Deep Learning

Revisão de *Machine Learning*

Lucas Silveira Kupssinskü

Machine Learning Theory and Applications Laboratory
PUCRS

agosto de 2026

Overview

Visão Geral

A importância de Machine Learning

- “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates)
- “Machine learning is the next Internet” (Tony Tether, Director, DARPA)
- “Machine learning is going to result in a real revolution” (Greg Papadopoulos, Former CTO, Sun)
- “Machine learning is today’s discontinuity” (Jerry Yang, Founder, Yahoo)
- “Machine learning is the new electricity” (Andrew NG, Stanford Professor and Coursera founder)

Software sem *Machine Learning*

É uma sequência de instruções que:

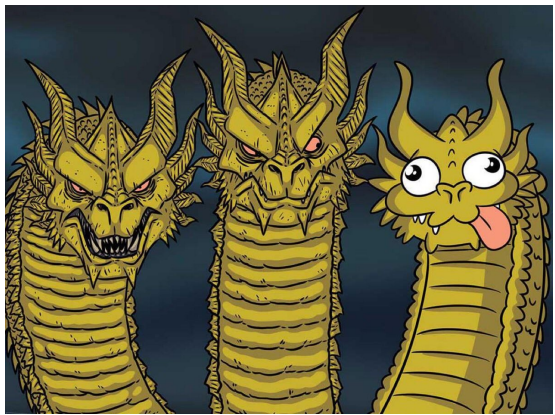
- Recebe uma entrada
- Realiza um processamento
- Retorna uma saída

Software sem *Machine Learning*
Exemplo

1. Se você ou o seu oponente tiverem dois em sequência, jogue no espaço restante
2. Caso contrário, se existir um movimento que crie duas sequencias de dois, faça essa jogada
3. Caso contrário, se o centro estiver vazio, jogue no centro
4. Caso contrário, se o seu oponente jogou em um canto, jogue no canto oposto
5. Caso contrário, se houver um canto livre jogue nesse canto
6. Caso contrário, jogue em uma posição vazia qualquer

Algumas vezes não sabemos como resolver

- Complexidade de Tempo
- Complexidade de Espaço
- Complexidade “Humana”



E no Aprendizado de Máquina?

- Mesmo quando não é possível criar um algoritmo para resolver o problema diretamente, pode ser possível criar algoritmos que *encontram padrões*
- Usamos raciocínio indutivo para resolver os problemas

Definição de *Machine Learning*

"A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**" (Mitchell, 1997)

Tipos de Raciocínio

- **Dedutivo**
  - Assumindo algumas premissas, a conclusão obrigatoriamente segue
- **Não-dedutivo**
  - **Abdução**
    - "Raciocínio do Detetive"
  - **Indução**
    - Dadas observações (limitadas em escopo e quantidade), podemos chegar a conclusões plausíveis

Exemplo de Dedução

- Premissa 1: Todos os humanos são mortais
- Premissa 2: Sócrates é humano
- Conclusão: Sócrates é mortal

Tipos de Raciocínio

- **Dedutivo**
  - Assumindo algumas premissas, a conclusão obrigatoriamente segue
- **Não-dedutivo**
  - **Abdução**
    - "Raciocínio do Detetive"
  - **Indução**
    - Dadas observações (limitadas em escopo e quantidade), podemos chegar a conclusões plausíveis

Exemplo de Abdução

- Tem restos de frango assado pelo chão
- O cachorro está com restos de frango no pelo, especialmente ao redor da boca
- Logo, o cachorro comeu o frango assado

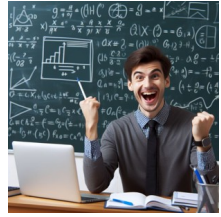
Tipos de Raciocínio

- **Dedutivo**
  - Assumindo algumas premissas, a conclusão obrigatoriamente segue
- **Não-dedutivo**
  - **Abdução**
    - "Raciocínio do Detetive"
  - **Indução**
    - Dadas observações (limitadas em escopo e quantidade), podemos chegar a conclusões plausíveis

Exemplo de Indução

Qual o próximo número na sequência abaixo?

$$1, 3, 5, 7, x$$



Tipos de Raciocínio

- **Dedutivo**
  - Assumindo algumas premissas, a conclusão obrigatoriamente segue
- **Não-dedutivo**
  - **Abdução**
    - "Raciocínio do Detetive"
  - **Indução**
    - Dadas observações (limitadas em escopo e quantidade), podemos chegar a conclusões plausíveis

Exemplo de Indução

Qual o próximo número na sequência abaixo?

