Bayesian Linear Regression

1. Khái niệm hồi quy tuyến tính:

- Hàm hồi quy tuyến tính là hàm dự báo các kết quả dựa trên những dữ liệu đã cho trước, theo quan hệ tuyến tính.
- Xét hàm $f(x) = \hat{y}$ với x là đầu vào còn $\hat{y}$ là kết quả của hàm, hay còn gọi là kết quả của mô hình.
- Khi đó ta mong muốn các giá trị thực y sẽ có kết quả xấp xĩ với $\hat{y}$.
- Nếu hàm f là một hàm tuyến tính, khi đó mô hình ta xét chính là mô hình hồi quy tuyến tính.

Mô tả toán học của hồi quy tuyến tính:

- Cho $y \in R$ và x là một vector dữ liệu, nếu:

  $y = xw^T$ với w cũng là một vector

Khi đó, ta nói phương trình trên là một mô hình hồi quy tuyến tính theo w.

- Xét (x_i, y_i) là một cặp dữ liệu input – outcome và n là số các cặp dữ liệu có dạng trên.
- Để output của mô hình ăn khớp nhất có thể với những bộ input – outcome này, ta cần sự chênh lệch giữa các output và outcome phải đủ nhỏ:

$$e^2 = (y - \hat{y})^2$$

Đạt giá trị nhỏ nhất.

- Nói cách khác, bài toán linear regression sẽ trở thành một bài toán ước lượng quen thuộc, đó là tìm w sao cho hàm mất mát:

$$L(Y, \hat{Y}) = \sum(\hat{y}_i - y_i)^2 = \sum(w^T x_i - y_i)^2$$

Đạt giá trị nhỏ nhất.

- Các bài toán sử dụng các mô hình hồi quy hay mô hình xác suất nói chung đều dẫn tới việc phải áp dụng các phương pháp ước lượng điểm.
- Có rất nhiều thuật toán liên quan đến Machine Learning phải sử dụng tham số. Trong bài này, ta sẽ tìm hiểu về trường hợp đơn giản nhất hàm hồi quy tuyến tính.

2. Cách tiếp cận theo trường phái cổ điển:

- Xét trường hợp bài toán hồi quy tuyến tính không có hệ số chặn:

$$y = \theta x$$

Trong đó x và y là data của bài toán, theta là hệ số góc của hàm hồi quy tuyến tính.

Nhắc lại:

- Trường phái cổ điển sẽ tính toán tham số cho bài toán thông qua Maximum likelihood.
- Cách tính: xét mô hình xác suất có data D, tham số θ, khi đó, ta sẽ ước lượng tham số thông qua likelihood của mô hình

$$\mathrm{L}(\theta) = \mathrm{Log}\,\mathrm{P}(\mathrm{D}|\theta)$$

- Cụ thể, ta tìm: $\mathrm{argmax}_\theta \mathrm{L}(\theta) = \mathrm{argmin}_\theta - \mathrm{L}(\theta)$

- Quay trở lại với bài toán hồi quy tuyến tính ở trên.
- Ta giả sử bộ dữ liệu có phân phối Gauss:

$$\mathrm{y}|\mathrm{x};\theta \sim \mathrm{N}(\theta\mathrm{x}, \sigma^2)$$

Lưu ý: thông thường ta kí hiệu là μ, trong trường hợp này thì $\mu = \mathrm{y} = \theta\mathrm{x}$.

