UNIVERSIDADE FEDERAL DO CEARÁ

DEPARTAMENTO DE COMPUTAÇÃO

LogIA

Aprendizagem de Máquina Probabilística

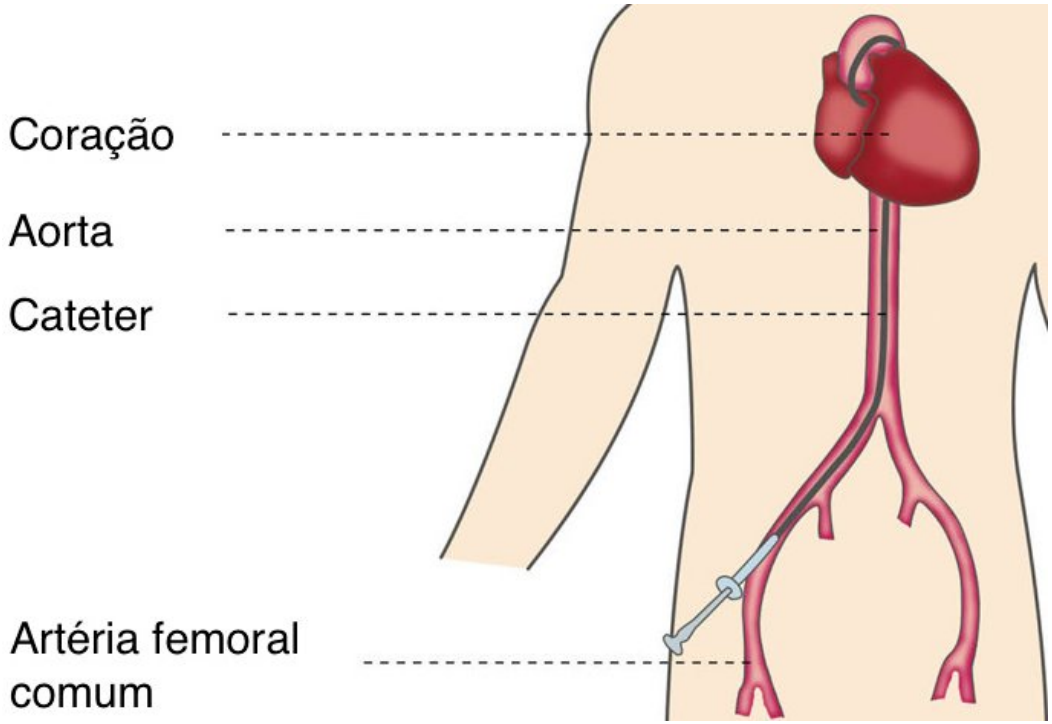
César Lincoln Cavalcante Mattos

2025

Agenda

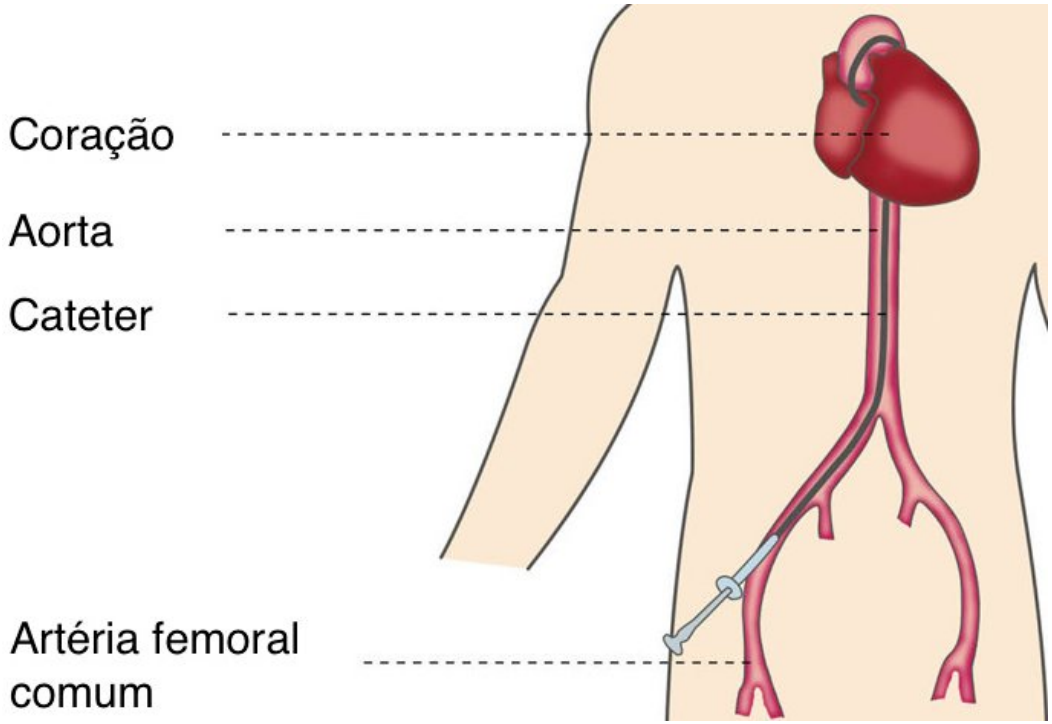
Regressão linear simples

- **Cataterismo cardíaco**: procedimento de inserção de um cateter (usualmente pela artéria femural) para diagnosticar problemas de obstrução no coração.



Regressão linear simples

- **Cataterismo cardíaco**: procedimento de inserção de um cateter (usualmente pela artéria femural) para diagnosticar problemas de obstrução no coração.



- **Problema**: Dada a **altura** de um paciente, qual o **comprimento** do cateter necessário para alcançar seu coração?

Regressão linear simples

- Considere a tabela a seguir relacionando alturas de jovens pacientes e comprimentos do cateter correspondente:

Altura (m)	**Comprimento (cm)**
1.087	37
1.613	50
0.953	34
1.003	36
1.156	43
0.978	28
1.092	37
0.572	20
0.940	34
0.597	30
0.838	38
1.473	47

Regressão linear simples

- Considere a tabela a seguir relacionando alturas de jovens pacientes e comprimentos do cateter correspondente:

Altura (m)	**Comprimento (cm)**
1.087	37
1.613	50
0.953	34
1.003	36
1.156	43
0.978	28
1.092	37
0.572	20
0.940	34
0.597	30
0.838	38
1.473	47

- **Problema**: Dado uma altura **não presente** na tabela, qual deverá ser o comprimento do cateter?

Regressão linear simples

- A coluna **Altura** é a **entrada** do nosso modelo.
- A coluna **Comprimento** é a **saída** do nosso modelo.
- Nosso **conjunto de dados** é formado por 12 alturas e 12 comprimentos correspondentes.

Regressão linear simples

- A coluna **Altura** é a **entrada** do nosso modelo.
- A coluna **Comprimento** é a **saída** do nosso modelo.
- Nosso **conjunto de dados** é formado por 12 alturas e 12 comprimentos correspondentes.
- Matematicamente, temos:

$$\mathcal{D} = \{(x_1, y_1), \cdots, (x_2, y_2)\} = \{(x_i, y_i)\}_{i=1}^{12},$$

em que x_i é a i-ésima entrada e y_i é a i-ésima saída.

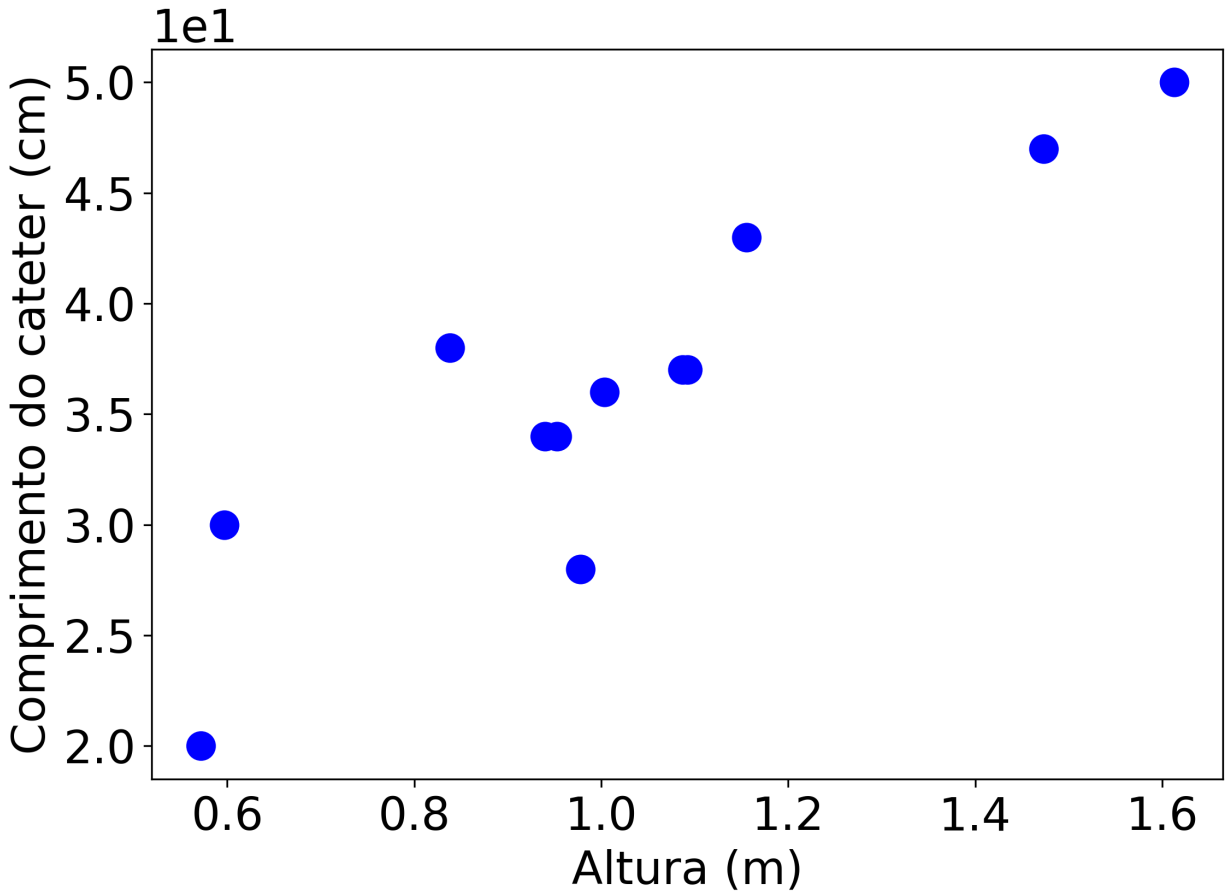
Regressão linear simples

- A coluna **Altura** é a **entrada** do nosso modelo.
- A coluna **Comprimento** é a **saída** do nosso modelo.
- Nosso **conjunto de dados** é formado por 12 alturas e 12 comprimentos correspondentes.
- Matematicamente, temos:

$$\mathcal{D} = \{(x_1, y_1), \cdots, (x_2, y_2)\} = \{(x_i, y_i)\}_{i=1}^{12},$$

  em que x_i é a i-ésima entrada e y_i é a i-ésima saída.
- **Objetivo**: Encontrar uma relação entre x_i e y_i que forneça uma predição $\hat{y}_i$ o mais próximo possível da saída real y_i.

Regressão linear simples



Regressão linear simples

Terminologia

- **Atributo (feature)**: Uma dada característica de um padrão.

Regressão linear simples

Terminologia

- **Atributo (feature)**: Uma dada característica de um padrão.
- **Padrão (pattern)**: Um vetor de atributos que representa um exemplo.

Regressão linear simples

Terminologia

- **Atributo (feature)**: Uma dada característica de um padrão.
- **Padrão (pattern)**: Um vetor de atributos que representa um exemplo.
- **Modelo**: Uma função que expressa a relação entre um padrão de entrada e sua saída correspondente.

Regressão linear simples

Terminologia

- **Atributo (feature)**: Uma dada característica de um padrão.
- **Padrão (pattern)**: Um vetor de atributos que representa um exemplo.
- **Modelo**: Uma função que expressa a relação entre um padrão de entrada e sua saída correspondente.
- **Função custo (ou função objetivo)**: Indica o quão mal (ou o quão bem) um modelo aproxima os dados disponíveis.

Regressão linear simples

Terminologia

- **Atributo (feature)**: Uma dada característica de um padrão.
- **Padrão (pattern)**: Um vetor de atributos que representa um exemplo.
- **Modelo**: Uma função que expressa a relação entre um padrão de entrada e sua saída correspondente.
- **Função custo (ou função objetivo)**: Indica o quão mal (ou o quão bem) um modelo aproxima os dados disponíveis.
- **Parâmetros**: Variáveis que caracterizam o modelo proposto.

Regressão linear simples

Terminologia

- **Atributo (feature)**: Uma dada característica de um padrão.
- **Padrão (pattern)**: Um vetor de atributos que representa um exemplo.
- **Modelo**: Uma função que expressa a relação entre um padrão de entrada e sua saída correspondente.
- **Função custo (ou função objetivo)**: Indica o quão mal (ou o quão bem) um modelo aproxima os dados disponíveis.
- **Parâmetros**: Variáveis que caracterizam o modelo proposto.
- **Risco empírico**: Estimativa do risco (custo) obtida a partir dos dados disponíveis.