$$1, 3, 5, 7, 217341$$

$$f(x) = \frac{18111}{2}x^4 - 90555x^3 + \frac{633885}{2}x^2 - 452773x + 217331$$

Visão Geral

Desenvolvimento de uma aplicação com AM

O projeto de um sistema que usa *Machine Learning* pode ser descrito pela figura abaixo[1]

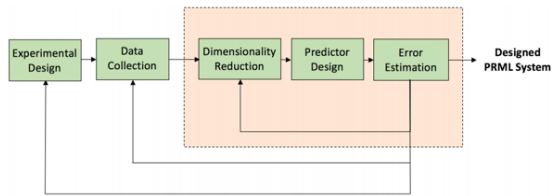


Figura: Representação simplificada do processo de *Machine Learning*. Adaptado de (Braga-Neto, 2020)

[1]Fluxos diferentes existem na literatura de *Knowledge Discovery in Databases* e em Mineração de Dados.

Desenvolvimento de uma aplicação com AM

Experimental Design

- Modelar o problema como uma tarefa de *Machine Learning*
- Como você definiria o problema de treinar um reconhecedor de rostos que é capaz de reconhecer 100 pessoas diferentes?
  - Como um problema de classificação multiclasse com 100 classes será muito difícil coletar uma base dados grande o bastante
  - Como um problema de classificação binário onde o sistema recebe um par de imagens é mais fácil coletar dados

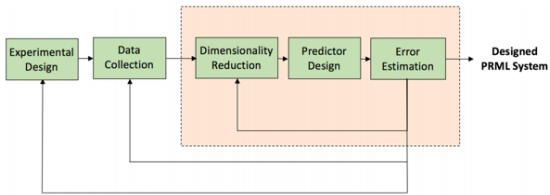


Figura: Representação simplificada do processo de *Machine Learning*. Adaptado de (Braga-Neto, 2020)

Desenvolvimento de uma aplicação com AM

Data Collection

- É possível coletar dados que sejam representativos do problema em questão?
- Existe algum dataset aberto que pode nos auxiliar?
- Quando a coleta de dados ou a anotação é cara, existem alternativas?

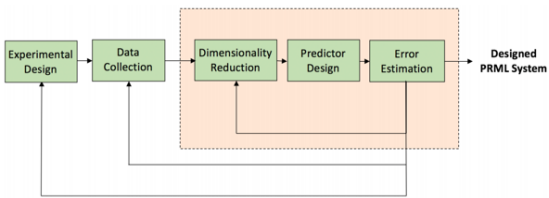


Figura: Representação simplificada do processo de *Machine Learning*. Adaptado de (Braga-Neto, 2020)

Desenvolvimento de uma aplicação com AM

Dimensionality Reduction

- Todos os atributos coletados são úteis para resolver a tarefa?
- Como selecionar apenas os atributos relacionados a variável alvo?
- Existem formas de criar novos atributos que sejam melhores preditores para a variável alvo?
- Posso resumir meus atributos em um espaço menor?

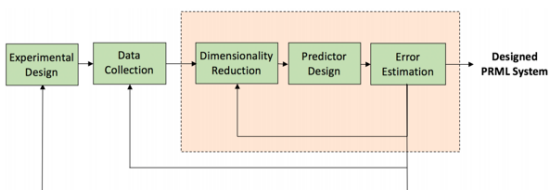


Figura: Representação simplificada do processo de *Machine Learning*. Adaptado de (Braga-Neto, 2020)

Desenvolvimento de uma aplicação com AM

Predictor Design

- Existe algum modelo cujo viés indutivo seja aderente ao meu problema?
- Qual será a métrica utilizada para avaliação do desempenho do modelo?
- Qual meu *budget* computacional?
- Qual a melhor hiperparametrização dos modelos selecionados?

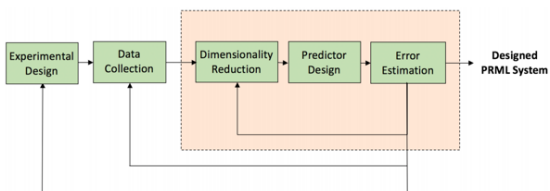


Figura: Representação simplificada do processo de *Machine Learning*. Adaptado de (Braga-Neto, 2020)

Desenvolvimento de uma aplicação com AM

Error Estimation

- Meu modelo treinado generaliza?
- Qual é o erro esperado do meu modelo para novos dados?
- Coletar mais dados pode melhorar meus resultados?