- Ta sẽ ước lượng tham số mô hình trên như sau:

$$\hat{\theta} = \text{argmax}_{\theta} \text{L}(\theta)$$

$$= \text{argmax}_{\theta} \text{Log P}(\text{y}|\text{x};\theta)$$

$$= \text{argmax}_{\theta} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \text{e}^{-\frac{(\text{y}-\theta\text{x})^2}{2\sigma^2}}\right)$$

- Như vậy, ta được:

$$\hat{\theta} = \text{argmax}_{\theta} \left(\text{Log} \frac{1}{\sqrt{2\pi}\sigma} + \text{Log e}^{-\frac{(\text{y}-\theta\text{x})^2}{2\sigma^2}} \right)$$

$$= \text{argmax}_{\theta} -\frac{(\text{y} - \theta\text{x})^2}{2\sigma^2}$$

- Sau khi nhận được theta bằng việc training bộ dữ liệu, ta nhận được mô hình dự đoán cho các điểm data $(\mathrm{x}^*, \mathrm{y}^*)$ với:

$$\mathrm{y}^* = \theta \mathrm{x}^*$$

- Ở đây $(\mathrm{x}^*, \mathrm{y}^*)$ là những điểm dữ liệu sẽ xuất hiện trong tương lai và ta hi vọng nó sẽ thoả phương trình tuyến tính theo theta trên.

- Cách tiếp cận cổ điển như trên không có tính toán trực tiếp cho theta, nên phải sử dụng các ước lượng bằng khoảng tin cậy để dò ra được tham số cần thiết.
- Phương pháp trên khó tính toán chính xác.
- Tuy nhiên lợi thế của nó là có thể làm việc với trường hợp bộ dữ liệu rất lớn.

- Như ta đã biết, phương pháp Bayes là một phương pháp tiếp cận khác cho các bài toán ước lượng kiểu trên để có những tính toán chính xác hơn.
- Theo như bài trước, ước lượng tham số thông qua kiểm định kiểu Bayes không cần dựa vào khoảng tin cậy.
- Tuy nhiên phương pháp Bayes luôn đối mặt với vấn đề nhiều tính toán phức tạp.
- Chỉ có thể dùng tốt trong các trường hợp dữ liệu vừa phải.

3. Hồi quy tuyến tính Bayes:

- Với bài toán hồi quy tuyến tính, ta luôn có bộ dữ liệu có phân phối Gauss:

$$\mathrm{y}|\mathrm{x};\theta \sim \mathrm{N}(\theta\mathrm{x}, \sigma^2)$$

- Tuy nhiên, ta không ước lượng θ như trước, mà coi θ là một biến ngẫu nhiên có phân phối xác suất.

- Việc chọn xác suất tiên nghiệm thông thường dựa trên kinh nghiệm của người làm, thông thường ta sẽ chọn xác suất tiên nghiệm sao cho xác suất hậu nghiệm về sau sẽ có dạng giống với xác suất tiên nghiệm và likelihood.
- Trong trường hợp phân phối Gauss, ta chọn xác suất tiên nghiệm là một phân phối chuẩn.

- Như vậy, bài toán của ta sẽ có những thông tin cần thiết sau:

$$\mathrm{y}|\mathrm{x};\theta \sim \mathrm{N}(\theta\mathrm{x},\sigma^2) \text{ và } \mathrm{P}(\mathrm{D}|\theta) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(\mathrm{y}-\theta\mathrm{x})^2}{2\sigma^2}}$$

$$\theta \sim \mathrm{N}(0,1) \text{ và } \mathrm{P}(\theta) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{\theta^2}{2}}$$

- Bước tiếp theo, ta tính xác suất hậu nghiệm $\mathrm{P}(\theta|\mathrm{D})$:

$$\mathrm{P}(\theta|\mathrm{D}) = \frac{\mathrm{P}(\theta, \mathrm{D})}{\mathrm{P}(\mathrm{D})}$$

$$= \frac{\mathrm{P}(\mathrm{D}|\theta).\mathrm{P}(\theta)}{\mathrm{P}(\mathrm{D})}$$

- Bằng các công thức tính xác suất biên, ta có:

$$P(\theta|D) = \frac{P(y|x;\ \theta).P(\theta)}{\int P(y|x).P(\theta)}$$

$$= \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\theta x)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}}}{\int \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\theta x)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta^2}{2}} \right) d\theta}$$

- Như vậy, ta được:

$$\mathrm{P}(\theta|\mathrm{D}) \propto \frac{1}{2\pi\sigma} \mathrm{e}^{-\frac{(\mathrm{y}-\theta\mathrm{x})^2}{2\sigma^2}-\frac{\theta^2}{2}}$$

- Chú ý ta có thể coi $\mathrm{P}(\mathrm{D}) = \mathrm{P}(\mathrm{y}|\mathrm{x})$ là một hằng số dựa trên data đã cho từ trước.