Regressão linear simples

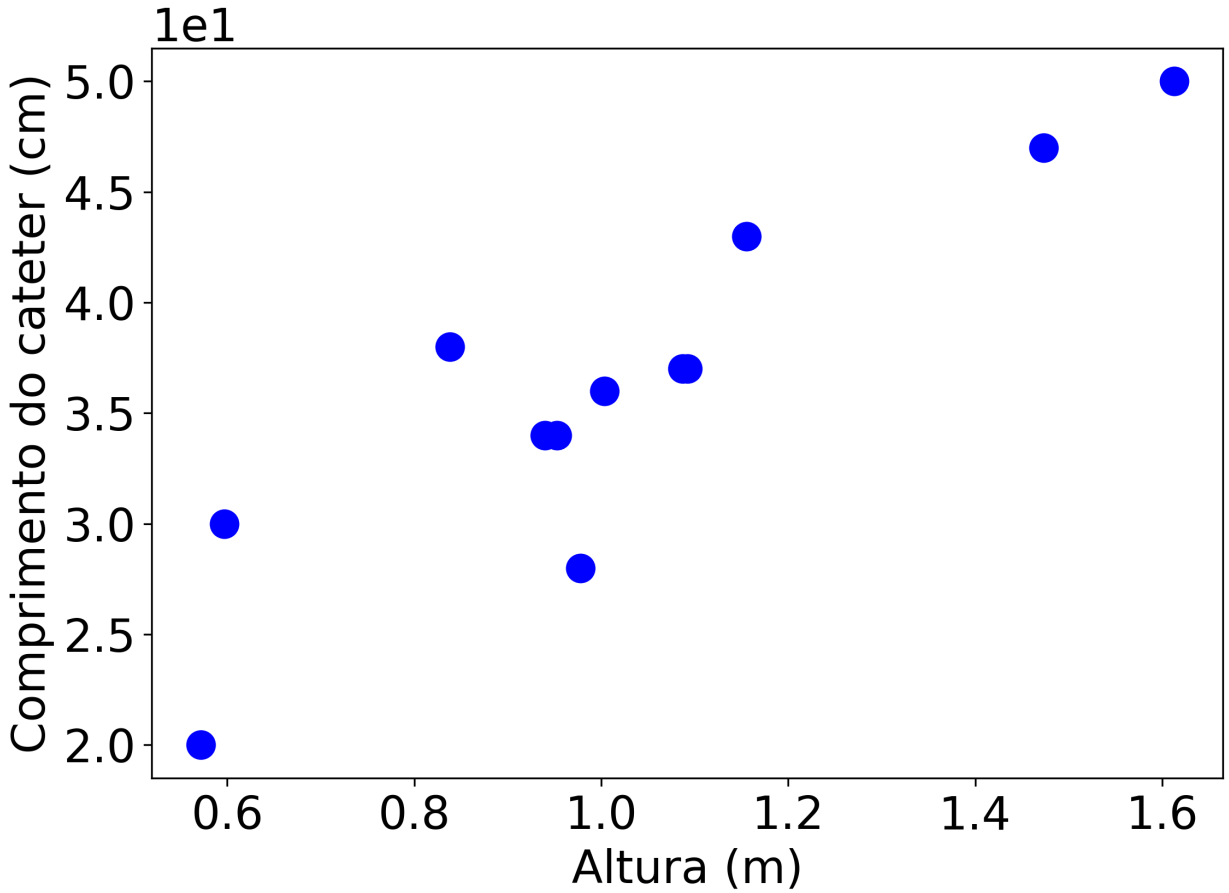
Terminologia

- **Atributo (feature)**: Uma dada característica de um padrão.
- **Padrão (pattern)**: Um vetor de atributos que representa um exemplo.
- **Modelo**: Uma função que expressa a relação entre um padrão de entrada e sua saída correspondente.
- **Função custo (ou função objetivo)**: Indica o quão mal (ou o quão bem) um modelo aproxima os dados disponíveis.
- **Parâmetros**: Variáveis que caracterizam o modelo proposto.
- **Risco empírico**: Estimativa do risco (custo) obtida a partir dos dados disponíveis.
- **Otimização (ou treinamento, aprendizagem)**: Algoritmo de obtenção dos parâmetros do modelo que minimizem uma função custo (ou maximizem uma função objetivo).

Regressão linear simples

- Considere uma relação linear entre x_i (entrada do modelo) e $\hat{y}_i$ (saída do modelo):

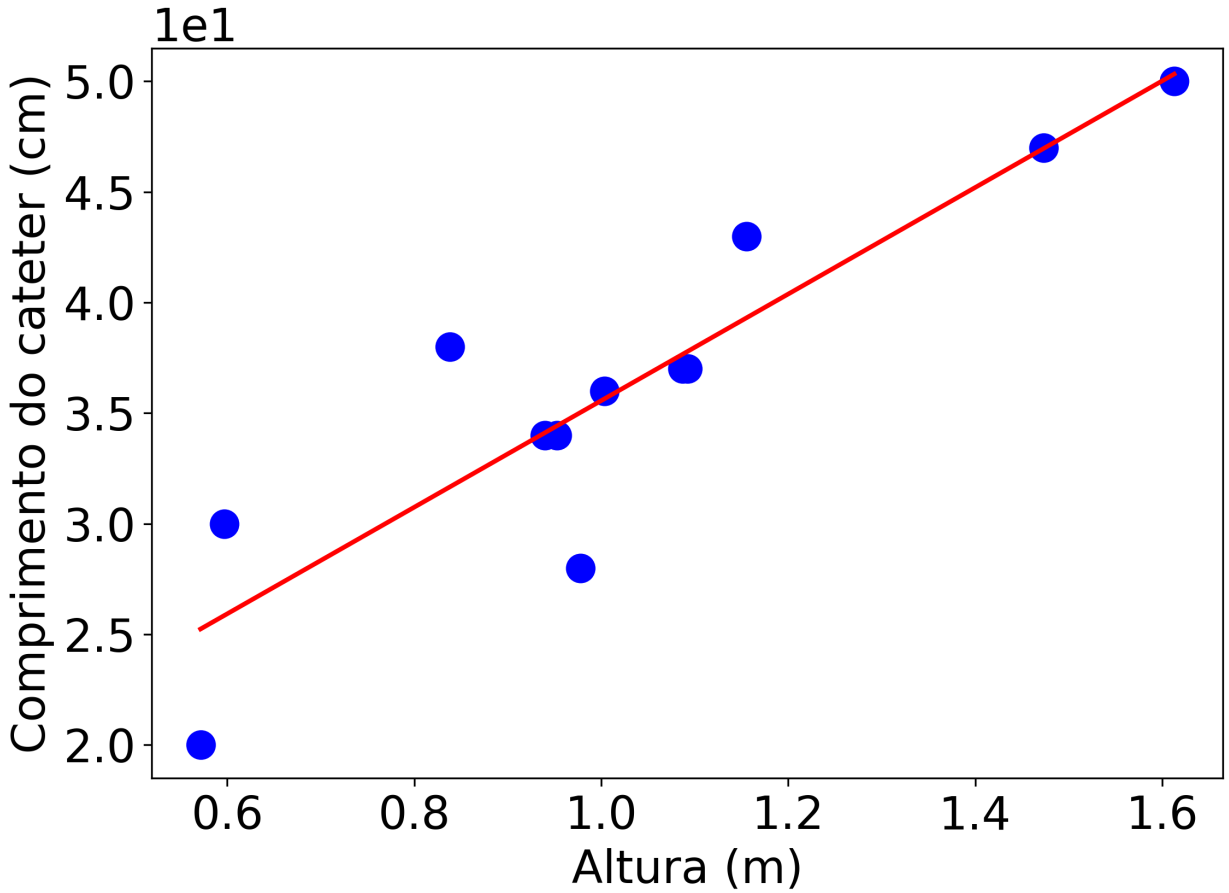
$$\hat{y}_i = w_0 + w_1 x_i.$$



Regressão linear simples

- Considere uma relação linear entre x_i (entrada do modelo) e $\hat{y}_i$ (saída do modelo):

$$\hat{y}_i = w_0 + w_1 x_i.$$



Regressão linear simples

- Escolhemos uma função custo quadrática para os erros obtidos pelo modelo:

$$\mathcal{J}(w_0, w_1) = \frac{1}{2N} \sum_{i=1}^{N} e_i^2,$$

$$e_i = y_i - \hat{y}_i.$$

Regressão linear simples

- Escolhemos uma função custo quadrática para os erros obtidos pelo modelo:

$$\mathcal{J}(w_0, w_1) = \frac{1}{2N} \sum_{i=1}^{N} e_i^2,$$

$$e_i = y_i - \hat{y}_i.$$

- Esse custo é chamado **Erro Quadrático Médio (MSE, Mean Squared Error)**.

Regressão linear simples

- Desejamos minimizar a função custo em relação aos parâmetros do modelo:

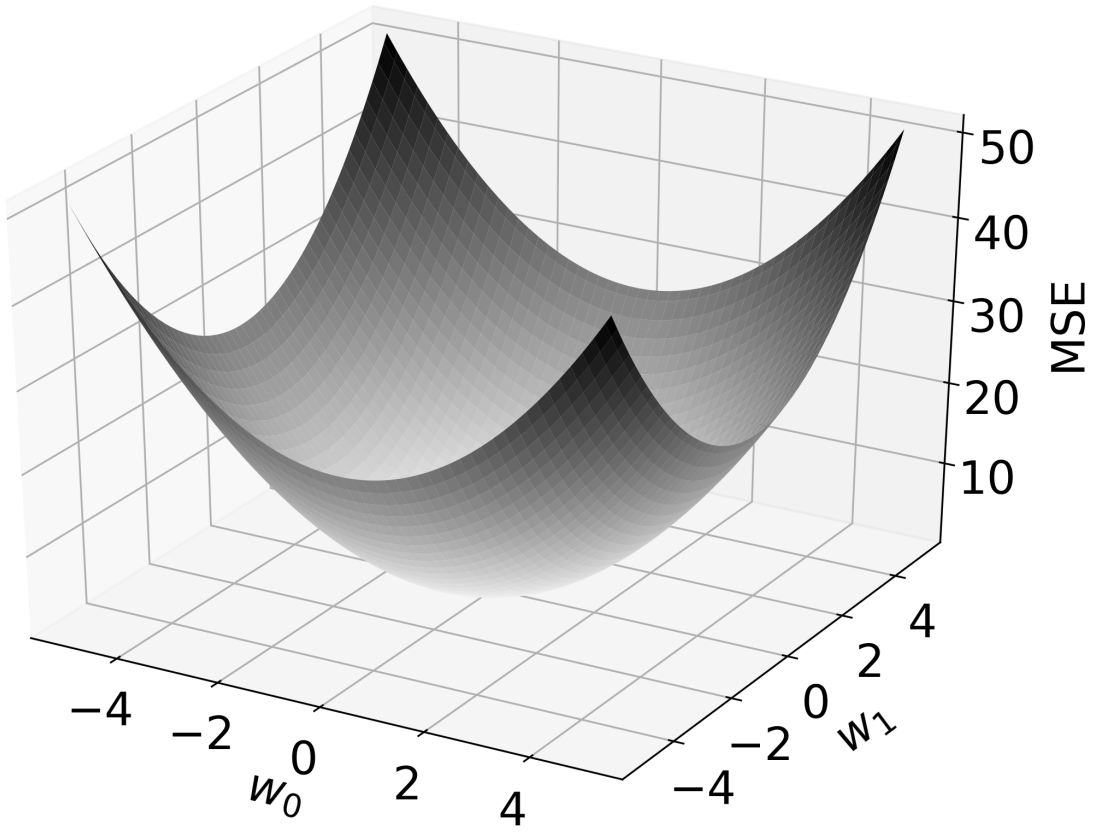
$$\min_{w_0, w_1} \mathcal{J}(w_0, w_1)$$

$$\min_{w_0, w_1} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$\min_{w_0, w_1} \frac{1}{2N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2$$

Regressão linear simples

$$\min_{w_0, w_1} \frac{1}{2N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2$$



Regressão linear simples

- Escolhemos valores iniciais para w_0 e w_1.
- Movimentamos os parâmetros na direção que diminui a função custo $\mathcal{J}(w_0, w_1)$:

$$w_0 \leftarrow w_0 - \alpha \frac{\partial \mathcal{J}}{\partial w_0},$$
$$w_1 \leftarrow w_1 - \alpha \frac{\partial \mathcal{J}}{\partial w_1}$$

- $\alpha > 0$ é um **passo de aprendizado**.

