Universidade Federal do Ceará

Departamento de Computação

Aprendizagem de Máquina

César Lincoln Cavalcante Mattos

2024

Agenda

Regressão polinomial

- 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 polinomial

- 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**: Podemos criar atributos a partir do atributo já existente (**Altura**)?

Regressão polinomial

- Criamos novos atributos via **transformações não-lineares**:

Altura	**Altura^2**	**Comprimento**
1.087	1.087^2	37
1.613	1.613^2	50
0.953	0.953^2	34
1.003	1.003^2	36
1.156	1.156^2	43
0.978	0.978^2	28
1.092	1.092^2	37
0.572	0.572^2	20
0.940	0.940^2	34
0.597	0.597^2	30
0.838	0.838^2	38
1.473	1.473^2	47

Regressão polinomial

- Criamos novos atributos via **transformações não-lineares**:

Altura	**Altura^2**	**Comprimento**
1.087	1.087^2	37
1.613	1.613^2	50
0.953	0.953^2	34
1.003	1.003^2	36
1.156	1.156^2	43
0.978	0.978^2	28
1.092	1.092^2	37
0.572	0.572^2	20
0.940	0.940^2	34
0.597	0.597^2	30
0.838	0.838^2	38
1.473	1.473^2	47

- Modelo **não-linear nos dados** mas **linear nos parâmetros**:

$$\hat{y}_1 = w_0 + w_1 1500 + w_2 1500^2$$
$$\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i$$

Regressão polinomial

- De maneira geral, para um polinômio de ordem P:

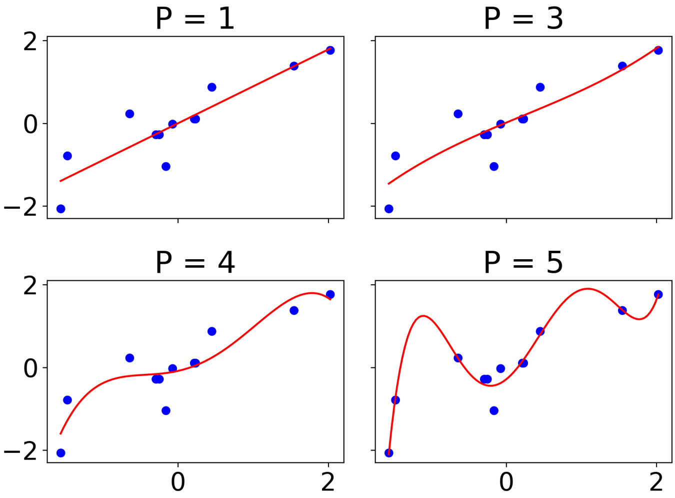
$$\begin{aligned}
\hat{y}_i &= \boldsymbol{w}^\top \boldsymbol{x}_i, \\
\boldsymbol{x}_i &= [1 \ x_i \ x_i^2 \ \cdots \ x_i^P]^\top, \\
\boldsymbol{w} &= [w_0 \ w_1 \ w_2 \ \cdots \ w_P]^\top.
\end{aligned}$$

- Na forma matricial, temos:

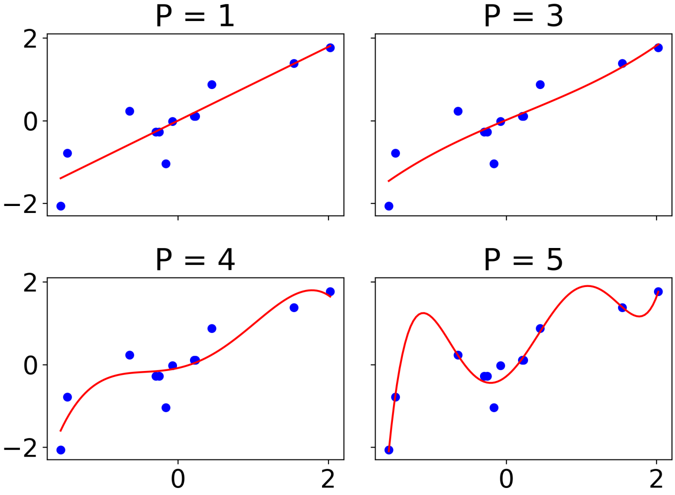
$$\begin{aligned}
\hat{\boldsymbol{y}} &= \boldsymbol{X}\boldsymbol{w}, \\
\hat{\boldsymbol{y}} &= [\hat{y}_1 \ \hat{y}_2 \ \hat{y}_3 \ \cdots \ \hat{y}_N]^\top, \\
\boldsymbol{X} &= \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \boldsymbol{x}_3^\top \\ \vdots \\ \boldsymbol{x}_N^\top \end{bmatrix}.
\end{aligned}$$

- Os parâmetros $\boldsymbol{w}$ podem ser estimados via GD, SGD ou OLS.