- Cũng như phương pháp cổ điển, ta cần đưa ra mô hình dự báo theo cách tiếp cận Bayes.
- Mô hình dự báo theo cách tiếp cận Bayes sẽ được mô tả như sau:

$$\mathrm{P}(\mathrm{y}^*|\mathrm{x}^*, \mathrm{D}) = \int \mathrm{P}(\mathrm{y}^*|\mathrm{x}^*, \mathrm{D}, \theta)\mathrm{d}\theta$$

$$= \int \mathrm{P}(\mathrm{y}^*|\mathrm{x}^*, \theta).\,\mathrm{P}(\theta|\mathrm{D})\mathrm{d}\theta$$

- Lưu ý: ta có thể viết lại đơn giản thành:

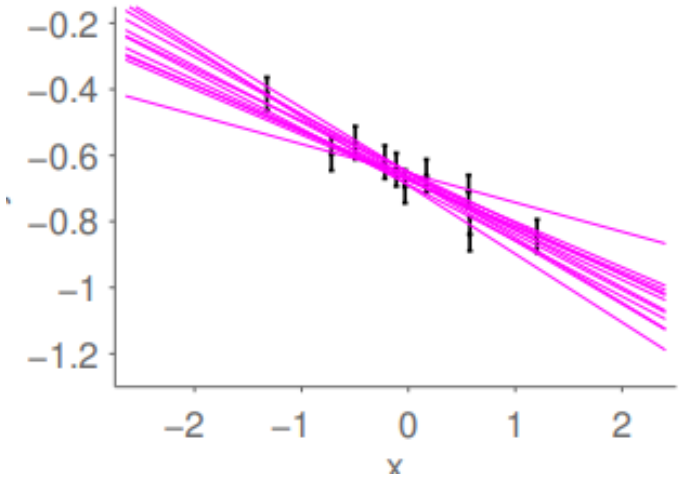
$$\mathrm{P}(\mathrm{x}^*|\mathrm{D}) = \int \mathrm{P}(\mathrm{x}^*, \theta|\mathrm{D})\mathrm{d}\theta$$

Vì xác suất của chúng ta hiện tại là xác suất có điều kiện D. Vì thế ta được kết quả:

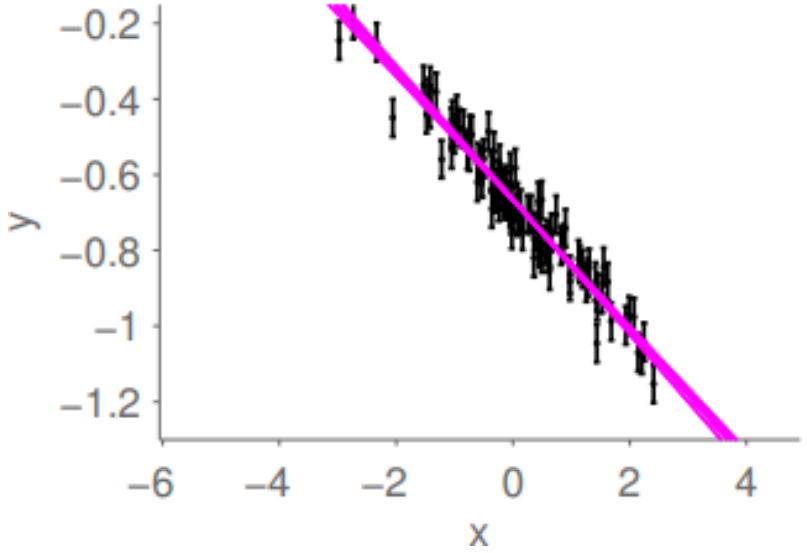
$$\mathrm{P}(\mathrm{x}^*|\mathrm{D}) = \int \mathrm{P}(\mathrm{x}^*, \theta|\mathrm{D})\mathrm{d}\theta = \int \mathrm{P}(\mathrm{x}^*|\mathrm{D}).\mathrm{P}(\theta|\mathrm{D})\mathrm{d}\theta$$