Regressão linear simples

$$\frac{\partial \mathcal{J}}{\partial w_0} = \frac{\partial}{\partial w_0} \frac{1}{2N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2$$

$$\frac{\partial \mathcal{J}}{\partial w_0} = \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)(-1)$$

$$\frac{\partial \mathcal{J}}{\partial w_0} = -\frac{1}{N} \sum_{i=1}^{N} e_i$$

Regressão linear simples

$$\frac{\partial \mathcal{J}}{\partial w_0} = \frac{\partial}{\partial w_0} \frac{1}{2N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2$$

$$\frac{\partial \mathcal{J}}{\partial w_0} = \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)(-1)$$

$$\frac{\partial \mathcal{J}}{\partial w_0} = -\frac{1}{N} \sum_{i=1}^{N} e_i$$

- Logo:

$$w_0 \leftarrow w_0 + \alpha \frac{1}{N} \sum_{i=1}^{N} e_i$$

Regressão linear simples

$$\frac{\partial \mathcal{J}}{\partial w_1} = \frac{\partial}{\partial w_1} \frac{1}{2N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2$$

$$\frac{\partial \mathcal{J}}{\partial w_1} = \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)(-x_i)$$

$$\frac{\partial \mathcal{J}}{\partial w_1} = -\frac{1}{N} \sum_{i=1}^{N} e_i x_i$$

Regressão linear simples

$$\frac{\partial \mathcal{J}}{\partial w_1} = \frac{\partial}{\partial w_1} \frac{1}{2N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2$$

$$\frac{\partial \mathcal{J}}{\partial w_1} = \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)(-x_i)$$

$$\frac{\partial \mathcal{J}}{\partial w_1} = -\frac{1}{N} \sum_{i=1}^{N} e_i x_i$$

- Logo:

$$w_1 \leftarrow w_1 + \alpha \frac{1}{N} \sum_{i=1}^{N} e_i x_i$$

Regressão linear simples

Gradiente Descendente (GD, gradient descent)

1. Escolha um valor α positivo e pequeno.
2. Inicialize os parâmetros do modelo na iteração $t = 0$.
3. Repita por diversas iterações (épocas):
   1. $t \leftarrow t + 1$;
   2. Calcule os erros do modelo:

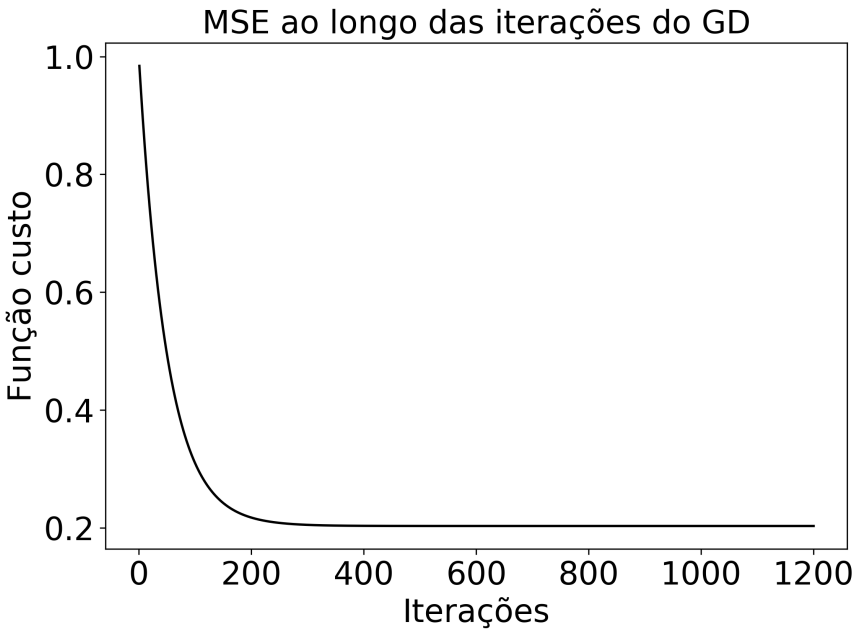
$$\hat{y}_i(t-1) = w_0(t-1) + w_1(t-1)x_i, \quad \forall i,$$
$$e_i(t-1) = y_i - \hat{y}_i(t-1), \quad \forall i.$$

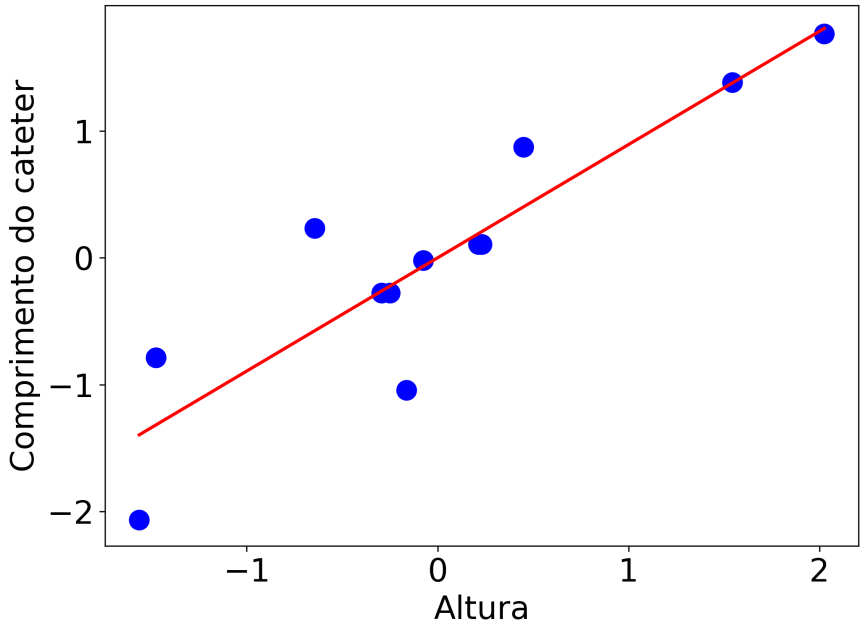
   3. Atualize os parâmetros:

$$w_0(t) = w_0(t-1) + \alpha \frac{1}{N} \sum_{i=1}^{N} e_i(t-1)$$

$$w_1(t) = w_1(t-1) + \alpha \frac{1}{N} \sum_{i=1}^{N} e_i(t-1)x_i$$

Regressão linear simples - Otimização via GD





Regressão linear simples

Gradiente Descendente Estocástico (SGD, stochastic gradient descent)

1. Escolha um valor α positivo e pequeno.
2. Inicialize os parâmetros do modelo na iteração $t = 0$.
3. Repita por diversos ciclos (épocas):
   1. Permute aleatoriamente a ordem dos dados.
   2. Para cada padrão de entrada, $i = 1, \ldots, N$, repita:
      1. Faça $t \leftarrow t + 1$.
      2. Calcule os erros do modelo:

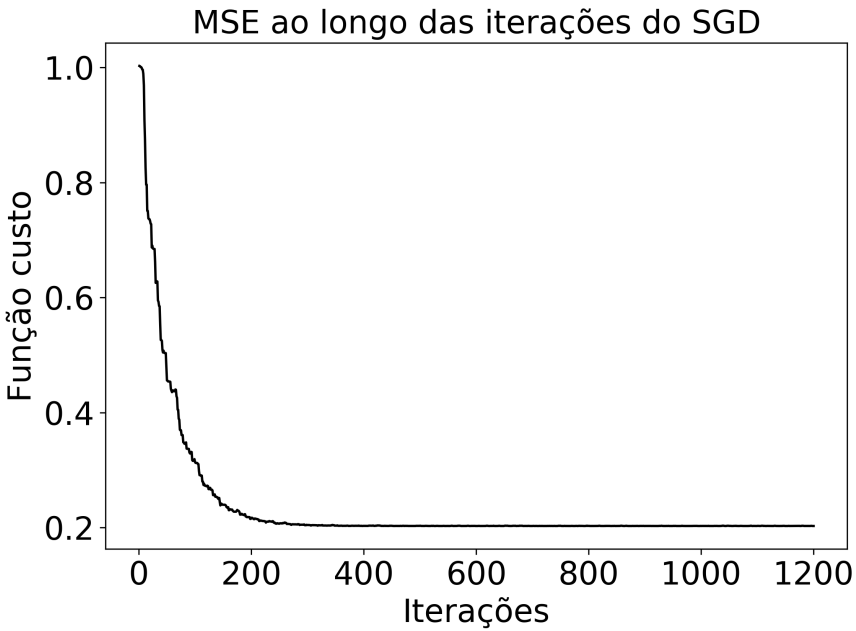
$$\hat{y}_i(t-1) = w_0(t-1) + w_1(t-1)x_i,$$
$$e_i(t-1) = y_i - \hat{y}_i(t-1).$$

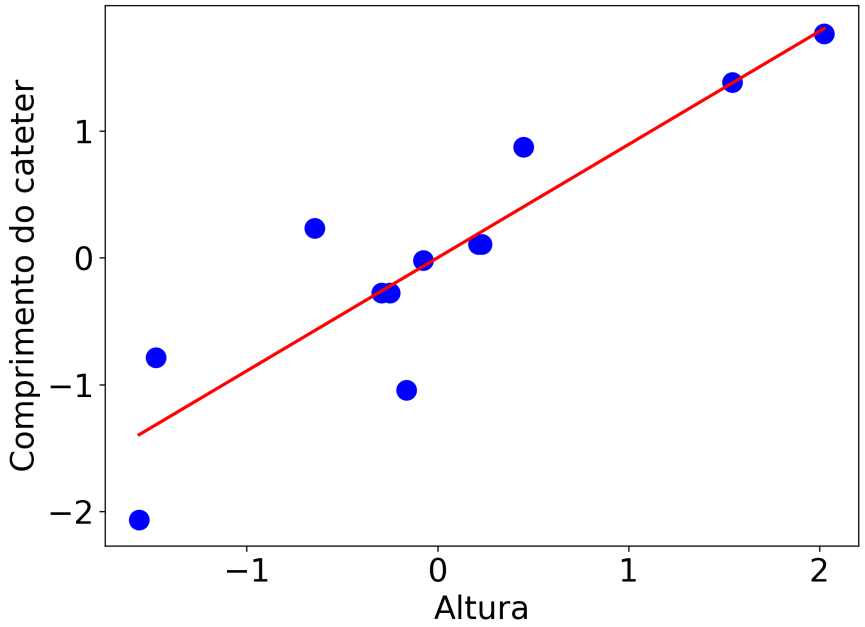
      3. Atualize os parâmetros:

$$w_0(t) = w_0(t-1) + \alpha e_i(t-1)$$
$$w_1(t) = w_1(t-1) + \alpha e_i(t-1)x_i$$

- Também chamado de **algoritmo LMS** (*Least Mean Squares*).

Regressão linear simples - Otimização via SGD





Regressão linear múltipla

- Podemos reconsiderar o problema inserindo o peso do paciente:

Altura (m)	**Peso (Kg)**	**Comprimento (cm)**
1.087	18.141	37
1.613	42.404	50
0.953	16.100	34
1.003	13.605	36
1.156	23.583	43
0.978	7.710	28
1.092	17.460	37
0.572	3.855	20
0.940	14.966	34
0.597	4.308	30
0.838	9.524	38
1.473	35.828	47

Regressão linear múltipla

- Podemos reconsiderar o problema inserindo o peso do paciente:

Altura (m)	**Peso (Kg)**	**Comprimento (cm)**
1.087	18.141	37
1.613	42.404	50
0.953	16.100	34
1.003	13.605	36
1.156	23.583	43
0.978	7.710	28
1.092	17.460	37
0.572	3.855	20
0.940	14.966	34
0.597	4.308	30
0.838	9.524	38
1.473	35.828	47

- Novo modelo linear **múltiplo**:

$$\hat{y}_i = w_0 + w_1 x_{i1} + w_2 x_{i2}$$

- x_{i1} é a i-ésima **Altura** e x_{i2} é o i-ésimo **Peso**.