Regressão polinomial



Regressão polinomial



- **Problema**: Como escolher a ordem do polinômio?

Agenda

Generalização

- Estamos interessados na capacidade do modelo em predizer corretamente dados não utilizados para ajustar seus parâmetros.

Generalização

- Estamos interessados na capacidade do modelo em predizer corretamente dados não utilizados para ajustar seus parâmetros.
- Procedimento básico:
  1. Separe os dados em dois conjuntos diferentes: **treinamento** e **teste**;
  2. Ajuste (**treine**) os parâmetros do modelo usando somente o conjunto de treinamento;
  3. Verifique a capacidade de **generalização** do modelo no conjunto de teste.

Generalização

Overfitting (sobreajuste)

- Ocorre quando o modelo se ajusta demasiadamente aos dados usados para encontrar seus parâmetros.

Generalização

Overfitting (sobreajuste)

- Ocorre quando o modelo se ajusta demasiadamente aos dados usados para encontrar seus parâmetros.

Underfitting (subajuste)

- Ocorre quando o modelo não possui expressividade suficiente para se ajustar aos dados disponíveis.

Generalização

Overfitting (sobreajuste)

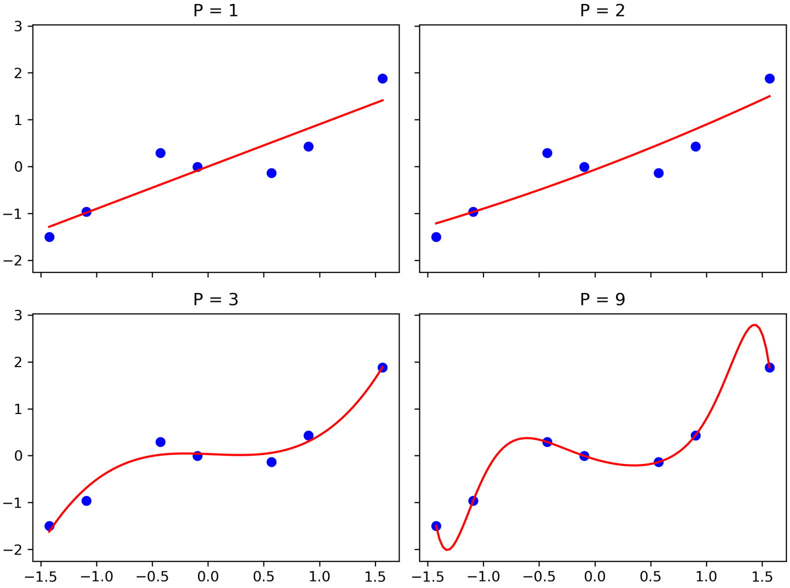
- Ocorre quando o modelo se ajusta demasiadamente aos dados usados para encontrar seus parâmetros.

Underfitting (subajuste)

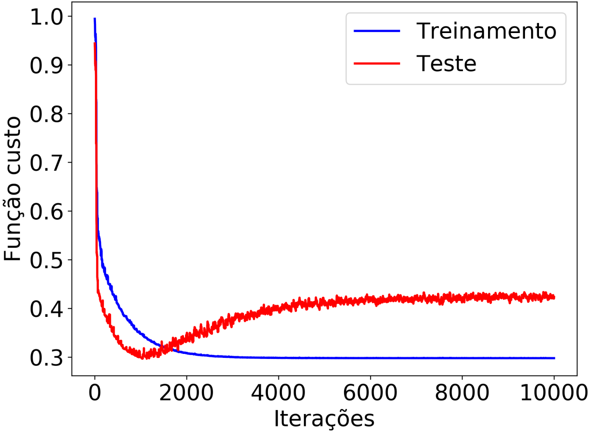
- Ocorre quando o modelo não possui expressividade suficiente para se ajustar aos dados disponíveis.

- Em ambos os cenários o modelo pode apresentar dificuldade de **generalização**.

Generalização



Generalização



Curva de aprendizagem em um cenário de overfitting.

Generalização

- Polinômios de **ordem baixa** podem resultar em **underfitting**.
  - O modelo é **pouco flexível/expressivo**.