Ví dụ bằng hình ảnh:

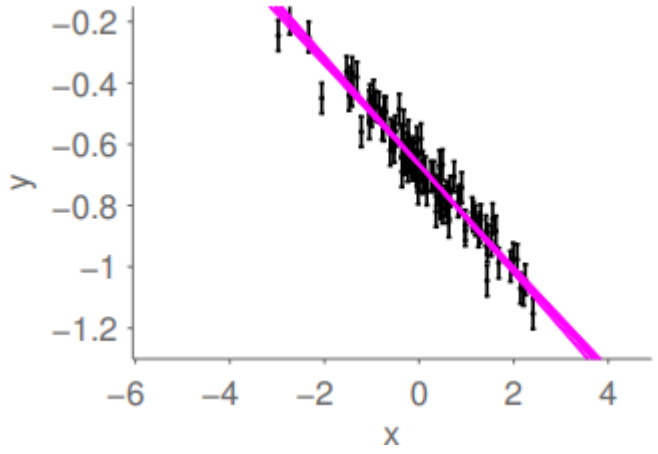
- Giả sử chúng ta có rất nhiều bộ dữ liệu, chúng ta muốn tìm hiểu rằng liệu có một phương trình nào đó mô tả toàn bộ các dữ liệu hiện đang có hay không.
- Ví dụ như hình dưới:



- Ta thường muốn một đường duy nhất, tuy nhiên, bằng cảm nhận của bản thân, ta hoàn toàn có thể vẽ ra rất nhiều đường và tìm trong đó đường ta cho là tốt nhất:



- Quan sát thêm rất nhiều điểm, ta có thể thấy các đường thẳng này có vẻ như gộp lại thành một đường duy nhất:



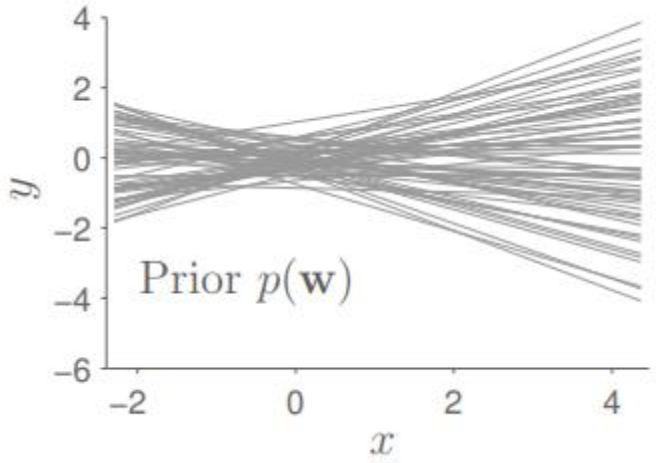
- Thế nên, ta có thể đặt ra câu hỏi: làm thế nào để có thể liên tục vẽ ra thật nhiều đường và cạnh như trên một cách có hệ thống và chính xác ?
- Để ý rằng ta có phương trình tuyến tính:

$$\mathrm{y} = \theta_1 \mathrm{x} + \theta_2$$

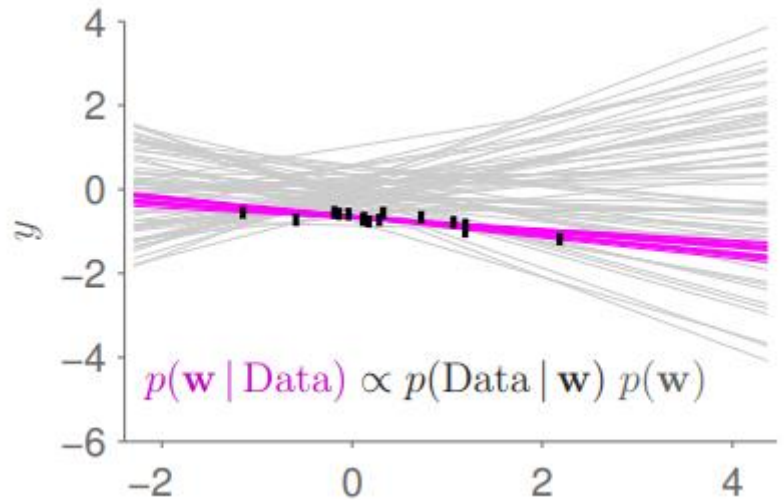
- Như vậy thay đổi $\theta = (\theta_1, \theta_2)$, ta sẽ được các đường thẳng khác nhau.

- Ý tưởng để các đường này di chuyển: thay vì coi θ cố định, ta cho θ thay đổi. Vì bài toán hiện tại làm việc với mô hình xác suất, ta hi vọng θ là một biến ngẫu nhiên.
- Như vậy, để tìm ra những chùm đường thẳng như vậy, ý tưởng sử dụng phương pháp tiếp cận Bayes là điều cần thiết.
- Như thế, khi hữu hình hoá các dữ liệu (visualize data), phương pháp hồi quy tuyến tính Bayes sẽ hoạt động khác đi.

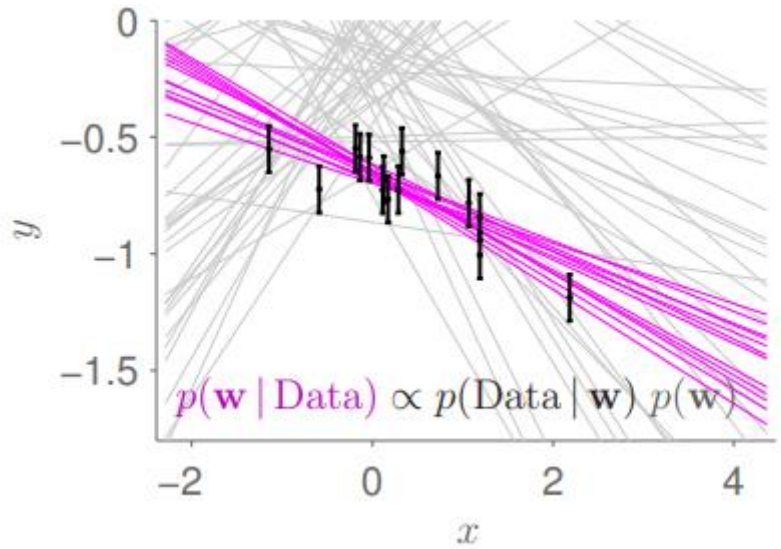
- Tiên nghiệm (prior):



- G


4
2
0
−2
−4
−6
y
−2
0
2
4
$p(\mathbf{w} \mid \text{Data}) \propto p(\text{Data} \mid \mathbf{w})\ p(\mathbf{w})$

- p


0
−0.5
−1
−1.5
y
−2
0
2
4
x
$p(\mathbf{w} \mid \text{Data}) \propto p(\text{Data} \mid \mathbf{w})\, p(\mathbf{w})$

- Ta sẽ xét dạng tổng quát của bài toán:

$$y_i = x_i w^T + \epsilon_i$$

Ta đặt thêm ϵ_i vì trên thực tế việc $y_i = x_i w^T$ là không thể xảy ra với mọi cặp (x_i, y_i).