Regressão linear múltipla

- Caso façamos $x_{i0} = 1$, temos:

$$\hat{y}_i = w_0 x_{i0} + w_1 x_{i1} + w_2 x_{i2}$$
$$\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i.$$

- Note que:

$$\boldsymbol{w} = [w_0, w_1, w_2]^\top,$$
$$\boldsymbol{x}_i = [1, x_{i1}, x_{i2}]^\top.$$

Regressão linear múltipla

Gradiente Descendente

- Regra de atualização:

$$\boldsymbol{w}(t) = \boldsymbol{w}(t-1) + \alpha \frac{1}{N} \sum_{i=1}^{N} e_i(t-1)\boldsymbol{x}_i$$

Gradiente Descendente Estocástico

- Regra de atualização:

$$\boldsymbol{w}(t) = \boldsymbol{w}(t-1) + \alpha e_i(t-1)\boldsymbol{x}_i$$

Diferença probabilística entre o GD e o SGD

- O erro quadrático $\mathcal{J}$ é uma **variável aleatória**:

$$\min_{\boldsymbol{w}} \frac{1}{2}\mathbb{E}\{\mathcal{J}(\boldsymbol{w})\}, \quad \mathcal{J}(\boldsymbol{w}) = e^2 = (y - \hat{y})^2,$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\alpha}{2}\frac{\partial}{\partial \boldsymbol{w}}\mathbb{E}\{\mathcal{J}\}$$

Diferença probabilística entre o GD e o SGD

- O erro quadrático $\mathcal{J}$ é uma **variável aleatória**:

$$\min_{\boldsymbol{w}} \frac{1}{2}\mathbb{E}\{\mathcal{J}(\boldsymbol{w})\}, \quad \mathcal{J}(\boldsymbol{w}) = e^2 = (y - \hat{y})^2,$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\alpha}{2}\frac{\partial}{\partial \boldsymbol{w}}\mathbb{E}\{\mathcal{J}\}$$

- A **média amostral** resulta no algoritmo GD:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\alpha}{2}\frac{\partial}{\partial \boldsymbol{w}}\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right)$$

Diferença probabilística entre o GD e o SGD

- O erro quadrático $\mathcal{J}$ é uma **variável aleatória**:

$$\min_{\boldsymbol{w}} \frac{1}{2}\mathbb{E}\{\mathcal{J}(\boldsymbol{w})\}, \quad \mathcal{J}(\boldsymbol{w}) = e^2 = (y - \hat{y})^2,$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\alpha}{2}\frac{\partial}{\partial \boldsymbol{w}}\mathbb{E}\{\mathcal{J}\}$$

- A **média amostral** resulta no algoritmo GD:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\alpha}{2}\frac{\partial}{\partial \boldsymbol{w}}\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right)$$

- Uma **aproximação estocástica** resulta no algoritmo SGD:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\alpha}{2}\frac{\partial}{\partial \boldsymbol{w}}(y_i - \hat{y}_i)^2$$

Agenda

Regressão linear analítica

- Reunimos todos os padrões de entrada $\boldsymbol{x}_i$ em uma matriz $\boldsymbol{X}$:

$$\boldsymbol{X} = \left[\boldsymbol{x}_1 \ \boldsymbol{x}_2 \ \cdots \ \boldsymbol{x}_N\right]^\top \in \mathbb{R}^{N \times (D+1)}.$$

  - N é o número de observações/amostras/vetores/padrões.
  - D é a dimensão da entrada (excluindo o termo $x_{i0} = 1$).

- Agrupamos as saídas disponíveis em um vetor $\boldsymbol{y}$:

$$\boldsymbol{y} = \left[y_1, y_2, \cdots, y_N\right]^\top \in \mathbb{R}^N.$$

- Agrupamos as saídas do modelo em um vetor $\hat{\boldsymbol{y}}$:

$$\hat{\boldsymbol{y}} = \left[\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_N\right]^\top \in \mathbb{R}^N.$$

- Agrupamos os parâmetros em um vetor $\boldsymbol{w}$:

$$\boldsymbol{w} = \left[w_0, w_1, \cdots, w_D\right]^\top \in \mathbb{R}^{D+1}.$$

Regressão linear analítica

- Dados do problema do cateterismo cardíaco em formato matricial:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1.087 & 18.141 \\ 1 & 1.613 & 42.404 \\ 1 & 0.953 & 16.100 \\ 1 & 1.003 & 13.605 \\ 1 & 1.156 & 23.583 \\ 1 & 0.978 & 7.710 \\ 1 & 1.092 & 17.460 \\ 1 & 0.572 & 3.855 \\ 1 & 0.940 & 14.966 \\ 1 & 0.597 & 4.308 \\ 1 & 0.838 & 9.524 \\ 1 & 1.473 & 35.828 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 37 \\ 50 \\ 34 \\ 36 \\ 43 \\ 28 \\ 37 \\ 20 \\ 34 \\ 30 \\ 38 \\ 47 \end{bmatrix}$$

Regressão linear analítica

- Reformulamos nosso modelo linear na forma matricial:

$$\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i,$$
$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w}.$$

Regressão linear analítica

- Reformulamos nosso modelo linear na forma matricial:

$$\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i,$$
$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w}.$$

- Reformulamos também a função custo:

$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}})$$
$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$

Regressão linear analítica

- O mínimo de $\mathcal{J}(\boldsymbol{w})$ ocorrerá em $\frac{\partial \mathcal{J}(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0$:

$$\begin{aligned}
\frac{\partial \mathcal{J}(\boldsymbol{w})}{\partial \boldsymbol{w}} &= \frac{1}{2} 2(-\boldsymbol{X}^\top)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \\
-\boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} &= 0 \\
\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} &= \boldsymbol{X}^\top \boldsymbol{y} \\
\boldsymbol{w} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.
\end{aligned}$$

Regressão linear analítica

- O mínimo de $\mathcal{J}(\boldsymbol{w})$ ocorrerá em $\frac{\partial \mathcal{J}(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0$:

$$\begin{aligned}
\frac{\partial \mathcal{J}(\boldsymbol{w})}{\partial \boldsymbol{w}} &= \frac{1}{2} 2(-\boldsymbol{X}^\top)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \\
-\boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} &= 0 \\
\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} &= \boldsymbol{X}^\top \boldsymbol{y} \\
\boldsymbol{w} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.
\end{aligned}$$

- $\boldsymbol{X}^+ = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$, em que $\boldsymbol{X}^+ \boldsymbol{X} = \boldsymbol{I}$, é chamada de **inversa de Moore-Penrose** ou **pseudo-inversa**.

Regressão linear analítica

Método dos mínimos quadrados ordinários (OLS, *ordinary least squares*)

- O vetor de parâmetros $\boldsymbol{w}$ que minimiza $\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}})$ é dado por

$$\boldsymbol{w} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y},$$

em quem $\boldsymbol{X}$ é a matriz de vetores de entrada (um por linha) e $\boldsymbol{y}$ é o vetor de saídas desejadas.

Regressão linear analítica

- OLS equivale ao método de Newton aplicado na função de custo quadrática $\mathcal{J}$:

$$\boldsymbol{w} = \boldsymbol{w}_0 - \left(\frac{\partial^2 \mathcal{J}(\boldsymbol{w}_0)}{\partial \boldsymbol{w}_0^2}\right)^{-1} \frac{\partial \mathcal{J}(\boldsymbol{w}_0)}{\partial \boldsymbol{w}_0},$$

$$\frac{\partial \mathcal{J}(\boldsymbol{w}_0)}{\partial \boldsymbol{w}_0} = -\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_0),$$

$$\frac{\partial^2 \mathcal{J}(\boldsymbol{w}_0)}{\partial \boldsymbol{w}_0^2} = \boldsymbol{X}^\top \boldsymbol{X},$$

$$\boldsymbol{w} = \boldsymbol{w}_0 - (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \left(-\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_0)\right)$$

- Podemos escolher $\boldsymbol{w}_0 = \boldsymbol{0}$ para obter o mínimo global.

$$\boldsymbol{w} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Regressão linear analítica

- De onde vem a função custo $\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}})$?

Regressão linear analítica

- De onde vem a função custo $\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}})$?
- Considerando um ruído independente $\epsilon \sim \mathcal{N}(0, \sigma^2)$:

$$y_i = \hat{y}_i + \epsilon = \boldsymbol{w}^\top \boldsymbol{x}_i + \epsilon,$$

$$p(y_i|\boldsymbol{x}_i, \boldsymbol{w}) = \mathcal{N}(y_i|\boldsymbol{w}^\top \boldsymbol{x}_i, \sigma^2),$$

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = \prod_{i=1}^{N} \mathcal{N}(y_i|\boldsymbol{w}^\top \boldsymbol{x}_i, \sigma^2)$$

$$\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2}{2\sigma^2}\right)$$

$$\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = \underbrace{-\frac{N}{2}\log(2\pi\sigma^2)}_{\text{const. em relação a } \boldsymbol{w}} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$$

Regressão linear analítica

- Queremos maximizar $\mathcal{L}(\boldsymbol{w}) = \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$, o que equivale a minimizar $\mathcal{J}(\boldsymbol{w}) = -\mathcal{L}(\boldsymbol{w})$.
- Ignorando os termos constantes:

$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$$

$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$

$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}})$$

Regressão linear analítica

- Queremos maximizar $\mathcal{L}(\boldsymbol{w}) = \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$, o que equivale a minimizar $\mathcal{J}(\boldsymbol{w}) = -\mathcal{L}(\boldsymbol{w})$.
- Ignorando os termos constantes:

$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$$

$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$

$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}})$$

Solução de máxima verossimilhança

A solução obtida via OLS (e aproximada via GD e SGD), chamada de solução de **máxima verossimilhança (maximum likelihood)**, é ótima quando o ruído é Gaussiano:

$$\boldsymbol{w}_{\text{OLS}} = \boldsymbol{w}_{\text{ML}} = \arg\max \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$$

Agenda

Propriedades da Distribuição Gaussiana

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Marginalização

A observação de uma coleção maior de variáveis não afeta a distribuição de subconjuntos menores, ou seja:

$$\boldsymbol{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \text{ e } \boldsymbol{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

Condicionamento

Condicionar Gaussianas resulta em uma Gaussiana:

$$p(\boldsymbol{x}_1 | \boldsymbol{x}_2 = \boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}_1 | \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$$

Linearidade

Uma combinação linear de Gaussianas resulta em uma Gaussiana:

$$p(a\boldsymbol{x}_1 + b\boldsymbol{x}_2) = \mathcal{N}(a\boldsymbol{\mu}_1 + b\boldsymbol{\mu}_2, a^2\boldsymbol{\Sigma}_{11} + b^2\boldsymbol{\Sigma}_{22})$$

Regressão Linear Bayesiana

- Considere um modelo linear com ruído Gaussiano:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}, \quad \epsilon_i \sim \mathcal{N}(\epsilon_i|0, \sigma^2).$$

Regressão Linear Bayesiana

- Considere um modelo linear com ruído Gaussiano:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}, \quad \epsilon_i \sim \mathcal{N}(\epsilon_i|0, \sigma^2).$$

- A verossimilhança do modelo é dada por:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, \sigma^2\boldsymbol{I}).$$

Regressão Linear Bayesiana

- Considere um modelo linear com ruído Gaussiano:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}, \quad \epsilon_i \sim \mathcal{N}(\epsilon_i|0, \sigma^2).$$

- A verossimilhança do modelo é dada por:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, \sigma^2\boldsymbol{I}).$$

- Por ser desconhecido, o vetor de parâmetros $\boldsymbol{w}$ recebe uma distribuição a priori, por exemplo, Gaussiana:

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0).$$

Regressão Linear Bayesiana

- Considere um modelo linear com ruído Gaussiano:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}, \quad \epsilon_i \sim \mathcal{N}(\epsilon_i|0, \sigma^2).$$

- A verossimilhança do modelo é dada por:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, \sigma^2\boldsymbol{I}).$$

- Por ser desconhecido, o vetor de parâmetros $\boldsymbol{w}$ recebe uma distribuição a priori, por exemplo, Gaussiana:

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0).$$

- Após a observação dos dados $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$, como calcular a distribuição a posteriori $p(\boldsymbol{w}|\mathcal{D})$?

Regressão Linear Bayesiana

- Voltamos às propriedades da Gaussiana (com nova notação):

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{w} \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}(\boldsymbol{m}_z, \boldsymbol{S}_z), \quad \boldsymbol{m}_z = \begin{bmatrix} \boldsymbol{\mu}_w \\ \boldsymbol{\mu}_y \end{bmatrix}, \quad \boldsymbol{S}_z = \begin{bmatrix} \boldsymbol{\Sigma}_w & \boldsymbol{\Sigma}_{wy} \\ \boldsymbol{\Sigma}_{yw} & \boldsymbol{\Sigma}_y \end{bmatrix}.$$

Marginalização

A observação de uma coleção maior de variáveis não afeta a distribuição de subconjuntos menores, ou seja:

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w) \text{ e } \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$$

Condicionamento

Condicionar Gaussianas resulta em uma Gaussiana:

$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_w + \boldsymbol{\Sigma}_{wy}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \boldsymbol{\Sigma}_w - \boldsymbol{\Sigma}_{wy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yw})$$

Regressão Linear Bayesiana

- Usamos a priori de $\boldsymbol{w}$, $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)$:

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{w} \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}(\boldsymbol{m}_z, \boldsymbol{S}_z), \quad \boldsymbol{m}_z = \begin{bmatrix} \boldsymbol{m}_0 \\ \boldsymbol{\mu}_y \end{bmatrix}, \quad \boldsymbol{S}_z = \begin{bmatrix} \boldsymbol{S}_0 & \boldsymbol{\Sigma}_{wy} \\ \boldsymbol{\Sigma}_{yw} & \boldsymbol{\Sigma}_y \end{bmatrix}.$$

Marginalização

A observação de uma coleção maior de variáveis não afeta a distribuição de subconjuntos menores, ou seja:

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{S}_0) \text{ e } \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$$

Condicionamento

Condicionar Gaussianas resulta em uma Gaussiana:

$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0 + \boldsymbol{\Sigma}_{wy}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \boldsymbol{S}_0 - \boldsymbol{\Sigma}_{wy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yw})$$

Regressão Linear Bayesiana

- Note que pela propriedade do condicionamento, podemos obter a posteriori de $\boldsymbol{w}$:

$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0 + \boldsymbol{\Sigma}_{wy}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \boldsymbol{S}_0 - \boldsymbol{\Sigma}_{wy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yw}).$$

Regressão Linear Bayesiana

- Note que pela propriedade do condicionamento, podemos obter a posteriori de $\boldsymbol{w}$:

$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0 + \boldsymbol{\Sigma}_{wy}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \boldsymbol{S}_0 - \boldsymbol{\Sigma}_{wy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yw}).$$

- Para calcular a expressão acima, precisamos de alguns valores ainda desconhecidos:
  - $\rightarrow$ $\boldsymbol{\mu}_y$ e $\boldsymbol{\Sigma}_y$, provenientes da verossimilhança marginal $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$.
  - $\rightarrow$ $\boldsymbol{\Sigma}_{wy}$, a matriz de covariância cruzada entre $\boldsymbol{w}$ e $\boldsymbol{y}$.