Generalização

- Polinômios de **ordem baixa** podem resultar em **underfitting**.
  - O modelo é **pouco flexível/expressivo**.
- Polinômios de **ordem alta** podem resultar em **overfitting**.
  - O modelo é **muito flexível/expressivo**.

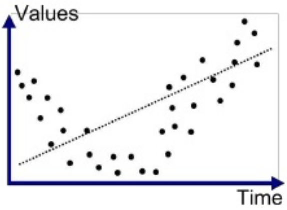
Generalização

- Polinômios de **ordem baixa** podem resultar em **underfitting**.
    - O modelo é **pouco flexível/expressivo**.
- Polinômios de **ordem alta** podem resultar em **overfitting**.
    - O modelo é **muito flexível/expressivo**.

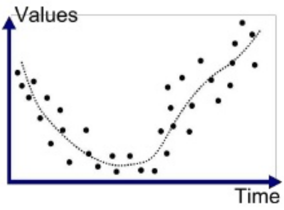
Dilema viés-variância

- **Underfitting** apresenta alto viés e baixa variância.
- **Overfitting** apresenta baixo viés e alta variância.
- O ideal é que viés e variância sejam equilibrados pelo modelo.

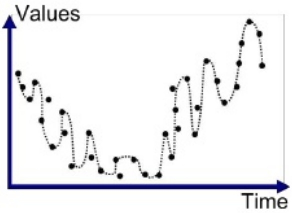
Generalização



Alto viés
Baixa variância



Médio viés
Média variância

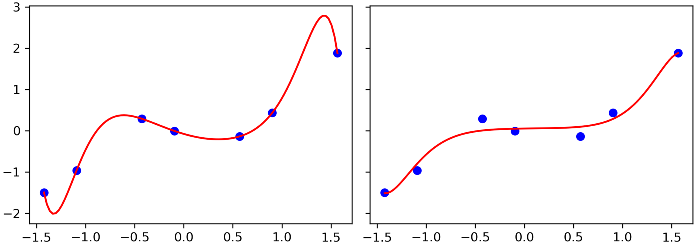


Baixo viés
Alta variância

Agenda

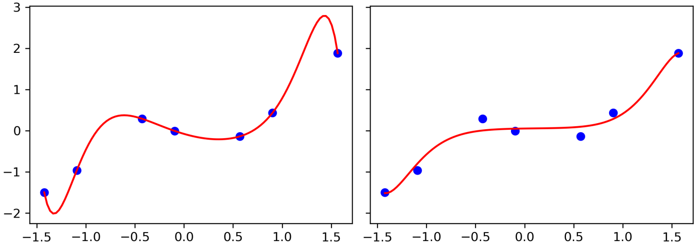
Regularização

- Modelos com **parâmetros grandes** (em módulo), tendem a **generalizar menos**.
  - Modelo com overfitting $\rightarrow \|\hat{\boldsymbol{w}}\|^2 = 2.372582$
- Modelos com **parâmetros pequenos** (em módulo), tendem a **generalizar mais**.
  - Modelo adequado $\rightarrow \|\hat{\boldsymbol{w}}\|^2 = 0.1221954$



Regularização

- Modelos com **parâmetros grandes** (em módulo), tendem a **generalizar menos**.
  - Modelo com overfitting $\rightarrow \|\hat{\boldsymbol{w}}\|^2 = 2.372582$
- Modelos com **parâmetros pequenos** (em módulo), tendem a **generalizar mais**.
  - Modelo adequado $\rightarrow \|\hat{\boldsymbol{w}}\|^2 = 0.1221954$



- **Ideia**: Como evitar o **overfitting** e melhorar a **generalização** mesmo em modelos flexíveis/expressivos?

Regularização

Regularização L2

- Adiciona à função custo um termo proporcional à norma quadrática dos parâmetros:

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

$$\|\boldsymbol{w}\|^2 = \boldsymbol{w}^\top\boldsymbol{w} = \sum_{d=1}^{D} w_d^2.$$

- O termo de regularização induz parâmetros menores.
- $\lambda > 0$ é um **hiperparâmetro de regularização**.
- O parâmetro w_0 não é regularizado.
- Também é chamada de **ridge regression** ou **weight decay**.