- Giả sử ta có các xác suất tiên nghiệm:

$$\mathrm{w} \sim \mathrm{N}(0, \lambda^{-1}\mathrm{I})$$

$$\epsilon_{\mathrm{i}} \sim \mathrm{N}(0, \sigma^2)$$

- Ta có phân phối xác suất cho likelihood:

$$y_i | x_i w^T, \sigma^2 \sim N(x_i w^T, \sigma^2)$$

- Để ý rằng:

$$p\left(y_i \middle| x_i w^T, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(y_i - x_i . w^T\right)^2}{2\sigma^2}}$$

- Ta giả sử n bộ dữ liệu (x_i, y_i) là độc lập, khi đó ta được:

$$p(y_1, \ldots, y_n | w, X) = \prod_{i=1}^{n} p(y_i | w, x_i)$$

Lưu ý vì đã “cố định” sigma nên ta không đưa sigma vào.

- Khi đó ta sẽ có xác suất hậu nghiệm:

$$P(w|X,y) = \frac{p(y|X,w)p(w)}{\int_{R^d} p(y|X,w)p(w)dw}$$

$$= \frac{\prod_i p(y_i|x_i,w)p(w)}{\int_{R^d} \prod_i p(y_i|x_i,w)p(w)\,dw}$$

- Sử dụng tính chất tỉ lệ thuận, ta có:

$$\mathrm{P}(\mathrm{w}|\mathrm{X},\mathrm{y}) \propto \prod_{\mathrm{i}} p(\mathrm{y_i}|\mathrm{x_i}, w)p(w)$$

$$\propto \prod_{\mathrm{i}} \mathrm{e}^{-\frac{\left(\mathrm{y_i}-\mathrm{x_i}.\mathrm{w}^{\mathrm{T}}\right)^2}{2\sigma^2}} . \mathrm{e}^{-\frac{\lambda}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w}}$$

- Kết quả của phép nhân ở trên là:

$$P(w|X,y) \propto e^{-\left(\sum \frac{\left(y_i - x_i . w^T\right)^2}{2\sigma^2} + \frac{\lambda}{2} w w^T\right)}$$

$$= e^{-\frac{1}{2}\left[w^T\left(\lambda I + \frac{1}{\sigma^2}\sum_i x_i . x_i^T\right)w - 2w^T\left(\frac{1}{\sigma^2}\sum_i y_i x_i\right) + \frac{1}{\sigma^2}\sum_i y_i^2\right]}$$

- Tiếp đó, ta sẽ biến đổi thành:

$$P(w|X,y) \propto e^{-\frac{1}{2}\left[w^T\left(\lambda I+\frac{1}{\sigma^2}\sum_i x_i.x_i^T\right)w-2w^T\left(\frac{1}{\sigma^2}\sum_i y_i x_i\right)+\frac{1}{\sigma^2}\sum_i y_i^2\right]}$$

$$\propto e^{-\frac{1}{2}\left[w^T\left(\lambda I+\frac{1}{\sigma^2}\sum_i x_i.x_i^T\right)w-2w^T\left(\frac{1}{\sigma^2}\sum_i y_i x_i\right)\right]}.\, e^{\frac{1}{\sigma^2}\sum_i y_i^2}$$

$$\propto e^{-\frac{1}{2}\left[w^T\left(\lambda I+\frac{1}{\sigma^2}\sum_i x_i.x_i^T\right)w-2w^T\left(\frac{1}{\sigma^2}\sum_i y_i x_i\right)\right]}$$

- Lưu ý, những biến đổi trên có được là vì ta làm việc với các phép tương đương tỉ lệ thuận.
- Ta hoàn toàn có thể sử dụng tính toán bằng likelihood, nhưng không khuyến khích vì với không gian vector nhiều chiều thì ta phải làm việc với quá nhiều dấu tích phân để tìm ra kết quả.
- Mục tiêu của các biến đổi trên là để cố gắng đưa xác suất hậu nghiệm về một dạng hàm mật độ của một phân phối xác suất.