Regressão Linear Bayesiana

- Começamos calculando a verossimilhança marginal de $\boldsymbol{y}$:

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{m}_0, \boldsymbol{S}_0, \sigma^2) &= \int p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2) p(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} \\
&= \int \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I}) \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0) \mathrm{d}\boldsymbol{w} \\
&= \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{m}_0, \boldsymbol{X}\boldsymbol{S}_0\boldsymbol{X}^\top + \sigma^2 \boldsymbol{I}).
\end{aligned}
$$

Regressão Linear Bayesiana

- Começamos calculando a verossimilhança marginal de $\boldsymbol{y}$:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{m}_0, \boldsymbol{S}_0, \sigma^2) &= \int p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2) p(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} \\ &= \int \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I}) \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0) \mathrm{d}\boldsymbol{w} \\ &= \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{m}_0, \boldsymbol{X}\boldsymbol{S}_0\boldsymbol{X}^\top + \sigma^2 \boldsymbol{I}). \end{aligned}$$

- Portanto, se $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$:

$$\begin{aligned} \boldsymbol{\mu}_y &= \boldsymbol{X}\boldsymbol{m}_0, \\ \boldsymbol{\Sigma}_y &= \boldsymbol{X}\boldsymbol{S}_0\boldsymbol{X}^\top + \sigma^2 \boldsymbol{I}. \end{aligned}$$

Regressão Linear Bayesiana

- Começamos calculando a verossimilhança marginal de $\boldsymbol{y}$:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{m}_0, \boldsymbol{S}_0, \sigma^2) &= \int p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \sigma^2) p(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} \\ &= \int \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I}) \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0) \mathrm{d}\boldsymbol{w} \\ &= \mathcal{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{m}_0, \boldsymbol{X}\boldsymbol{S}_0\boldsymbol{X}^\top + \sigma^2 \boldsymbol{I}). \end{aligned}$$

- Portanto, se $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$:

$$\begin{aligned} \boldsymbol{\mu}_y &= \boldsymbol{X}\boldsymbol{m}_0, \\ \boldsymbol{\Sigma}_y &= \boldsymbol{X}\boldsymbol{S}_0\boldsymbol{X}^\top + \sigma^2 \boldsymbol{I}. \end{aligned}$$

- Qual a covariância entre $\boldsymbol{w}$ e $\boldsymbol{y}$, isto é, $\boldsymbol{\Sigma}_{wy}$?

$$\begin{aligned} \boldsymbol{\Sigma}_{wy} = \mathrm{cov}[\boldsymbol{w}, \boldsymbol{y}] &= \mathrm{cov}[\boldsymbol{w}, \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}] \\ &= \mathrm{cov}[\boldsymbol{w}, \boldsymbol{X}\boldsymbol{w}] + \mathrm{cov}[\boldsymbol{w}, \boldsymbol{\epsilon}] \\ &= \mathrm{cov}[\boldsymbol{w}, \boldsymbol{X}\boldsymbol{w}] = \mathrm{cov}[\boldsymbol{w}, \boldsymbol{w}]\boldsymbol{X}^\top = \boldsymbol{S}_0\boldsymbol{X}^\top. \end{aligned}$$

Regressão Linear Bayesiana

- Agora podemos construir a distribuição conjunta $p(\boldsymbol{w}, \boldsymbol{y})$:

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{w} \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}(\boldsymbol{m}_z, \boldsymbol{S}_z), \quad \boldsymbol{m}_z = \begin{bmatrix} \boldsymbol{m}_0 \\ \boldsymbol{\mu}_y \end{bmatrix}, \quad \boldsymbol{S}_z = \begin{bmatrix} \boldsymbol{S}_0 & \boldsymbol{\Sigma}_{wy} \\ \boldsymbol{\Sigma}_{yw} & \boldsymbol{\Sigma}_y \end{bmatrix},$$

$$\begin{bmatrix} \boldsymbol{w} \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m}_0 \\ \boldsymbol{X}\boldsymbol{m}_0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{S}_0 & \boldsymbol{S}_0\boldsymbol{X}^\top \\ \boldsymbol{X}\boldsymbol{S}_0 & \boldsymbol{X}\boldsymbol{S}_0\boldsymbol{X}^\top + \sigma^2\boldsymbol{I} \end{bmatrix} \right).$$

Regressão Linear Bayesiana

- Finalmente, usamos a propriedade de condicionamento de Gaussianas para obter a posteriori de $\boldsymbol{w}$:

$$p(\boldsymbol{w}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0 + \boldsymbol{\Sigma}_{wy}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \boldsymbol{S}_0 - \boldsymbol{\Sigma}_{wy}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_{yw}),$$

$$p(\boldsymbol{w}|\mathcal{D}, \boldsymbol{m}_0, \boldsymbol{S}_0, \sigma^2) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

$$\begin{aligned}
\boldsymbol{\mu} &= \boldsymbol{m}_0 + \boldsymbol{S}_0\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{S}_0\boldsymbol{X}^\top + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{m}_0) \\
&= \boldsymbol{m}_0 + (\boldsymbol{S}_0\boldsymbol{X}^\top\boldsymbol{X} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{S}_0\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{m}_0), \\
\boldsymbol{\Sigma} &= \boldsymbol{S}_0 - \boldsymbol{S}_0\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{S}_0\boldsymbol{X}^\top + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{X}\boldsymbol{S}_0 \\
&= \boldsymbol{S}_0 - (\boldsymbol{S}_0\boldsymbol{X}^\top\boldsymbol{X} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{S}_0\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{S}_0.
\end{aligned}$$

  em que foi usada a identidade matricial abaixo (derivada a partir da identidade de Woodbury):

$$\boldsymbol{U}(\boldsymbol{V}\boldsymbol{U} + \boldsymbol{I})^{-1} = (\boldsymbol{U}\boldsymbol{V} + \boldsymbol{I})^{-1}\boldsymbol{U}$$

- Note que o modelo não é definido por um único valor para $\boldsymbol{w}$, mas uma distribuição a posteriori $p(\boldsymbol{w}|\mathcal{D}, \boldsymbol{m}_0, \boldsymbol{S}_0, \sigma^2)$.

Regressão Linear Bayesiana

- Com a distribuição a posteriori $p(\boldsymbol{w}|\mathcal{D}, \boldsymbol{m}_0, \boldsymbol{S}_0, \sigma^2)$, podemos fazer predições para novas entradas $\boldsymbol{X}_*$:

$$\begin{aligned}
\boldsymbol{y}_* &= \boldsymbol{X}_* \boldsymbol{w} + \boldsymbol{\epsilon}, \\
p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{D}, \boldsymbol{m}_0, \boldsymbol{S}_0, \sigma^2) &= \int p(\boldsymbol{y}_*|\boldsymbol{w}, \sigma^2) p(\boldsymbol{w}|\mathcal{D}, \boldsymbol{m}_0, \boldsymbol{S}_0, \sigma^2) \mathrm{d}\boldsymbol{w} \\
&= \int \mathcal{N}(\boldsymbol{y}_*|\boldsymbol{X}_* \boldsymbol{w}, \sigma^2 \boldsymbol{I}) \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{w} \\
&= \mathcal{N}(\boldsymbol{y}_*|\boldsymbol{X}_* \boldsymbol{\mu}, \boldsymbol{X}_* \boldsymbol{\Sigma} \boldsymbol{X}_*^\top + \sigma^2 \boldsymbol{I}).
\end{aligned}$$

- Note que a predição não é um único valor, mas uma distribuição de probabilidade bem definida.

Regressão Linear Bayesiana

- Caso escolhamos uma priori $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \sigma_w^2 \boldsymbol{I})$, temos:

$$
\begin{aligned}
p(\boldsymbol{w}|\mathcal{D}, \sigma_w^2, \sigma^2) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\
\boldsymbol{\mu} &= \sigma_w^2 (\sigma_w^2 \boldsymbol{X}^\top \boldsymbol{X} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \\
\boldsymbol{\Sigma} &= \sigma_w^2 \boldsymbol{I} - \sigma_w^2 (\sigma_w^2 \boldsymbol{X}^\top \boldsymbol{X} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{X} \sigma_w^2.
\end{aligned}
$$

- Podemos reescrever a média a posteriori $\boldsymbol{\mu}$ para obter a solução de mínimos quadrados regularizada (ridge regression):

$$
\begin{aligned}
\boldsymbol{\mu} &= \left(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\sigma^2}{\sigma_w^2} \boldsymbol{I} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\
&= \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \quad \lambda = \frac{\sigma^2}{\sigma_w^2}.
\end{aligned}
$$

Regressão Linear Bayesiana

Resumo do algoritmo

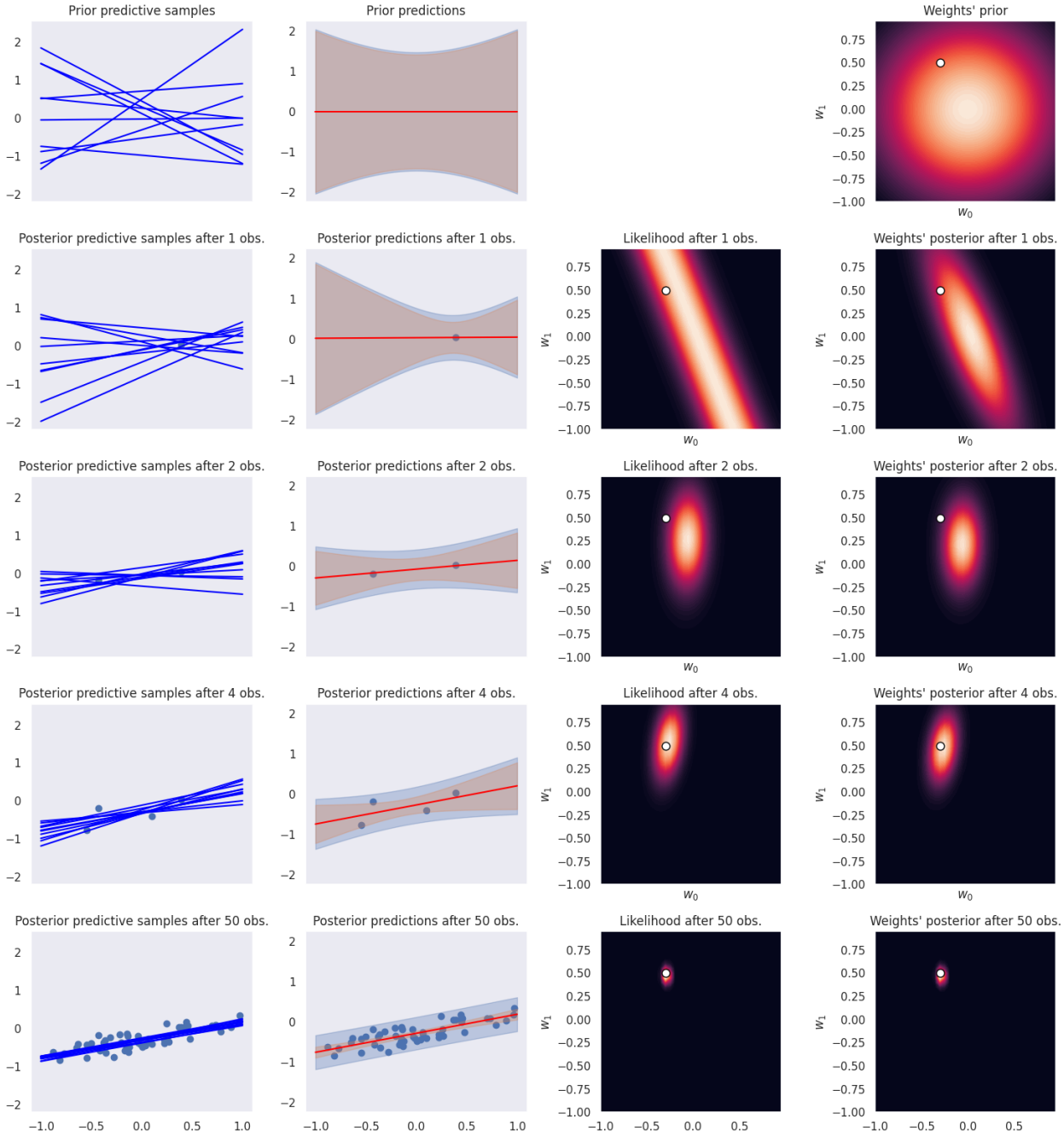
- Passo de estimação
  1. Defina a partir de conhecimentos/experimentos anteriores:
     - $\rightarrow$ os momentos da priori $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)$;
     - $\rightarrow$ a variância do ruído $p(\epsilon) = \mathcal{N}(\epsilon|0, \sigma^2)$.
  2. A partir dos dados $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$, calcule a posteriori de $\boldsymbol{w}$:

$$
\begin{aligned}
p(\boldsymbol{w}|\mathcal{D}) &= \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\
\boldsymbol{\mu} &= \boldsymbol{m}_0 + (\boldsymbol{S}_0\boldsymbol{X}^\top\boldsymbol{X} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{S}_0\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{m}_0), \\
\boldsymbol{\Sigma} &= \boldsymbol{S}_0 - (\boldsymbol{S}_0\boldsymbol{X}^\top\boldsymbol{X} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{S}_0\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{S}_0.
\end{aligned}
$$