Regularização

Regularização L2

- Pode ser aplicada aos métodos de otimização estudados:
  - **Gradiente descendente (GD) regularizado**:

$$\boldsymbol{w}(t) = \boldsymbol{w}(t-1) + \alpha \left[\frac{1}{N} \left(\sum_{i=1}^{N} e_i(t-1)\boldsymbol{x}_i \right) - \lambda \boldsymbol{w}(t-1) \right].$$

  - **Gradiente descendente estocástico (SGD) regularizado**:

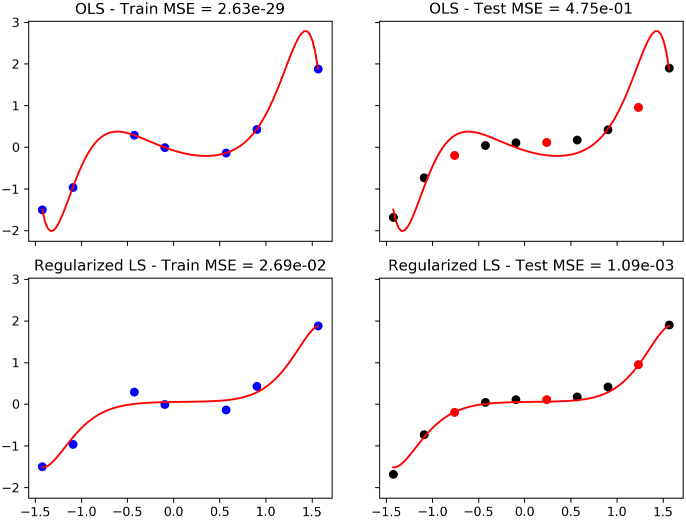
$$\boldsymbol{w}(t) = \boldsymbol{w}(t-1) + \alpha \left[e_i(t-1)\boldsymbol{x}_i - \lambda \boldsymbol{w}(t-1) \right].$$

  - **Mínimos quadrados regularizado**:

$$\boldsymbol{w} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

- Lembre-se que o parâmetro w_0 não deve ser regularizado.

Regressão polinomial ($P = 9$)



Dados de treinamento: pontos **azuis** (com ruído) e **pretos** (sem ruído).
Dados de teste: pontos **vermelhos**.

Regularização L2

- De onde vem o termo de regularização $\frac{\lambda}{2}\|\boldsymbol{w}\|^2$?

Regularização L2

- De onde vem o termo de regularização $\frac{\lambda}{2}\|\boldsymbol{w}\|^2$?
- Lembre que temos uma verossimilhança Gaussiana:

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

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

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

Regularização L2

- Em vez de maximizar $\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$ (solução ML), encontramos o vetor $\boldsymbol{w}$ que maximiza $\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w})$ (solução MAP).

Regularização L2

- Em vez de maximizar $\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$ (solução ML), encontramos o vetor $\boldsymbol{w}$ que maximiza $\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w})$ (solução MAP).
- Escolhendo $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \sigma_w^2 \boldsymbol{I})$, o termo adicional será:

$$\log p(\boldsymbol{w}) = \log \left[\frac{1}{(2\pi\sigma_w^2)^{D_w/2}} \exp\left(-\frac{1}{2\sigma_w^2} \boldsymbol{w}^\top \boldsymbol{w} \right) \right]$$

$$\log p(\boldsymbol{w}) = -\frac{D_w}{2} \log(2\pi\sigma_w^2) - \frac{1}{2\sigma_w^2} \boldsymbol{w}^\top \boldsymbol{w}.$$

Regularização L2

- Em vez de maximizar $\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$ (solução ML), encontramos o vetor $\boldsymbol{w}$ que maximiza $\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w})$ (solução MAP).
- Escolhendo $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \sigma_w^2 \boldsymbol{I})$, o termo adicional será:

$$\log p(\boldsymbol{w}) = \log \left[\frac{1}{(2\pi\sigma_w^2)^{D_w/2}} \exp\left(-\frac{1}{2\sigma_w^2} \boldsymbol{w}^\top \boldsymbol{w} \right) \right]$$
$$\log p(\boldsymbol{w}) = -\frac{D_w}{2} \log(2\pi\sigma_w^2) - \frac{1}{2\sigma_w^2} \boldsymbol{w}^\top \boldsymbol{w}.$$

- Ignorando o termo constante e fazendo $\lambda = \frac{1}{\sigma_w^2}$ e $\boldsymbol{w}^\top \boldsymbol{w} = \|\boldsymbol{w}\|^2$, recuperamos o termo $\frac{\lambda}{2}\|\boldsymbol{w}\|^2$ da regularização L2.