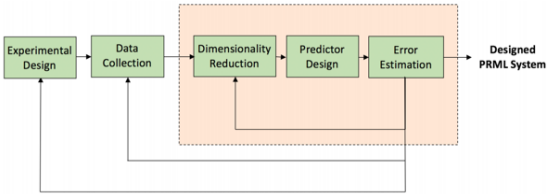


Figura: Representação simplificada do processo de *Machine Learning*. Adaptado de (Braga-Neto, 2020)

Tipos de Aprendizado

- **Supervisionado**
  - Dados de treino possuem rótulos
  - *Exemplo*: classificar fotos como gato ou cachorro a partir de 10 mil imagens rotuladas
- **Não Supervisionado**
  - Dados de treino não possuem rótulos
  - *Exemplo*: agrupar clientes em segmentos a partir do histórico de compras, sem rótulos
- **Semi Supervisionado**
  - Alguns dados de treino possuem rótulo outros não
  - *Exemplo*: detecção de fraude com poucas transações rotuladas e milhões sem rótulo
- **Aprendizado por Reforço**
  - Agente recebe "recompensa" devido a uma série de ações
  - *Exemplo*: AlphaGo aprende a jogar Go ganhando recompensa ao final da partida
- **Auto Supervisionado**
  - Parte das entradas servem como "supervisão"
  - *Exemplo*: pré-treino do BERT prevendo palavras mascaradas em uma frase

Notação

Considerando um dataset formado por exemplos de uma relação $\left(\mathbf{X}^{(i)}, \mathbf{y}^{(i)}\right)$

- $\mathbf{X}$ é nossa matriz de atributos $N \times M$ onde N é o número de instâncias e M é o número de atributos indexada por $\mathbf{X}_m^{(n)}$
- $\mathbf{y}$ é nossa variável alvo $N \times 1$
- Objetivo é estimar $\mathbf{y}^{(i)}$ para uma instância desconhecida $\mathbf{X}^{(i)}$
  - Quando $\mathbf{y}^{(i)}$ é uma variável discreta: *Problema de Classificação*
  - Quando $\mathbf{y}^{(i)}$ é uma variável contínua: *Problema de Regressão*

Viés Indutivo

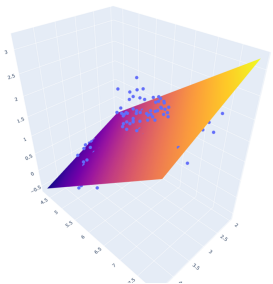


Figura: Regressão Linear

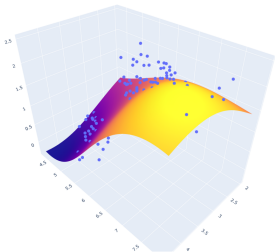


Figura: SVM

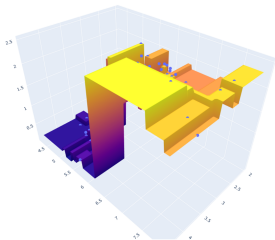


Figura: Árvore de Decisão

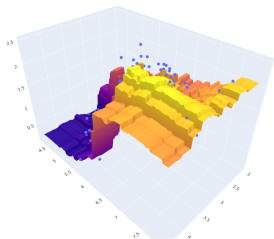


Figura: *Random Forest*

Efeito tesoura (*Scissor Effect*)

- Quando temos um conjunto de dados pequeno, regras de classificação mais simples tendem a obter erros de classificação menores
- Quando temos um conjunto de dados grande, regras de classificação mais complexas tendem a obter erros menores

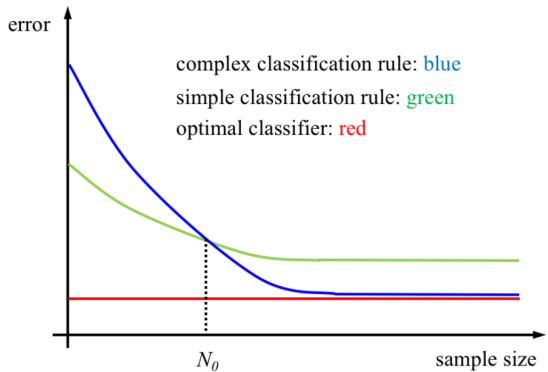
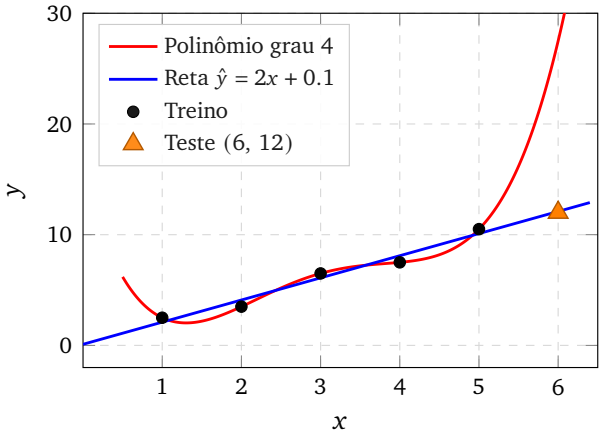


Figura: Adaptado de (Braga-Neto, 2020)

Efeito tesoura, exemplo numérico



Em $x = 6$: reta prediz 12.1 (acerta), polinômio prediz 27.5 (erra muito).