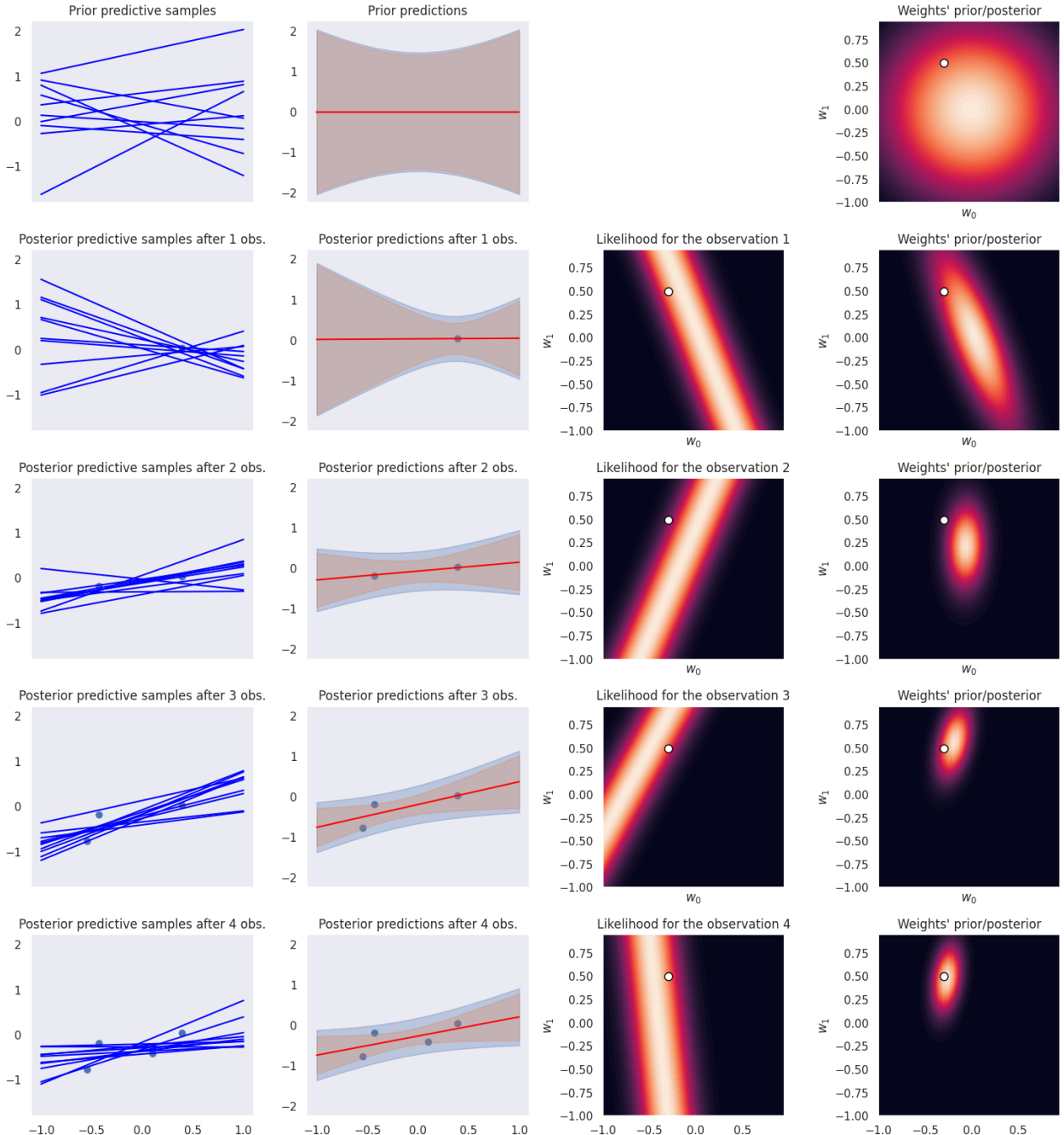
  3. Retorne a posteriori $p(\boldsymbol{w}|\mathcal{D})$ dos parâmetros.
- Passo de predição
  1. Dados padrões $\boldsymbol{X}_* \in \mathbb{R}^{(N_* \times D)}$, retorne a distribuição preditiva

$$
p(\boldsymbol{y}_*|\boldsymbol{X}_*) = \mathcal{N}(\boldsymbol{y}_*|\boldsymbol{X}_*\boldsymbol{\mu}, \boldsymbol{X}_*\boldsymbol{\Sigma}\boldsymbol{X}_*^\top + \sigma^2\boldsymbol{I}).
$$

Regressão Linear Bayesiana em batch



Regressão Linear Bayesiana sequencial



Regressão Linear Bayesiana com Funções de Base

- Podemos usar uma transformação arbitrária $\phi(\cdot)$ nos atributos:

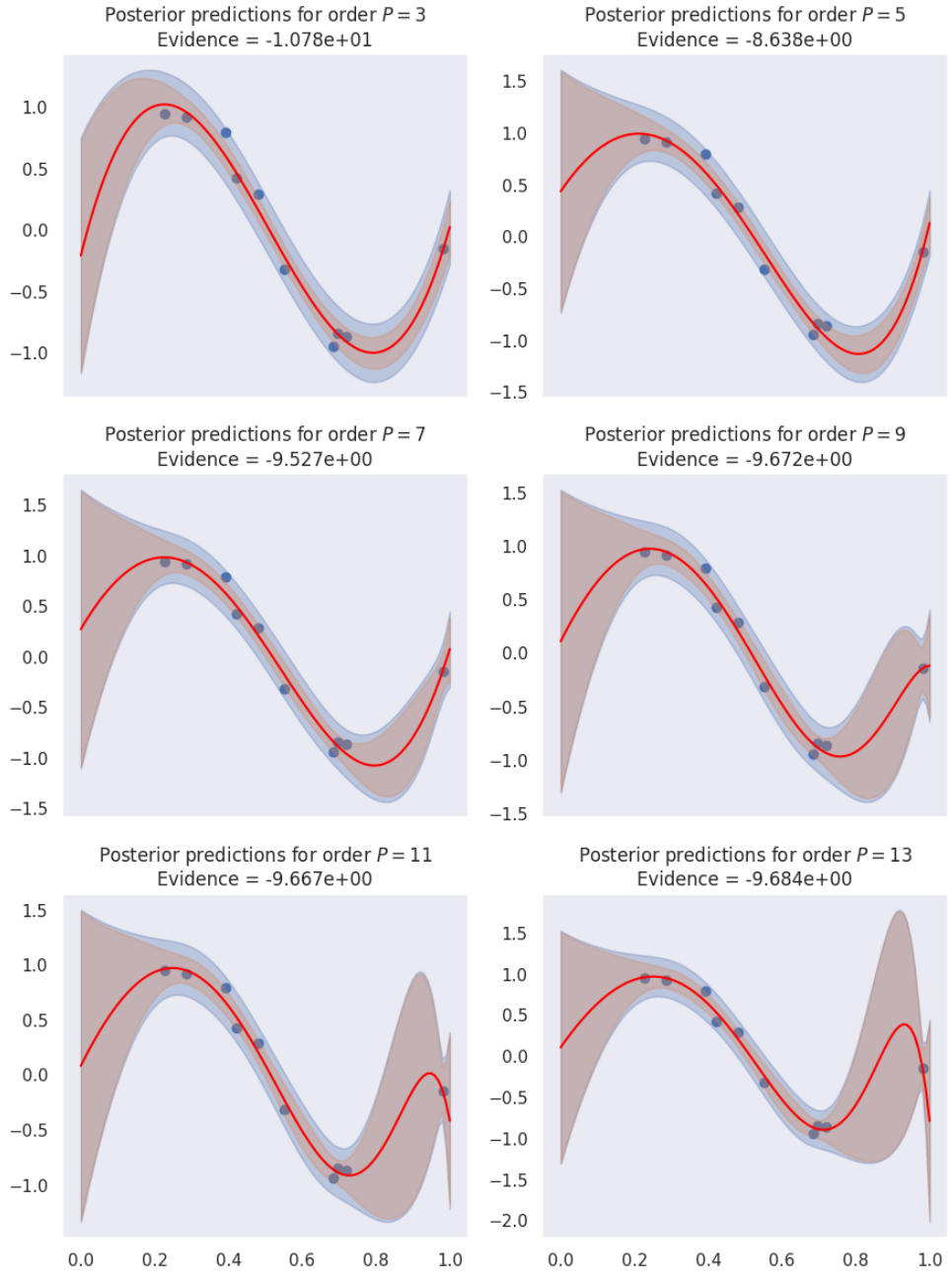
$$\boldsymbol{\Phi} = \phi(\boldsymbol{X}) = [\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2), \ldots, \phi(\boldsymbol{x}_N)]^\top, \quad \phi : \mathbb{R}^D \to \mathbb{R}^Q,$$
$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{\epsilon}, \quad \epsilon \sim \mathcal{N}(\epsilon|0, \sigma^2).$$

- As expressões do modelo linear Bayesiano continuam as mesmas:

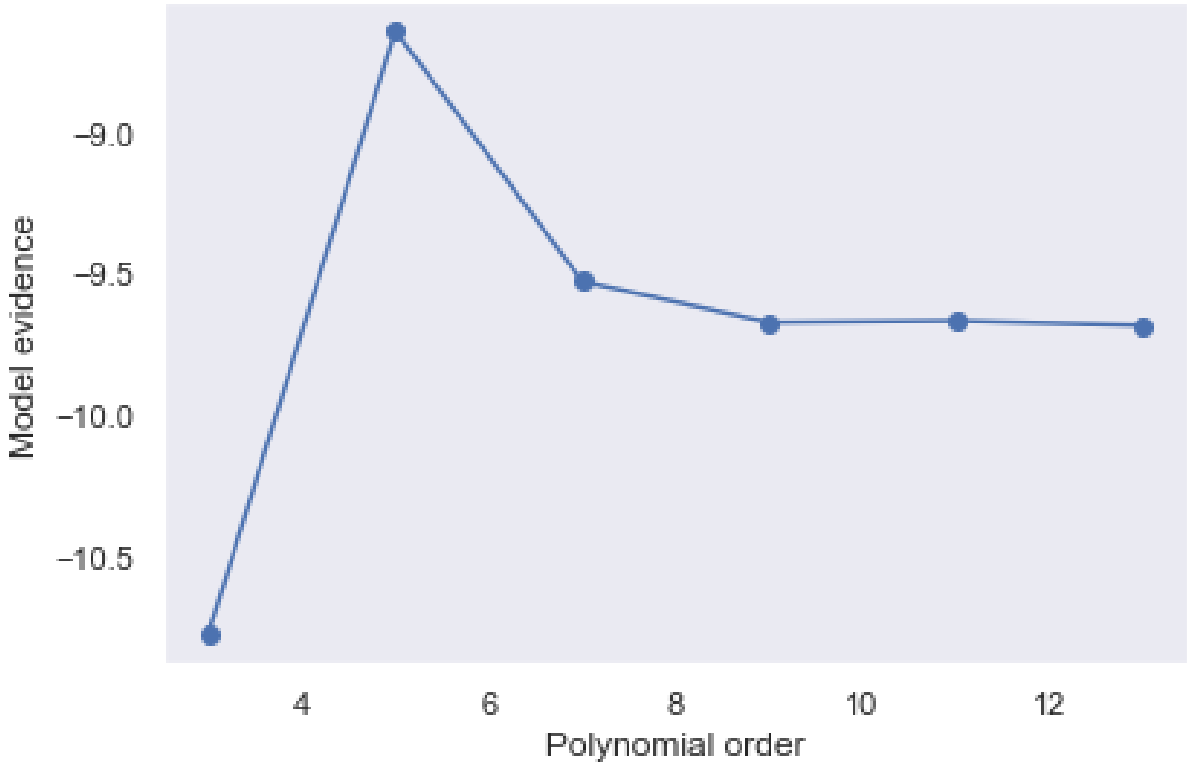
$$p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{w}, \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}),$$
$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0),$$
$$p(\boldsymbol{w}|\mathcal{D}, \boldsymbol{m}_0, \boldsymbol{S}_0, \sigma^2) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \mathcal{D} = (\boldsymbol{\Phi}, \boldsymbol{y}),$$
$$p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{D}, \boldsymbol{m}_0, \boldsymbol{S}_0, \sigma^2) = \mathcal{N}(\boldsymbol{y}_*|\boldsymbol{\Phi}_*\boldsymbol{\mu}, \boldsymbol{\Phi}_*\boldsymbol{\Sigma}\boldsymbol{\Phi}_*^\top + \sigma^2\boldsymbol{I}),$$
$$\boldsymbol{\mu} = \boldsymbol{m}_0 + (\boldsymbol{S}_0\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{S}_0\boldsymbol{\Phi}^\top(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{m}_0),$$
$$\boldsymbol{\Sigma} = \boldsymbol{S}_0 - (\boldsymbol{S}_0\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{S}_0\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{S}_0.$$

- Regressão polinomial Bayesiana de ordem P pode ser obtida fazendo $\phi(x_i) = [1, x_i, x_i^2, \ldots, x_i^P]^\top$.