Regularização L2

- Em vez de maximizar $\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$ (solução ML), encontramos o vetor $\boldsymbol{w}$ que maximiza $\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})p(\boldsymbol{w})$ (solução MAP).
- Escolhendo $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \sigma_w^2 \boldsymbol{I})$, o termo adicional será:

$$
\begin{aligned}
\log p(\boldsymbol{w}) &= \log \left[\frac{1}{(2\pi\sigma_w^2)^{D_w/2}} \exp\left(-\frac{1}{2\sigma_w^2} \boldsymbol{w}^\top \boldsymbol{w} \right) \right] \\
\log p(\boldsymbol{w}) &= -\frac{D_w}{2} \log(2\pi\sigma_w^2) - \frac{1}{2\sigma_w^2} \boldsymbol{w}^\top \boldsymbol{w}.
\end{aligned}
$$

- Ignorando o termo constante e fazendo $\lambda = \frac{1}{\sigma_w^2}$ e $\boldsymbol{w}^\top \boldsymbol{w} = \|\boldsymbol{w}\|^2$, recuperamos o termo $\frac{\lambda}{2}\|\boldsymbol{w}\|^2$ da regularização L2.
- **Observação**: A regularização L2 equivale a uma **priori Gaussiana** para os parâmetros e uma **solução MAP**.

Agenda

1. Regressão polinomial
2. Generalização
3. Regularização
4. Seleção de hiperparâmetros e avaliação de modelos
5. Normalização dos dados
6. Tópicos adicionais
7. Referências

Seleção de hiperparâmetros e avaliação de modelos

- Qual a diferença entre um **parâmetro** e um **hiperparâmetro**?

Seleção de hiperparâmetros e avaliação de modelos

- Qual a diferença entre um **parâmetro** e um **hiperparâmetro**?
- Como escolher o hiperparâmetro de regularização λ?

Seleção de hiperparâmetros e avaliação de modelos

- Qual a diferença entre um **parâmetro** e um **hiperparâmetro**?
- Como escolher o hiperparâmetro de regularização λ?

Validação cruzada (hold-out validation)

- Separe um terceiro conjunto de dados de **validação** para ajuste dos hiperparâmetros.
- O conjunto de validação pode ser usado para estimar a **generalização** do modelo antes de usar o conjunto de teste.
- O modelo final deve ser obtido a partir dos conjuntos de treinamento e validação.
- O conjunto de teste **não** deve ser usado até a obtenção do modelo final.

Seleção de hiperparâmetros e avaliação de modelos

- Qual a diferença entre um **parâmetro** e um **hiperparâmetro**?
- Como escolher o hiperparâmetro de regularização λ?

Validação cruzada (hold-out validation)

- Separe um terceiro conjunto de dados de **validação** para ajuste dos hiperparâmetros.
- O conjunto de validação pode ser usado para estimar a **generalização** do modelo antes de usar o conjunto de teste.
- O modelo final deve ser obtido a partir dos conjuntos de treinamento e validação.
- O conjunto de teste **não** deve ser usado até a obtenção do modelo final.

- Validação cruzada pode ser usada para escolher o hiperparâmetro de regularização λ, a ordem do polinômio P e/ou o passo de aprendizagem α.

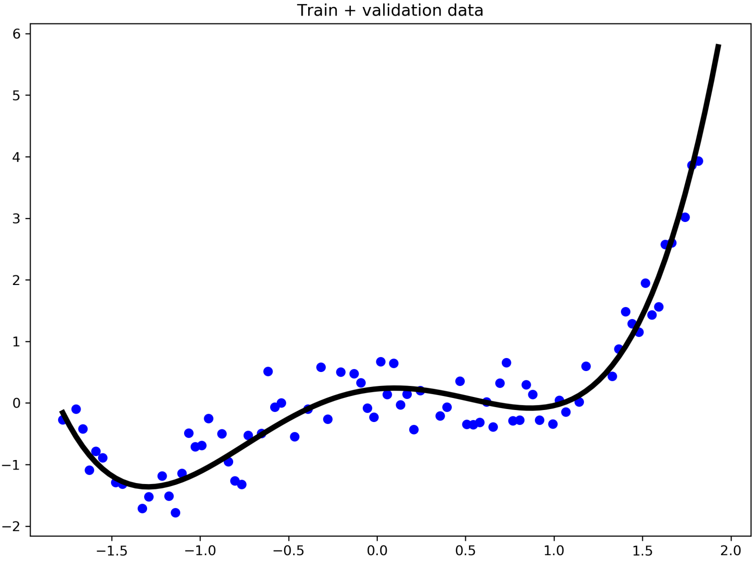
Seleção de hiperparâmetros

Grid search

1. Separe os dados em 3 conjuntos: **treinamento**, **validação** e **teste**;
2. Construa uma lista de hiperparâmetros candidados;
3. Treine o modelo para o primeiro hiperparâmetro candidado;
4. Verifique a qualidade do modelo obtido no conjunto de validação;
5. Repita os dois passos anteriores para os demais candidatos;
6. Escolha o hiperparâmetro com melhor avaliação no conjunto de validação;
7. Retreine o modelo usando o conjunto de treinamento e de validação;
8. Verifique a generalização do modelo no conjunto de teste.