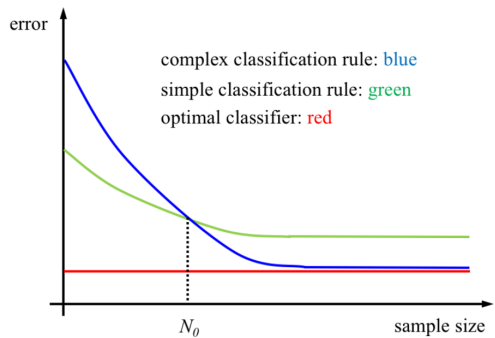


Figura: Adaptado de (Braga-Neto, 2020)

Trade-off Viés-Variância

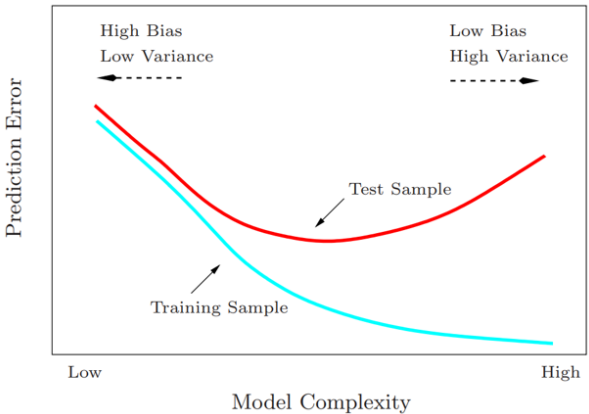


Figura: Adaptado de (Braga-Neto, 2020).

Decomposição

$$\text{Erro}_{\text{teste}}(x) = \text{Viés}^2(x) + \text{Variância}(x) + \sigma^2$$

- **Viés alto**: modelo simples demais, não captura o padrão.
- **Variância alta**: modelo sensível à amostra de treino.
- **Ruído** (σ^2): limite irredutível.

Derivação completa em *Regularização* (Aula 08).

Visão Geral

Modelos Lineares

- Modelos lineares baseados em variações dos Método dos Mínimos Quadrados talvez sejam o tipo de modelo mais estudado por todas as áreas do conhecimento.
- Faremos uma breve revisão com viés de *Machine Learning*

Intuição

Queremos ajustar um *hiperplano*[1] aos nossos dados

$$\hat{\mathbf{y}} = \mathbf{X}_1 \cdot \theta_1 + \theta_0$$

[1] No $\mathbb{R}^2$ um hiperplano é equivalente a uma reta

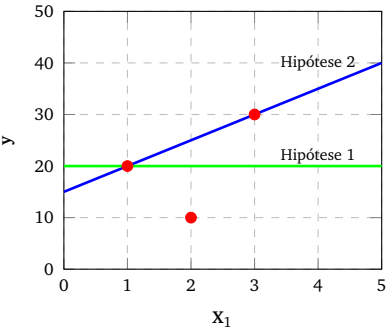
Intuição

Queremos ajustar um *hiperplano*[1] aos nossos dados

$$\hat{\mathbf{y}} = \mathbf{X}_1 \cdot \theta_1 + \theta_0$$

X	y
1	20
2	10
3	30

Tabela: Conjunto de Dados



Hipótese 1

$\theta_0 = 20, \theta_1 = 0$

$\hat{\mathbf{y}} = 0\mathbf{X}_1 + 20$

Hipótese 2

$\theta_0 = 15, \theta_1 = 5$

$\hat{\mathbf{y}} = 5\mathbf{X}_1 + 15$

[1] No $\mathbb{R}^2$ um hiperplano é equivalente a uma reta

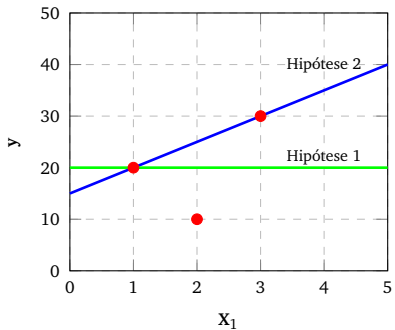
Intuição

Qual é a melhor hipótese?

$$J_{\text{SSR}} = \sum_{i=1}^{N} \left(\hat{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)} \right)^2$$