Regressão Polinomial Bayesiana variando a ordem



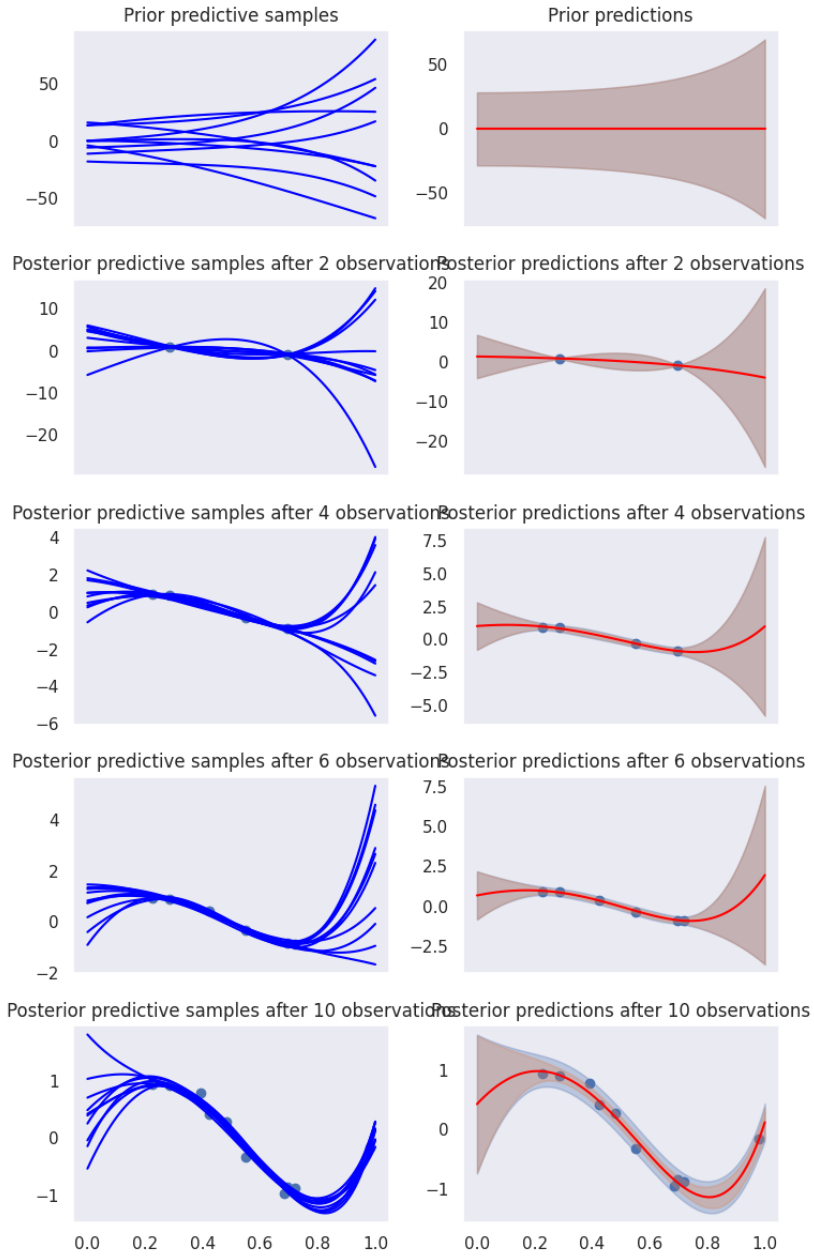
Regressão Polinomial Bayesiana variando a ordem



Lembrando que a evidência do modelo é dada por:

$$\begin{aligned}\log p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{m}_0, \boldsymbol{S}_0, \sigma^2) &= \log \int p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{w}, \sigma^2) p(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} \\ &= \log \int \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2 \boldsymbol{I}) \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0) \mathrm{d}\boldsymbol{w} \\ &= \log \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{m}_0, \boldsymbol{\Phi}\boldsymbol{S}_0\boldsymbol{\Phi}^\top + \sigma^2 \boldsymbol{I}).\end{aligned}$$

Regressão Polinomial Bayesiana em batch



Regressão Linear Bayesiana com Funções de Base

- Podemos escolher $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \boldsymbol{S}_0)$:

$$p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{D}, \boldsymbol{S}_0, \sigma^2) = \mathcal{N}(\boldsymbol{y}_*|\boldsymbol{\Phi}_*\boldsymbol{\mu}, \boldsymbol{\Phi}_*\boldsymbol{\Sigma}\boldsymbol{\Phi}_*^\top + \sigma^2\boldsymbol{I}),$$

$$\boldsymbol{\mu} = \underbrace{(\boldsymbol{S}_0\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{S}_0\boldsymbol{\Phi}^\top}_{\boldsymbol{A}}\boldsymbol{y},$$

$$\boldsymbol{\Sigma} = \boldsymbol{S}_0 - \underbrace{(\boldsymbol{S}_0\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{S}_0\boldsymbol{\Phi}^\top}_{\boldsymbol{A}}\boldsymbol{\Phi}\boldsymbol{S}_0.$$

- Usamos a seguinte identidade matricial (derivada da identidade de Woodbury):

$$\boldsymbol{A} = (\boldsymbol{S}_0\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{S}_0\boldsymbol{\Phi}^\top = \boldsymbol{S}_0\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{S}_0\boldsymbol{\Phi}^\top + \sigma^2\boldsymbol{I})^{-1}$$

- Reescrevemos a distribuição preditiva:

$$p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{D}, \boldsymbol{S}_0, \sigma^2) = \mathcal{N}(\boldsymbol{y}_*|\boldsymbol{\Phi}_*\boldsymbol{S}_0\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{S}_0\boldsymbol{\Phi}^\top + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y},$$
$$\boldsymbol{\Phi}_*(\boldsymbol{S}_0 - \boldsymbol{S}_0\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{S}_0\boldsymbol{\Phi}^\top + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{\Phi}\boldsymbol{S}_0)\boldsymbol{\Phi}_*^\top + \sigma^2\boldsymbol{I}),$$

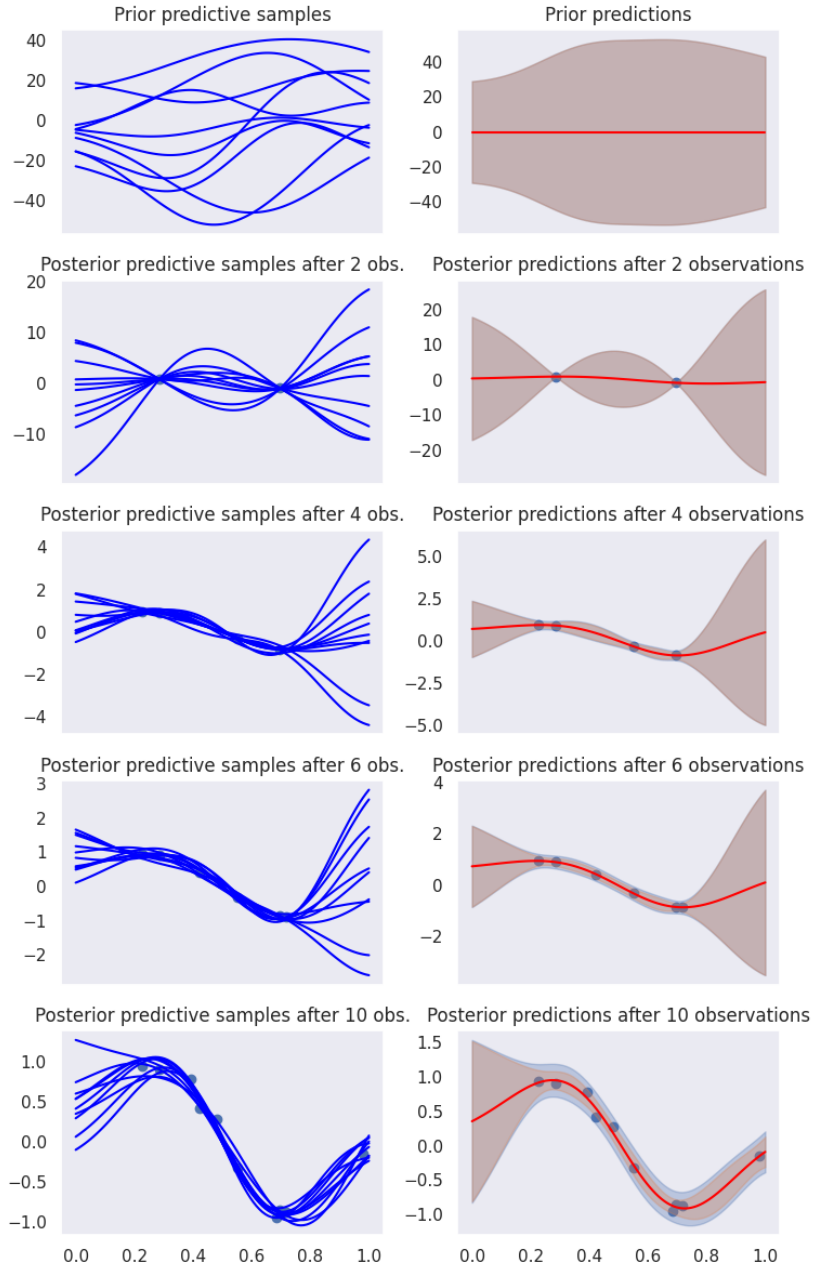
Regressão RBF Bayesiana

- Podemos usar uma função de base radial (radial basis function) para obter uma regressão RBF Bayesiana.
- Considerando M funções RBF com centros $c_m|_{m=1}^{M}$, largura de banda $\lambda > 0$ e entradas unidimensionais:

$$
\begin{aligned}
y_i &= \boldsymbol{w}^\top \boldsymbol{\phi}_i + \epsilon, \\
\boldsymbol{\phi}_i &= [1, \phi_1(x_i), \ldots, \phi_M(x_i)]^\top, \\
\phi_m(x_i) &= \exp\left(-\frac{(x_i - c_m)^2}{2\lambda}\right).
\end{aligned}
$$

- Por exemplo, podemos inicializar os centros via algoritmo de agrupamento e fazer $\lambda = \mathbb{V}(x)$.

Regressão RBF Bayesiana



Do espaço de atributos para funções de kernel

- Fazendo $\boldsymbol{\Psi} = \boldsymbol{\Phi}\boldsymbol{S}_0^{1/2}$ e $\boldsymbol{\Psi}_* = \boldsymbol{\Phi}_*\boldsymbol{S}_0^{1/2}$ a preditiva se torna:

$$p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{D}, \boldsymbol{S}_0, \sigma^2) = \mathcal{N}(\boldsymbol{y}_*|\boldsymbol{\Psi}_*\boldsymbol{\Psi}^\top(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y}, \\ (\boldsymbol{\Psi}_*\boldsymbol{\Psi}_*^\top - \boldsymbol{\Psi}_*\boldsymbol{\Psi}^\top(\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{\Psi}\boldsymbol{\Psi}_*^\top) + \sigma^2\boldsymbol{I}),$$

- Agora podemos aplicar o *kernel trick* (truque do kernel):

$$\begin{aligned}
\boldsymbol{\Psi}\boldsymbol{\Psi}^\top &= k(\boldsymbol{X}, \boldsymbol{X}) = \boldsymbol{K}, \\
\boldsymbol{\Psi}_*\boldsymbol{\Psi}^\top &= k(\boldsymbol{X}_*, \boldsymbol{X}) = \boldsymbol{K}_{*f}, \\
\boldsymbol{\Psi}\boldsymbol{\Psi}_*^\top &= k(\boldsymbol{X}, \boldsymbol{X}_*) = \boldsymbol{K}_{f*}, \\
\boldsymbol{\Psi}_*\boldsymbol{\Psi}_*^\top &= k(\boldsymbol{X}_*, \boldsymbol{X}_*) = \boldsymbol{K}_{**}.
\end{aligned}$$

- Finalmente, obtemos a versão abaixo da distribuição preditiva:

$$p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{D}, \boldsymbol{S}_0, \sigma^2) = \mathcal{N}(\boldsymbol{y}_*|\boldsymbol{K}_{*f}(\boldsymbol{K} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{y}, \\ \boldsymbol{K}_{**} - \boldsymbol{K}_{*f}(\boldsymbol{K} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{K}_{f*} + \sigma^2\boldsymbol{I}),$$

correspondente a um modelo de processo Gaussiano.