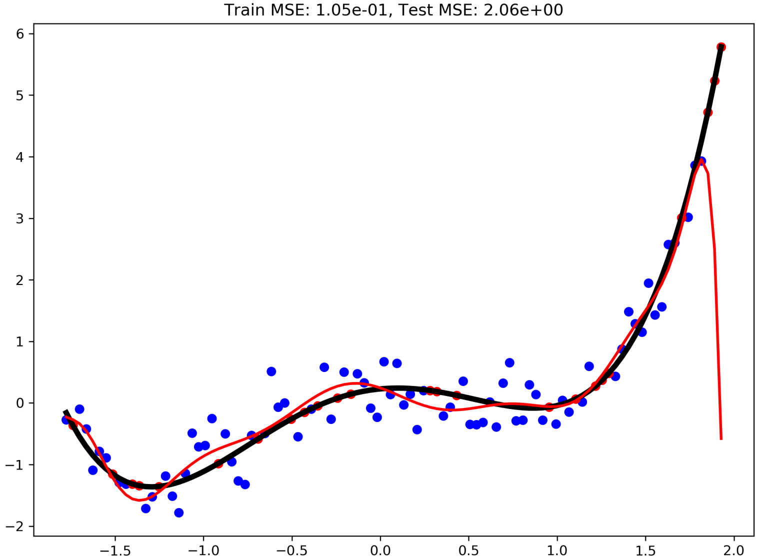
Seleção de hiperparâmetros

Dados de treinamento+validação em **azul**, função real em **preto**



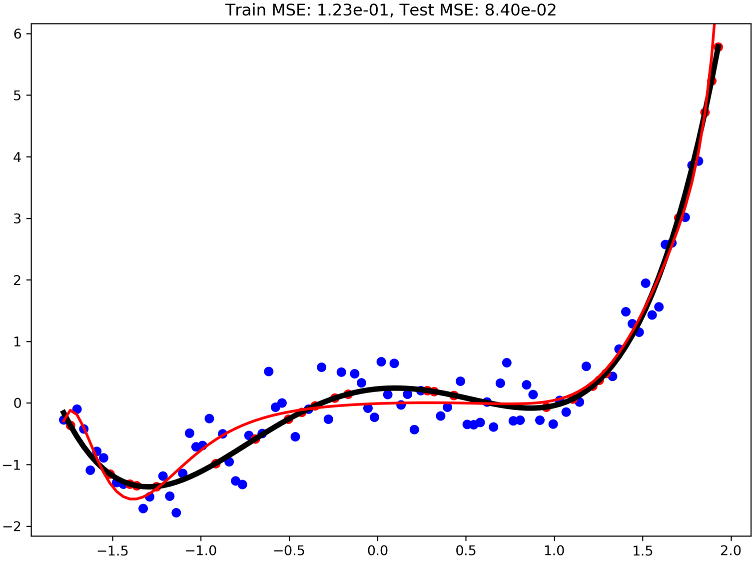
Seleção de hiperparâmetros

Regressão polinomial ($P = 15$) na curva **vermelha**, pontos de teste em **vermelho**