X	y
1	20
2	10
3	30

Tabela: Conjunto de Dados



Hipótese 1

$$\theta_0 = 20, \theta_1 = 0$$

$$\hat{\mathbf{y}} = 0\mathbf{X}_1 + 20$$

Hipótese 2

$$\theta_0 = 15, \theta_1 = 5$$

$$\hat{\mathbf{y}} = 5\mathbf{X}_1 + 15$$

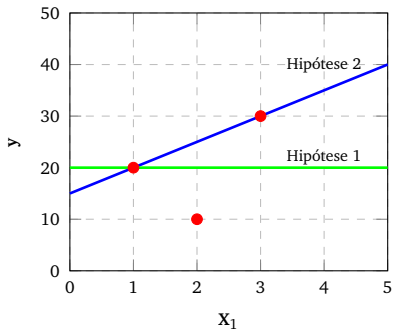
Intuição

Qual é a melhor hipótese?

$$J_{\text{SSR}} = \sum_{i=1}^{N} \left(\hat{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)} \right)^2$$

X	y
1	20
2	10
3	30

Tabela: Conjunto de Dados



✅ Hipótese 1

$$\theta_0 = 20, \theta_1 = 0$$

$$\hat{\mathbf{y}} = 0\mathbf{X}_1 + 20$$

$$J_{\text{SSR}} = (20-20)^2 + (10-20)^2 + (30-20)^2 = 200$$

❌ Hipótese 2

$$\theta_0 = 15, \theta_1 = 5$$

$$\hat{\mathbf{y}} = 5\mathbf{X}_1 + 15$$

$$J_{\text{SSR}} = (20-20)^2 + (10-25)^2 + (30-30)^2 = 225$$

Treinamento

Como encontrar a melhor hipótese?

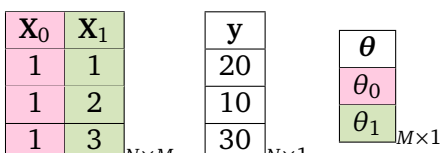
- $\text{argmin}_{\theta_0,\theta_1} [J_{\text{SSR}}]$
- $J_{\text{SSR}} = \sum_{i=1}^{N} \left(\hat{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}\right)^2 = (\hat{\mathbf{y}} - \mathbf{y})^\top (\hat{\mathbf{y}} - \mathbf{y})$
- $\hat{\mathbf{y}} = \mathbf{X}_1 \cdot \theta_1 + \theta_0$

$\mathbf{X}_1$	$\mathbf{y}$
1	20
2	10
3	30

θ_0 θ_1

Notation Trick

- Colocamos o parâmetro livre junto do vetor θ fazendo um *truque de notação*
- Imagine que existe um novo atributo cujo valor é 1 para todas as instâncias
- Assim todos os parâmetros treináveis podem ser tratados da mesma forma.
- $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta}$



Treinamento

Como encontrar a melhor hipótese?

- $\text{argmin}_{\theta_0,\theta_1} [J_{\text{SSR}}]$
- $J_{\text{SSR}} = \sum_{i=1}^{N} \left(\hat{\mathbf{y}}^{(i)} - \mathbf{y}^{(i)}\right)^2 = (\hat{\mathbf{y}} - \mathbf{y})^\top (\hat{\mathbf{y}} - \mathbf{y})$
- $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta}$

$\mathbf{X}_0$	$\mathbf{X}_1$
1	1
1	2
1	3

$N \times M$

$\mathbf{y}$
20
10
30

$N \times 1$

$\boldsymbol{\theta}$
θ_0
θ_1

$M \times 1$

Sabendo que..

- A soma dos quadrados dos resíduos é uma *função de custo convexa* para o problema da regressão linear
- Logo, *possui apenas um mínimo*
- Podemos utilizar o gradiente da função de custo em um processo iterativo de otimização

Treinamento

Vamos derivar J_{SSR} em relação a θ e igualar a 0

$$
\begin{aligned}
\frac{\partial J_{\text{SSR}}}{\partial \theta} &= \frac{\partial((\hat{\mathbf{y}} - \mathbf{y})^\top(\hat{\mathbf{y}} - \mathbf{y}))}{\partial \theta} \\
&= \frac{\partial(\hat{\mathbf{y}}^\top\hat{\mathbf{y}} - \hat{\mathbf{y}}^\top\mathbf{y} - \mathbf{y}^\top\hat{\mathbf{y}} + \mathbf{y}^\top\mathbf{y})}{\partial \theta} \\
&= \frac{\partial(\hat{\mathbf{y}}^\top\hat{\mathbf{y}} - 2\hat{\mathbf{y}}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y})}{\partial \theta} \\
&= \frac{\partial(\boldsymbol{\theta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta} - 2\boldsymbol{\theta}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y})}{\partial \theta} \\
&= \frac{\partial(\boldsymbol{\theta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta})}{\partial \theta} - \frac{\partial(2\boldsymbol{\theta}^\top\mathbf{X}^\top\mathbf{y})}{\partial \theta} \\
&= 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta} - 2\mathbf{X}^\top\mathbf{y} \\
&= 2\mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})
\end{aligned}
$$