Agenda

Comparação Bayesiana de modelos

- A abordagem Bayesiana evita otimizar parâmetros, preferindo marginalizá-los.
- Não há a necessidade de um conjunto de validação para comparar modelos.
- Dado um modelo $\mathcal{M}_j$ e um conjunto de dados $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$, desejamos computar a distribuição a posteriori do modelo:

$$p(\mathcal{M}_j|\mathcal{D}) = \frac{p(\mathcal{M}_j)p(\mathcal{D}|\mathcal{M}_j)}{p(\mathcal{D})} \propto p(\mathcal{M}_j)p(\mathcal{D}|\mathcal{M}_j).$$

Comparação Bayesiana de modelos

- A abordagem Bayesiana evita otimizar parâmetros, preferindo marginalizá-los.
- Não há a necessidade de um conjunto de validação para comparar modelos.
- Dado um modelo $\mathcal{M}_j$ e um conjunto de dados $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$, desejamos computar a distribuição a posteriori do modelo:

$$p(\mathcal{M}_j|\mathcal{D}) = \frac{p(\mathcal{M}_j)p(\mathcal{D}|\mathcal{M}_j)}{p(\mathcal{D})} \propto p(\mathcal{M}_j)p(\mathcal{D}|\mathcal{M}_j).$$

- Considere L modelos com a priori $p(\mathcal{M}_j)$ iguais.
- Evidência do modelo $\mathcal{M}_j$: $\propto p(\mathcal{D}|\mathcal{M}_j)$.
- **Fator de Bayes** entre modelos com priori iguais:

$$B_{j,k} = \frac{p(\mathcal{D}|\mathcal{M}_j)}{p(\mathcal{D}|\mathcal{M}_k)}.$$

Evidência e Navalha de Occam Bayesiana

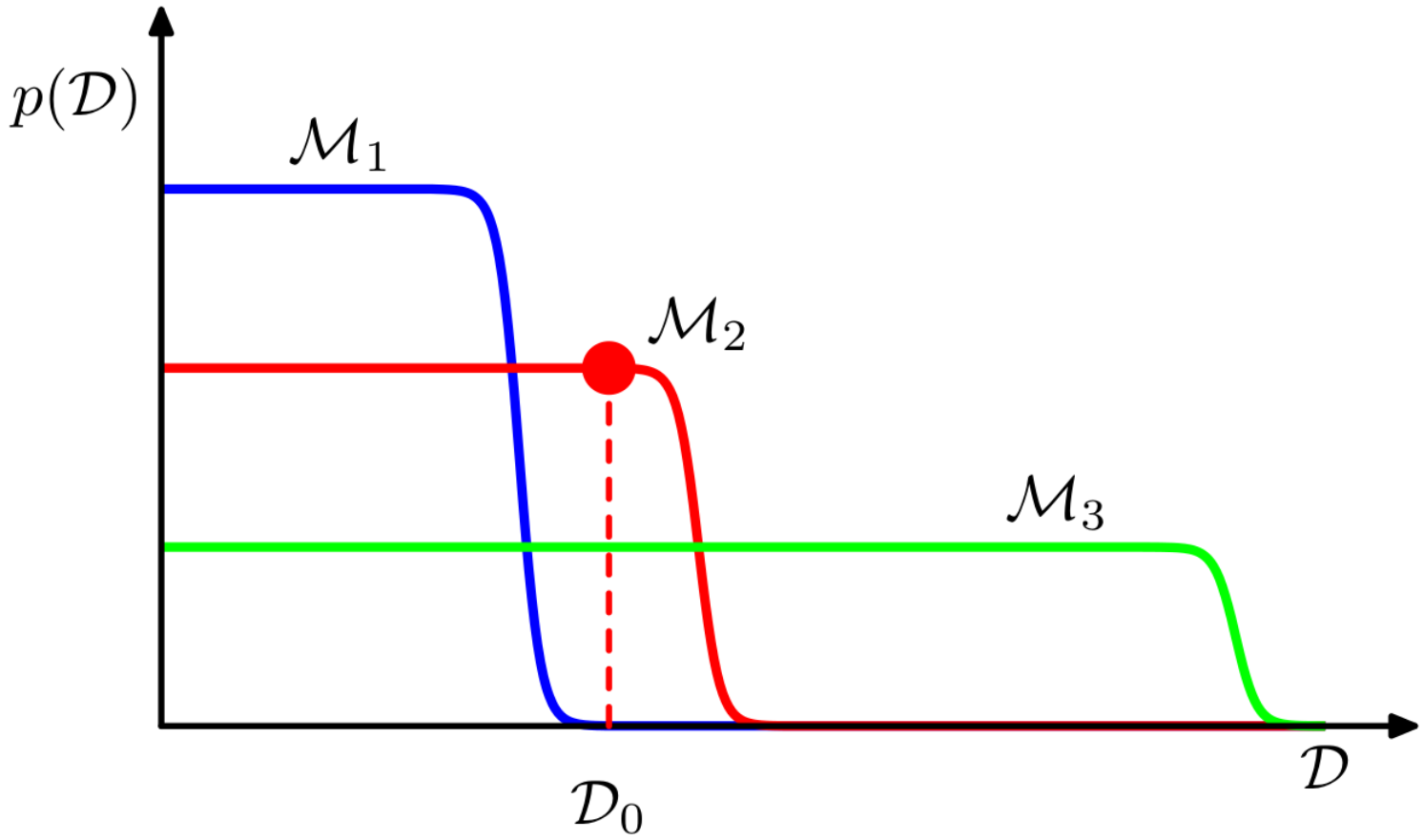
- A evidência é obtida pela marginalização dos parâmetros:

$$p(\mathcal{D}|\mathcal{M}_j) = \int p(\mathcal{D}|\boldsymbol{w}, \mathcal{M}_j) p(\boldsymbol{w}|\mathcal{M}_j) \mathrm{d}\boldsymbol{w}$$

$$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{w}) p(\boldsymbol{w}) \mathrm{d}\boldsymbol{w} \quad \text{(omitindo } \mathcal{M}_j\text{)}.$$

- Soluções ML ou MAP usam somente o ajuste aos dados $p(\mathcal{D}|\boldsymbol{w})$ na comparação de modelos, podendo resultar em overfitting.
- Ao marginalizar os parâmetros, penalizamos modelos muito complexos.
- **Navalha de Occam Bayesiana**: Equilíbrio entre ajuste aos dados e complexidade na comparação Bayesiana de modelos.

Conservação de massa de probabilidade



Mistura e seleção de modelos

- **Mistura de modelos**: Predições são feitas combinando todos os modelos (note que os parâmetros já foram marginalizados):

$$p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{D}) = \sum_{j=1}^{L} p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{M}_j, \mathcal{D}) p(\mathcal{M}_j|\mathcal{D}).$$

Mistura e seleção de modelos

- **Mistura de modelos**: Predições são feitas combinando todos os modelos (note que os parâmetros já foram marginalizados):

$$p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{D}) = \sum_{j=1}^{L} p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{M}_j, \mathcal{D}) p(\mathcal{M}_j|\mathcal{D}).$$

- **Seleção de modelos**: Somente o modelo com maior posteriori $p(\mathcal{M}_{\text{best}}|\mathcal{D})$ é usado na predição.
  - $\rightarrow$ Para modelos com priori iguais, equivale a usar aquele com maior verossimilhança marginal $p(\mathcal{D}|\mathcal{M}_{\text{best}})$.

Evidência aproximada

- Em um tratamento Bayesiano puro, colocaríamos priori nos hiperparâmetros $\boldsymbol{\varphi}$ (por exemplo σ_w^2 e σ^2), que seriam marginalizados junto com os parâmetros para fazer predições:

$$p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{D}) = \int p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{D}, \boldsymbol{w}, \boldsymbol{\varphi}) p(\boldsymbol{w}|\mathcal{D}, \boldsymbol{\varphi}) p(\boldsymbol{\varphi}|\mathcal{D}) \mathrm{d}\boldsymbol{w} \mathrm{d}\boldsymbol{\varphi}.$$

Evidência aproximada

- A integral anterior usualmente não é analítica, podendo ser aproximada por $p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{D}) \approx p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{D}, \hat{\boldsymbol{\varphi}})$:

$$p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{D}, \hat{\boldsymbol{\varphi}}) = \int p(\boldsymbol{y}_*|\boldsymbol{X}_*, \mathcal{D}, \boldsymbol{w}, \hat{\boldsymbol{\varphi}}) p(\boldsymbol{w}|\mathcal{D}, \hat{\boldsymbol{\varphi}}) \mathrm{d}\boldsymbol{w},$$

$$\hat{\boldsymbol{\varphi}} = \arg\max p(\mathcal{D}|\boldsymbol{\varphi}) = \arg\max \int p(\mathcal{D}|\boldsymbol{w}, \boldsymbol{\varphi}) p(\boldsymbol{w}|\boldsymbol{\varphi}) \mathrm{d}\boldsymbol{w}.$$

- O procedimento acima também é chamado de *empirical Bayes*, *type 2 maximum likelihood* ou *generalized maximum likelihood*.

Computando a evidência

- Considere o modelo de regressão abaixo:

$$p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{w}, \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}), \quad \boldsymbol{\Phi} = \phi(\boldsymbol{X}),$$
$$p(\boldsymbol{w}|\sigma_w^2) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \sigma_w^2\boldsymbol{I}).$$

Computando a evidência

- Considere o modelo de regressão abaixo:

$$p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{w}, \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}), \quad \boldsymbol{\Phi} = \phi(\boldsymbol{X}),$$
$$p(\boldsymbol{w}|\sigma_w^2) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \sigma_w^2\boldsymbol{I}).$$

- A evidência é dada por:

$$p(\boldsymbol{y}|\boldsymbol{\Phi}) = \int p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{w}, \sigma^2)p(\boldsymbol{w}|\sigma_w^2)p(\sigma^2, \sigma_w^2)\mathrm{d}\boldsymbol{w}\mathrm{d}\sigma_w^2\mathrm{d}\sigma^2$$
$$\approx p(\boldsymbol{y}|\boldsymbol{\Phi}, \sigma^2, \sigma_w^2) = \int p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{w}, \sigma^2)p(\boldsymbol{w}|\sigma_w^2)\mathrm{d}\boldsymbol{w},$$
$$p(\boldsymbol{y}|\boldsymbol{\Phi}, \sigma^2, \sigma_w^2) = \int \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \sigma_w^2\boldsymbol{I})\mathrm{d}\boldsymbol{w}$$
$$= \mathcal{N}(\boldsymbol{y}|\boldsymbol{0}, \sigma_w^2\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \sigma^2\boldsymbol{I}).$$

- A otimização pode ser feita iterativamente a partir dos dados de treinamento e de valores iniciais (Bishop, pg. 168).

Maximizando a evidência

- Lembrando a diferença entre ML e ML-II:
  - → Máxima verossimilhança (ML):

$$\hat{\boldsymbol{w}} = \arg\max \log p(\mathcal{D}|\boldsymbol{w}).$$

  - → Máxima verossimilhança do tipo II (ML-II):

$$\hat{\boldsymbol{\phi}} = \arg\max \log \int p(\mathcal{D}|\boldsymbol{w}, \boldsymbol{\phi}) p(\boldsymbol{w}|\boldsymbol{\phi}) \mathrm{d}\boldsymbol{w}.$$

- Encontrar os hiperparâmetros $\hat{\boldsymbol{\phi}}$ via ML-II pode ser feito por:
  - → Uso dos gradientes da evidência.
  - → Algoritmo Expectation-Maximization (EM).
  - → Computar a evidência para uma grid de candidatos.

Agenda

Tópicos adicionais

- Regressão Bayesiana com variância σ^2 do ruído desconhecida com priori Gamma inversa $\mathrm{IG}(\sigma^2|a_0, b_0)$:

$$p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{w}, \sigma^2) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}), \quad \boldsymbol{\Phi} = \phi(\boldsymbol{X}),$$
$$p(\boldsymbol{w}, \sigma^2|\boldsymbol{m}_0, \boldsymbol{S}_0, a_0, b_0) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)\mathrm{IG}(\sigma^2|a_0, b_0).$$

  - $\rightarrow$ A posteriori $p(\boldsymbol{w}, \sigma^2|\mathcal{D})$ é analítica (Murphy, pgs. 235 e 236), sendo as marginais $p(\boldsymbol{w}|\mathcal{D})$ e $p(\sigma^2|\mathcal{D})$ respectivamente uma t de Student e uma Gamma inversa.
  - $\rightarrow$ A preditiva $p(\boldsymbol{y}_*|\boldsymbol{X}_*)$ também é analítica (Murphy, pg. 236), sendo uma t de Student.
- Modelos não-conjugados.

Agenda

Referências bibliográficas

- **Cap. 9** - DEISENROTH, M. *et al.* **Mathematics for machine learning**. 2019.
- **Caps. 1 e 7** - MURPHY, Kevin P. **Machine learning: a probabilistic perspective**, 2012.
- **Caps. 2 e 3** - HAYKIN, Simon. **Neural Networks and Learning Machines**, 3ed., 2010.
- **Cap. 3** - BISHOP, Christopher M. **Pattern recognition and machine learning**, 2006.