Regularização

Regularização é inimiga do overfitting

Mas o que é a regularização?

Regularização ⟷ Aplicar "crenças" pessoais

Um exemplo de "crenças pessoais":

- uma moeda, vários lançamentos
- qual valor de p = Prob(dar "cara")

lançamento	p, só dados	p, com crença
—	indefinido	1/2
cara	1	≈ 1/2
cara	1	≈ 1/2
coroa	2/3	≈ 1/2
cara	3/4	≈ 1/2
coroa	3/5	≈ 1/2

matematicamente: H: nº caras (heads)

T: nº coroas (tails)

1. só dados: $p = \frac{H}{H+T}$

2. com crença a priori: $p = \frac{(H+\alpha)}{(H+\alpha)+(T+\alpha)}$ $\alpha > 0$

E como se antes de lançar a moeda, já tivéssemos $H = \alpha$, $T = \alpha$

Regularização em modelos lineares

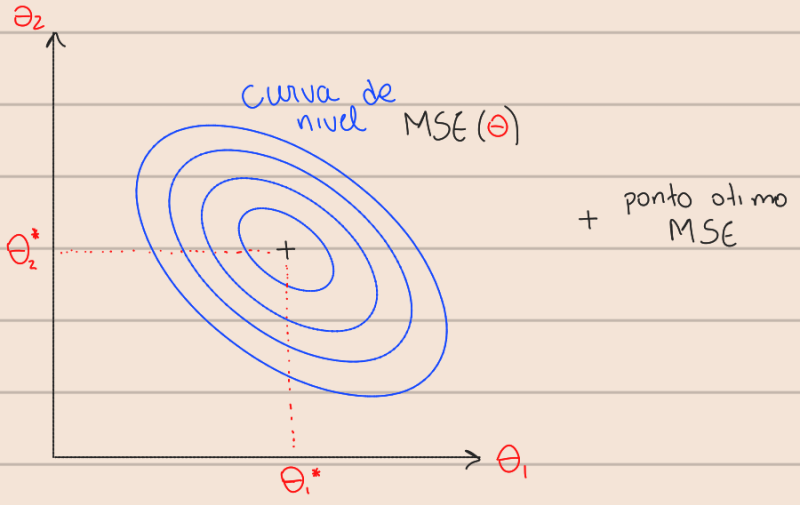
- sem regularização
- Ridge
- LASSO
- Elastic Net

* vol dados pequeno regularização ajuda

1) Sem regularização

$$L(\vec{\theta}) = MSE(\vec{\theta})$$

$\underset{\vec{\theta}}{\text{argmin}}\ L(\vec{\theta}) \rightarrow \vec{\theta}^{\text{ótimo}}_{\text{sem regularização}}$



2) Regularização Ridge

* combate a colinearidade → por causa da penalidade sobre o $\vec{\theta}^2$
* ñ busco pelo ótimo, mas que eles sejam pequenos
↳ vc quebra a possibilidade da minimização ser igual em vários casos

$$\vec{\theta}^{\text{ótimo}}_{\text{Ridge}} = \underset{\vec{\theta}}{\text{argmin}}\ MSE(\vec{\theta})$$

sujeito a $\|\vec{\theta}\|^2 \leq \beta$ $\Longleftrightarrow$ multiplicadores de Lagrange

$$\vec{\theta}^{\text{ótimo}}_{\text{Ridge}} = \underset{\vec{\theta}}{\text{argmin}}\ MSE(\vec{\theta}) + \alpha \|\vec{\theta}\|^2$$

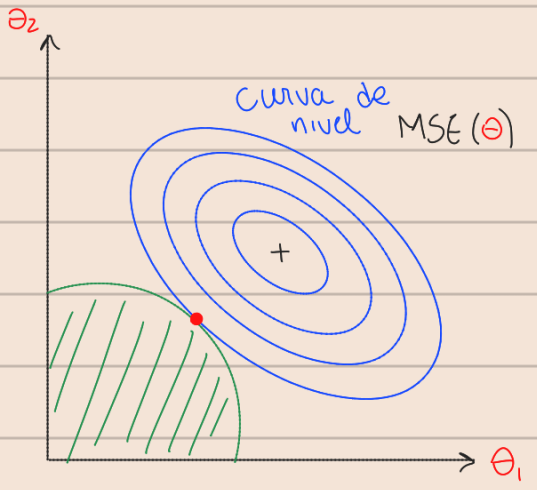
α: hiperparâmetro; $\|\vec{\theta}\|^2$: sem θ_0

normal $l_2 = \sqrt{\sum_{i=1}^{n} \|\theta_i\|^2}$

* antes de uma regularização Ridge usar Standard Scaler → normar X_1 (milhões) não X_n (micro) foi sentido

Soluções:
- gradient descent
- equação normal

$$\vec{\theta}^{\text{ótimo}}_{\text{Ridge}} = (X^T X + \alpha I)^{-1} X^T y$$



Vantagens do Ridge

· combate a colinearidade

ex: $X_1 = X_2$

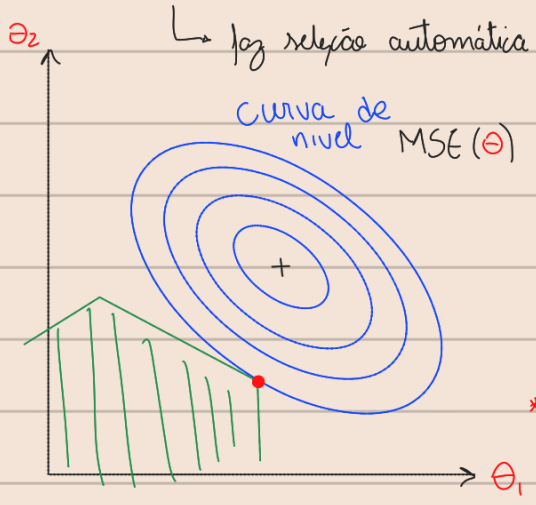
$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

1) Sem regularização, qualquer solução c/ $\theta_1 + \theta_2 = \theta_{12}^{\text{ótimo}}$ dá o mesmo modelo → ambíguo

2) Ridge: $\theta_1 + \theta_2 = \frac{1}{2}\theta_{12}$ é a melhor opção

3) LASSO (least absolute shrinkage and selection operator)

↳ faz seleção automática de features



+ ponto ótimo MSE

* aqui toca na aresta mas por causo de R^2 em graus maiores a chance de tocar um vértice é muito grande

$$\vec{\theta}_{\text{LASSO}}^{\text{ótimo}} = \underset{\vec{\theta}}{\operatorname{argmin}}\ MSE(\vec{\theta})$$

sujeito à $\|\vec{\theta}\|_1 \leq \beta$

norma $\ell_1 = \sum_{i=1}^{n} |\theta_i|$

$$\vec{\theta}_{\text{LASSO}}^{\text{ótimo}} = \underset{\vec{\theta}}{\operatorname{argmin}}\ MSE(\vec{\theta}) + \alpha \|\vec{\theta}\|_1$$

4) Reg. Elastic Net

$$\text{Loss} = MSE(\vec{\theta}) + \alpha \cdot (1-r) \cdot \|\vec{\theta}\|_2^2 + \alpha \cdot r \cdot \|\vec{\theta}\|_1$$

α: penalidade

r: balanceamento entre Ridge e LASSO

· conforma vantagem dos dois modelos