Treinamento

Vamos derivar J_{SSR} em relação a θ e igualar a 0

$$
\begin{aligned}
2\mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) &= 0 \\
\mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) &= 0 \\
\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta} - \mathbf{X}^\top\mathbf{y} &= 0 \\
\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta} &= \mathbf{X}^\top\mathbf{y} \\
\boldsymbol{\theta} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}
\end{aligned}
$$

Solução analítica para θ

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

$$\left(\underset{M \times N}{\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}} \times \underset{N \times M}{\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}} \right)^{-1} \times \underset{M \times N}{\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}} \times \underset{N \times 1}{\begin{bmatrix} 20 \\ 10 \\ 30 \end{bmatrix}} = \underset{M \times 1}{\boldsymbol{\theta}}$$

Solução analítica para θ

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

$$\left(\underset{M \times M}{\begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}} \right)^{-1} \times \underset{M \times N}{\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}} \times \underset{N \times 1}{\begin{bmatrix} 20 \\ 10 \\ 30 \end{bmatrix}} = \underset{M \times 1}{\boldsymbol{\theta}}$$

Solução analítica para θ

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

$$\underset{M \times M}{\begin{bmatrix} 2.33 & -1 \\ -1 & 0.5 \end{bmatrix}} \times \underset{M \times N}{\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}} \times \underset{N \times 1}{\begin{bmatrix} 20 \\ 10 \\ 30 \end{bmatrix}} = \underset{M \times 1}{\boldsymbol{\theta}}$$

Solução analítica para θ

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

$$\underbrace{\begin{bmatrix} 1.33 & 0.333 & -0.666 \\ -0.5 & 0 & 0.5 \end{bmatrix}}_{M \times N} \times \underbrace{\begin{bmatrix} 20 \\ 10 \\ 30 \end{bmatrix}}_{N \times 1} = \underset{M \times 1}{\boldsymbol{\theta}}$$

Solução analítica para θ

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

$$\underset{M \times 1}{\begin{bmatrix} 10 \\ \hline 5 \end{bmatrix}} = \underset{M \times 1}{\boldsymbol{\theta}}$$

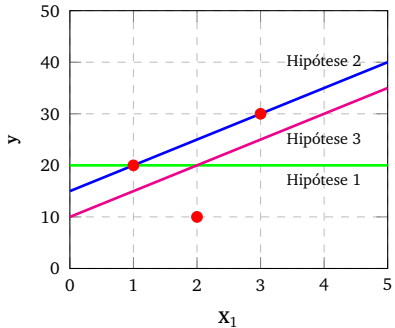
Sobre a solução analítica para θ

- Esse é o vetor de pesos que minimiza a soma dos quadrados dos resíduos (J_{SSR}).
- Repare que existe uma operação de inversão da matriz $\mathbf{X}^\top \mathbf{X}$. Essa operação gera um custo computacional $O(M^3)$ em relação à quantidade de atributos do problema
  - Esse custo pode inviabilizar a solução analítica

Qual é a melhor hipótese?

X	y
1	20
2	10
3	30

Tabela: Conjunto de Dados



❌ Hipótese 1

$$\theta_0 = 20, \theta_1 = 0, \hat{\mathbf{y}} = 0\mathbf{X}_1 + 20$$

$$J_{\text{SSR}} = (20-20)^2 + (10-20)^2 + (30-20)^2 = 200$$

❌ Hipótese 2

$$\theta_0 = 15, \theta_1 = 5, \hat{\mathbf{y}} = 5\mathbf{X}_1 + 15$$

$$J_{\text{SSR}} = (20-20)^2 + (10-25)^2 + (30-30)^2 = 225$$

✅ Hipótese 3

$$\theta_0 = 10, \theta_1 = 5, \hat{\mathbf{y}} = 5\mathbf{X}_1 + 10$$

$$J_{\text{SSR}} = (20-15)^2 + (10-20)^2 + (30-25)^2 = 150$$

Descida de Gradiente

- Algoritmo Iterativo
- $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla_J$
- $\boldsymbol{\theta}_0$ inicializado com valores aleatórios
- $\eta \in (0, 1]$ taxa de aprendizado
- Valores típicos para η são $10^{-3}, 10^{-4}$ ou 10^{-5}