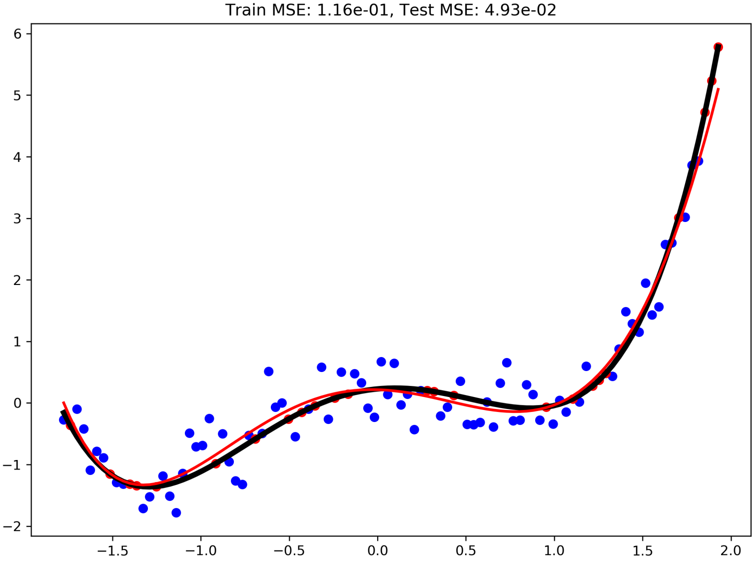
Seleção de hiperparâmetros

Regressão polinomial ($P = 15$) com grid search para $\lambda = 7.196857$



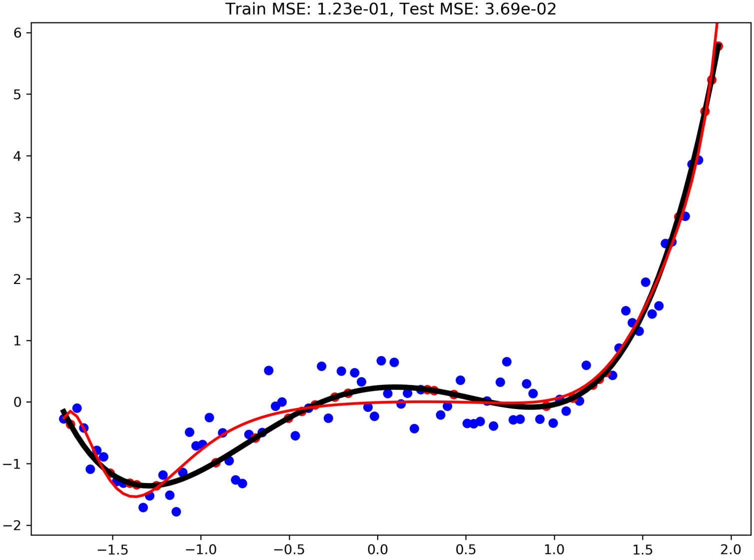
Seleção de hiperparâmetros

Regressão polinomial com grid search para $P = 5$



Seleção de hiperparâmetros

Regressão polinomial com grid search para $P = 13$ e $\lambda = 7.196857$



Seleção de hiperparâmetros e avaliação de modelos

K-fold cross-validation

1. Divida o conjunto de treinamento aleatoriamente em K partições iguais;
2. Uma das partições é mantida para teste, enquanto as $K-1$ demais são usadas para treinar o modelo;
3. Repita o passo anterior para cada uma das partições;
4. Retorne o resultado médio obtido para cada partição.

Seleção de hiperparâmetros e avaliação de modelos

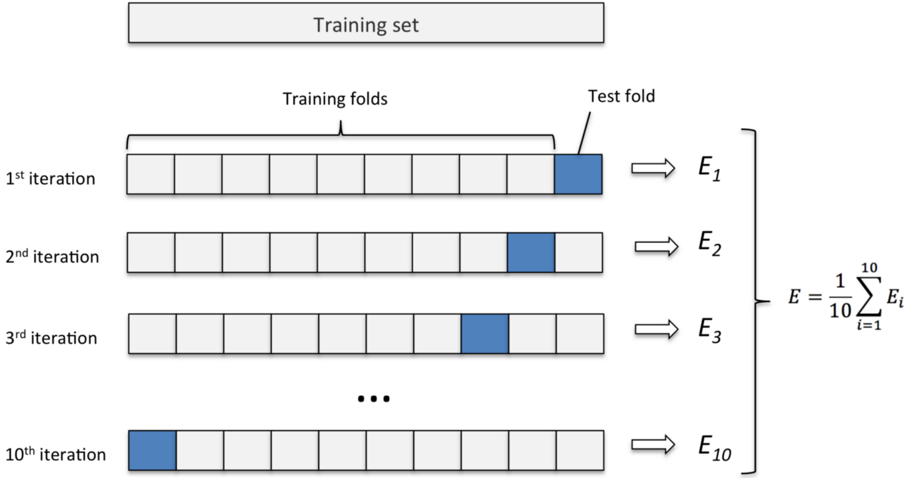
K-fold cross-validation

1. Divida o conjunto de treinamento aleatoriamente em K partições iguais;
2. Uma das partições é mantida para teste, enquanto as $K-1$ demais são usadas para treinar o modelo;
3. Repita o passo anterior para cada uma das partições;
4. Retorne o resultado médio obtido para cada partição.

Leave-one-out (LOO)

- O mesmo que K-fold cross-validation quando $K = N$.

