做修正，還記得吧！本來我們推導的式子是這樣

$$P(\mathbf{w}\,|\,D)\;\alpha\;P(D\,|\,\mathbf{w})P(\mathbf{w})\;\;\alpha\;\;e^{\frac{-a}{2}\sum_i(y_i-X_i\mathbf{w})^2+\frac{-1}{2}(\mathbf{w}-0)bI(\mathbf{w}-0)}=e^{\frac{-a}{2}\sum_i(y_i-X_i\mathbf{w})^2+\frac{-b}{2}\mathbf{w}^T\mathbf{w}}$$

紅圓圈的地方是我們假設 mean 為零向量，covariance matrix 為對角矩陣，故公式需重新推導，假設 prior 的 mean 為 **m**，covariance matrix 的 inverse 為 S

$$P(\mathbf{w}\mid D)\,\alpha\,P(D\mid \mathbf{w})P(\mathbf{w})\ \ \alpha\ \ e^{\frac{-a}{2}\sum_i (y_i - X_i\mathbf{w})^2 + \frac{-1}{2}(\mathbf{w}-\mathbf{m})^T S(\mathbf{w}-\mathbf{m})} = e^{\frac{-a}{2}\sum_i (y_i - X_i\mathbf{w})^2 + \frac{-1}{2}(\mathbf{w}-\mathbf{m})^T S(\mathbf{w}-\mathbf{m})}$$

寫成矩陣表示式為

$$\frac{-a}{2}\sum_i (y_i - X_i\mathbf{w})^2 + \frac{-1}{2}(\mathbf{w}-\mathbf{m})^T S(\mathbf{w}-\mathbf{m}) = \frac{-a}{2}\| X\mathbf{w}-\mathbf{y}\|^2 + \frac{-1}{2}(\mathbf{w}-\mathbf{m})^T S(\mathbf{w}-\mathbf{m})$$
$$= \frac{-a}{2}(\| X\mathbf{w}-\mathbf{y}\|^2 + \frac{1}{a}(\mathbf{w}-\mathbf{m})^T S(\mathbf{w}-\mathbf{m}))$$

忽略係數，整理得

$$a\| X\mathbf{w}-\mathbf{y}\|^2 + (\mathbf{w}-\mathbf{m})^T S(\mathbf{w}-\mathbf{m}) = a(X\mathbf{w}-\mathbf{y})^T (X\mathbf{w}-\mathbf{y}) + (\mathbf{w}-\mathbf{m})^T S(\mathbf{w}-\mathbf{m})$$
$$= a(\mathbf{w}^T X^T X\mathbf{w} - 2\mathbf{w}^T X^T\mathbf{y} + \mathbf{y}^T\mathbf{y}) + (\mathbf{w}^T S\mathbf{w} - 2\mathbf{w}^T S\mathbf{m} + \mathbf{m}^T S\mathbf{m})$$
$$= \mathbf{w}^T(aX^T X + S)\mathbf{w} - 2\mathbf{w}^T(aX^T\mathbf{y} + S\mathbf{m}) + a\mathbf{y}^T\mathbf{y} + \mathbf{m}^T \mathrm{S}\mathbf{m}$$

和 quadratic form 比較
$$(\mathbf{w}-\boldsymbol{\mu})^T \Lambda(\mathbf{w}-\boldsymbol{\mu})$$
$$= \mathbf{w}^T\Lambda\mathbf{w} - 2\mathbf{w}^T\Lambda\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\mu}$$

得
$$\Lambda = aX^T X + \mathrm{S}$$
$$\boldsymbol{\mu} = \Lambda^{-1}(aX^T\mathbf{y} + S\mathbf{m})$$

我們就推導出 posterior 的 mean vector 和 covariance matrix 了！