- Để đưa được về dạng hàm mật độ, ta cần nhân biểu thức

$$P(w|X,y) \propto e^{-\frac{1}{2}\left[w^T\left(\lambda I+\frac{1}{\sigma^2}\sum_i x_i.x_i^T\right)w-2w^T\left(\frac{1}{\sigma^2}\sum_i y_i x_i\right)\right]}$$

Với biểu thức

$$e^{-\frac{1}{2}\left[\left(\frac{1}{\sigma^2}\sum_i y_i x_i\right)^T\left(\lambda I+\frac{1}{\sigma^2}\sum_i x_i x_i^T\right)^{-1}\left(\frac{1}{\sigma^2}\sum_i y_i x_i\right)\right]}$$

- Ta được :

$$P(w|X,y)$$

$$\propto e^{-\frac{1}{2}\left[w^T\left(\lambda I+\frac{1}{\sigma^2}\sum_i x_i.x_i^T\right)w-2w^T\left(\frac{1}{\sigma^2}\sum_i y_i x_i\right)+\left(\frac{1}{\sigma^2}\sum_i y_i x_i\right)^T\left(\lambda I+\frac{1}{\sigma^2}\sum_i x_i x_i^T\right)^{-1}\left(\frac{1}{\sigma^2}\sum_i y_i x_i\right)\right]}$$

$$\propto e^{-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)}$$

- Trong đó:

$$\Sigma = \left(\lambda \mathrm{I} + \frac{1}{\sigma^2} \sum_{\mathrm{i}} \mathrm{x_i} . \mathrm{x_i^T} \right)^{-1}$$

$$\mu = \Sigma \left(\frac{1}{\sigma^2} \sum_{\mathrm{i}} \mathrm{y_i x_i} \right)$$

- Đưa về dạng tính toán chính xác, ta có:

$$P(w|X,y) = \frac{e^{-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)}}{\int_{R^d} e^{-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)} dw}$$

- Kết quả thu nhận được:

$$\mathrm{P}(\mathrm{w}|\mathrm{X},\mathrm{y}) = \mathrm{P}(\mathrm{w}|\Sigma,\mu)$$

$$= \frac{\mathrm{e}^{-\frac{1}{2}(\mathrm{w}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathrm{w}-\mu)}}{(2\pi)^{\frac{\mathrm{d}}{2}}.|\Sigma|^{\frac{1}{2}}}$$

- Như vậy, ta có thể kết luận rằng:

$$\mathrm{w}|\mathrm{X}, \mathrm{y} \sim \ \mathrm{N}(\mu, \Sigma)$$

Trong đó:

$$\Sigma = \left(\lambda \mathrm{I} + \frac{1}{\sigma^2} \sum_{\mathrm{i}} \mathrm{x_i}.\mathrm{x_i^T}\right)^{-1}$$

$$\mu = \Sigma \left(\frac{1}{\sigma^2} \sum_{\mathrm{i}} \mathrm{y_i x_i}\right)$$

Mô hình dự báo:

- Ta sẽ đến với việc dự báo các điểm dữ liệu mới.
- Giả sử $\hat{x}$ là dữ liệu tương lai, khi đó ta có:

$$\hat{y}|\hat{x} \sim p(y|x, w)$$

Và

$$P(\hat{y}|\hat{x}, y, X) = \int p(\hat{y}|\hat{x}, w)p(w|y, X)dw$$

- Bằng các phương pháp tính toán và biến đổi phức tạp (tương tự như phần trên), ta được:

$$p(\hat{y}|\hat{x}, y, X) = \frac{e^{-\frac{1}{2(\sigma^2+\hat{x}^T\Sigma\hat{x})}(\hat{y}-\hat{x}^T\mu)^2}}{(2\pi)^{\frac{1}{2}}(\sigma^2+\hat{x}^T\Sigma\hat{x})^{\frac{1}{2}}}$$