K-fold cross-validation



Agenda

Normalização dos dados

- Os dados disponíveis podem estar em escalas muito diferentes.
  - Idade, altura, peso, salário, etc.
- Usualmente é recomendado colocar os dados (de entrada e saída) em uma escala comum.
- Podemos normalizar entre os intervalos $[0, 1]$ ou $[-1, 1]$ ou forçar média 0 e variância unitária.
- **Vantagens**: Maior controle dos valores dos parâmetros; maior facilidade em ajustar os hiperparâmetros.
- **Importante**: Os dados de validação/teste devem ser atualizados de acordo com as estatísticas dos dados de treinamento.

Normalização dos dados

Exemplo de normalização dos dados para o intervalo $[0, 1]$

1. Calcule o valor máximo do vetor $\boldsymbol{y}$ e da matriz $\boldsymbol{X}$, coluna a coluna:

$$y_{\text{max}} = \max(\boldsymbol{y}), \quad [\boldsymbol{x}_{\text{max}}]_d = \max([\boldsymbol{X}]_{:d}), \forall d.$$

2. Calcule o valor mínimo do vetor $\boldsymbol{y}$ e da matriz $\boldsymbol{X}$, coluna a coluna:

$$y_{\text{min}} = \min(\boldsymbol{y}), \quad [\boldsymbol{x}_{\text{min}}]_d = \min([\boldsymbol{X}]_{:d}), \forall d.$$

3. Conclua a normalização dos dados:

$$\boldsymbol{y} \leftarrow \frac{\boldsymbol{y} - y_{\text{min}}}{y_{\text{max}} - y_{\text{min}}}, \quad [\boldsymbol{X}]_{:d} \leftarrow \frac{[\boldsymbol{X}]_{:d} - [\boldsymbol{x}_{\text{min}}]_d}{[\boldsymbol{x}_{\text{max}}]_d - [\boldsymbol{x}_{\text{min}}]_d}, \forall d.$$

Normalização dos dados

Exemplo de normalização dos dados via z-score

1. Calcule a média do vetor $\boldsymbol{y}$ e da matriz $\boldsymbol{X}$, coluna a coluna:

$$\mu_y = \mathbb{E}[y], \quad [\boldsymbol{\mu}_x]_d = \mathbb{E}[x_d], \forall d.$$

2. Retire a média dos dados (as operações seguem o Numpy):

$$\boldsymbol{y} \leftarrow \boldsymbol{y} - \mu_y, \quad [\boldsymbol{X}]_{:d} \leftarrow [\boldsymbol{X}]_{:d} - [\boldsymbol{\mu}_x]_d, \forall d.$$

3. Calcule o desvio padrão de $\boldsymbol{y}$ e da matriz $\boldsymbol{X}$, coluna a coluna:

$$\sigma_y = \sqrt{\mathbb{V}[y]}, \quad [\boldsymbol{\sigma}_x]_d = \sqrt{\mathbb{V}[x_d]}, \forall d.$$

4. Conclua a normalização dos dados:

$$\boldsymbol{y} \leftarrow \frac{\boldsymbol{y}}{\sigma_y}, \quad [\boldsymbol{X}]_{:d} \leftarrow \frac{[\boldsymbol{X}]_{:d}}{[\boldsymbol{\sigma}_x]_d}, \forall d.$$

Normalização dos dados

- Antes de calcular o erro de teste, é interessante "desnormalizar" as predições e usar os dados de teste não-normalizados.
- No caso de normalização via z-score teríamos:

$$\hat{\boldsymbol{y}} \leftarrow (\hat{\boldsymbol{y}} \times \sigma_y) + \mu_y.$$

- **Importante**: Utilize as médias e desvios computados com os dados de treinamento para fazer a normalização e a "desnormalização".

Agenda

Tópicos adicionais

- MAP $\times$ regularização.
- Regularização L1 (least absolute shrinkage and selection operator - lasso).
- Regularização via elastic net (L1 $+$ L2).
- Esparsidade via regularização.
- Métodos robustos a outliers (valores discrepantes).
- Double descent e generalização em modelos sobreparametrizados.

Agenda

Referências bibliográficas

- **Caps. 7 e 13*** - MURPHY, Kevin P. **Machine learning: a probabilistic perspective**, 2012.
- **Caps. 1 e 4** - MURPHY, Kevin P. **Probabilistic Machine Learning: An Introduction**, 2021.
- **Cap. 9** - DEISENROTH, M. *et al.* **Mathematics for machine learning**. 2019.