Descida de Gradiente

- Ao lado temos um algoritmo para *Batch Gradient Descent*
- Se selecionarmos apenas uma instância $\mathbf{X}^{(i)}$ para estimar o valor do gradiente teremos o *Stochastic Gradient Descent*
- A diferença entre J_{SSR} de diferentes iterações pode ser usado como condição de convergência

Require: $\mathbf{X}_{n \times m}$ matriz de instâncias
Require: $\mathbf{y}_{n \times 1}$ variável alvo
Require: $\eta \in (0, 1]$ taxa de aprendizado

1: **procedure** GRADIENT DESCENT
2: $\quad \boldsymbol{\theta} \sim \mathcal{N}(0, 1)$
3: $\quad$ **repeat**
4: $\quad\quad \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \mathbf{X}^\top (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})$
5: $\quad\quad \hat{\mathbf{y}} \leftarrow \mathbf{X}\boldsymbol{\theta}$
6: $\quad\quad J_{\text{SSR}} \leftarrow (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}})$
7: $\quad$ **until** Condição de convergência
8: $\quad$ **return** $\boldsymbol{\theta}$
9: **end procedure**

Descida de Gradiente

Descida de Gradiente	Equação Normal
$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \mathbf{X}^\top (\mathbf{X}\boldsymbol{\theta}_t - \mathbf{y})$	$\boldsymbol{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$
solução iterativa	solução analítica
Possui o hiperparâmetro η	Sem hiperparâmetros
arbitrariamente próxima ao melhor $\boldsymbol{\theta}$	encontra exatamente o melhor $\boldsymbol{\theta}$
$O(i \cdot N \cdot M)$ em implementações vetoriais	$O(M^3)$

Como aplicar Regressão Linear para Classificação?

A *imagem* da Regressão Linear é incompatível com uma classificação

- $p\left(\hat{\mathbf{y}}^{(i)} = c|\mathbf{X}^{(i)}\right) \in [0, 1]$
- $\mathbf{X}^{(i)}\boldsymbol{\theta} \in (-\infty, +\infty)$

Precisamos encontrar algum valor relacionado a probabilidade de pertencimento a uma classe $p\left(\hat{\mathbf{y}}^{(i)} = c|\mathbf{X}^{(i)}\right)$ que tenha um contradomínio similar a Regressão Linear $\mathbf{X}^{(i)}\boldsymbol{\theta}$

Como aplicar Regressão Linear para Classificação?

- $p\left(\hat{\mathbf{y}}^{(i)} = c|\mathbf{X}^{(i)}\right) \in [0, 1]$: probabilidade de pertencimento a classe c
- $\frac{p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)}{1-p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)} \in [0, +\infty)$: chance de pertencimento a classe c
- $\ln\left(\frac{p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)}{1-p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)}\right) \in (-\infty, +\infty)$: log da chance de pertencimento a classe c:

Como aplicar Regressão Linear para Classificação?

- $p\left(\hat{\mathbf{y}}^{(i)} = c|\mathbf{X}^{(i)}\right) \in [0, 1]$: probabilidade de pertencimento a classe c
- $\frac{p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)}{1-p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)} \in [0, +\infty)$: chance de pertencimento a classe c
- $\ln\left(\frac{p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)}{1-p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)}\right) \in (-\infty, +\infty)$: log da chance de pertencimento a classe c:

Como aplicar Regressão Linear para Classificação?

- $p\left(\hat{\mathbf{y}}^{(i)} = c|\mathbf{X}^{(i)}\right) \in [0, 1]$: probabilidade de pertencimento a classe c
- $\frac{p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)}{1-p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)} \in [0, +\infty)$: chance de pertencimento a classe c
- $\ln\left(\frac{p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)}{1-p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)}\right) \in (-\infty, +\infty)$: log da chance de pertencimento a classe c:

Aplicando Regressão Linear para estimar o log das chances

$$\ln\left(\frac{p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)}{1-p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)}\right)=\mathbf{X}^{(i)}\boldsymbol{\theta}$$

$$\frac{p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)}{1-p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)}=e^{\mathbf{X}^{(i)}\boldsymbol{\theta}}$$

$$p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)=e^{\mathbf{X}^{(i)}\boldsymbol{\theta}}\left(1-p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)\right)$$

$$p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)+p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)e^{\mathbf{X}^{(i)}\boldsymbol{\theta}}=e^{\mathbf{X}^{(i)}\boldsymbol{\theta}}$$

$$p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)(1+e^{\mathbf{X}^{(i)}\boldsymbol{\theta}})=e^{\mathbf{X}^{(i)}\boldsymbol{\theta}}$$

$$p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)=\frac{e^{\mathbf{X}^{(i)}\boldsymbol{\theta}}}{(1+e^{\mathbf{X}^{(i)}\boldsymbol{\theta}})}=\frac{1}{e^{-\mathbf{X}^{(i)}\boldsymbol{\theta}}(1+e^{\mathbf{X}^{(i)}\boldsymbol{\theta}})}$$

$$p\left(\hat{\mathbf{y}}^{(i)}=c|\mathbf{X}^{(i)}\right)=\frac{1}{\left(1+e^{-\mathbf{X}^{(i)}\boldsymbol{\theta}}\right)}$$

Regressão Logística

- Ao adaptar a Regressão Linear para a tarefa de classificação binária obtemos a Regressão Logística
- $\hat{\mathbf{y}} = \frac{1}{\left(1+e^{-\mathbf{x}^{(i)}\boldsymbol{\theta}}\right)}$
- Precisamos de uma nova função de custo
  - J_{SSR} não é convexo para a Regressão Logística
  - $J_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\mathbf{y}^{(i)}\ln\hat{\mathbf{y}}^{(i)} + (1-\mathbf{y}^{(i)})\ln(1-\hat{\mathbf{y}}^{(i)})\right]$

Treinamento

Vamos derivar J_{BCE} em relação aos $\boldsymbol{\theta}$

$$
\begin{aligned}
\frac{\partial J_{\text{BCE}}}{\partial \boldsymbol{\theta}} &= \frac{\partial(-\mathbf{y}\ln\hat{\mathbf{y}} - (1-\mathbf{y})\ln(1-\hat{\mathbf{y}}))}{\partial \boldsymbol{\theta}} \\
&= -\frac{\partial(\mathbf{y}\ln\hat{\mathbf{y}})}{\partial \boldsymbol{\theta}} - \frac{\partial((1-\mathbf{y})\ln(1-\hat{\mathbf{y}}))}{\partial \boldsymbol{\theta}} \\
&= -\mathbf{y}\frac{\partial(\ln\hat{\mathbf{y}})}{\partial \boldsymbol{\theta}} - (1-\mathbf{y})\frac{\partial \ln(1-\hat{\mathbf{y}})}{\partial \boldsymbol{\theta}} \\
&= -\mathbf{y}\frac{\partial(\ln\hat{\mathbf{y}})}{\partial \hat{\mathbf{y}}}\frac{\partial \hat{\mathbf{y}}}{\partial \boldsymbol{\theta}} - (1-\mathbf{y})\frac{\partial \ln(1-\hat{\mathbf{y}})}{\partial \hat{\mathbf{y}}}\frac{\partial \hat{\mathbf{y}}}{\partial \boldsymbol{\theta}} \\
&= -\mathbf{X}^\top \mathbf{y}\frac{1}{\hat{\mathbf{y}}}\hat{\mathbf{y}}(1-\hat{\mathbf{y}}) - \mathbf{X}^\top(1-\mathbf{y})\left(-\frac{1}{1-\hat{\mathbf{y}}}\right)\hat{\mathbf{y}}(1-\hat{\mathbf{y}}) \\
&= \mathbf{X}^\top(-\mathbf{y} + \mathbf{y}\hat{\mathbf{y}} + \hat{\mathbf{y}} - \hat{\mathbf{y}}\mathbf{y}) \\
&= \mathbf{X}^\top(\hat{\mathbf{y}} - \mathbf{y})
\end{aligned}
$$

Regressão Logística

- Função de Custo J_{BCE} é convexa, mas não admite solução analítica
- O gradiente tem exatamente a mesma forma da Regressão Linear
- Gera uma fronteira de decisão linear (hiperplano) no espaço de atributos

Leituras Recomendadas

- Seções 3.1 a 3.3 de: Y.S. Abu-Mostafa, M. Magdon-Ismail e H.T. Lin.
Learning from Data: A Short Course. AMLBook.com, 2012. ISBN: 978-1-60049-006-4

Referências

[1] Y.S. Abu-Mostafa, M. Magdon-Ismail e H.T. Lin. Learning from Data: A Short Course. AMLBook.com, 2012. ISBN: 978-1-60049-006-4.

[2] U. Braga-Neto. Fundamentals of Pattern Recognition and Machine Learning. Springer International Publishing AG, 2020. ISBN: 978-3-030-27655-3.

[3] T.M. Mitchell. Machine Learning. McGraw-Hill International Editions. McGraw-Hill, 1997. ISBN: 978-0-07-115467-3.