Exercício 1 — HRO na prática com `scikit-learn`

Objetivos de aprendizagem

- Identificar, em código Python, os três ingredientes do princípio HRO (hipótese, risco, otimização).
- Familiarizar-se com a API básica do `scikit-learn` para tarefas de regressão.

Nas aulas você viu o princípio **HRO** — um modelo de machine learning é determinado por três ingredientes: o espaço de hipóteses $\mathcal{H}$ (família de funções candidatas), uma função de risco $R_{\text{emp}}(\boldsymbol{\theta})$ (como medimos o erro) e um procedimento de otimização (como encontramos o melhor candidato). Neste exercício, você vai localizar esses três ingredientes dentro do `scikit-learn` olhando para um *regression tree*.

Instale/importe o `scikit-learn` e rode o seguinte código em um script ou notebook Python:

```
from sklearn.datasets import load_iris
from sklearn.tree import DecisionTreeRegressor

iris = load_iris()
print(type(iris))
print(iris.feature_names)
print(iris.target[:5])
print(iris.data.shape)

learner = DecisionTreeRegressor()
print(learner.get_params())
```

Com base no que você observar ao executar, responda:

(a) Quais atributos o objeto `iris` armazena? Descreva em uma ou duas frases o papel de cada um dos seguintes: `data`, `target`, `feature_names`.

(b) Cite **três** hiperparâmetros retornados por `learner.get_params()` e diga, em uma frase cada, o que eles controlam (pode consultar a documentação do `scikit-learn`).

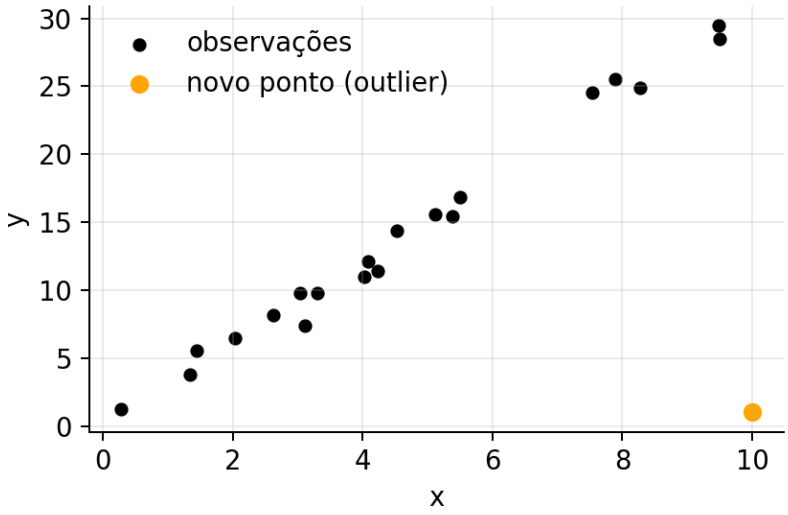
(c) Relacione cada um dos três ingredientes do HRO (espaço de hipóteses, risco e otimização) a algo que apareceu no código: qual objeto ou chamada corresponde a cada ingrediente?

Exercício 2 — Funções de loss em regressão

Objetivos de aprendizagem

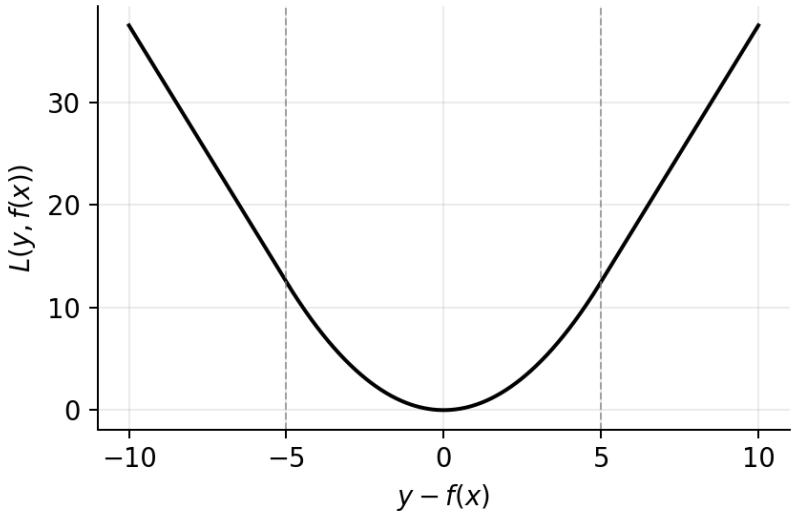
- Avaliar como outliers afetam modelos treinados com diferentes losses.
- Ler uma loss function a partir de sua representação visual.

Considere uma tarefa de regressão linear simples. A figura abaixo mostra 20 observações geradas em torno de uma reta, mais um novo ponto (em laranja) que claramente foge do padrão.



(a) Suponha que vamos ajustar uma regressão linear nesses dados minimizando o risco empírico. Como os parâmetros do modelo serão afetados ao incluir o ponto laranja se usarmos a **L2 loss**? E se usarmos a **L1 loss**? Argumente sem calcular valores — basta explicar a direção do efeito e por quê.

(b) A segunda figura mostra uma terceira loss bastante usada em regressão, a *Huber loss* (com parâmetro $\epsilon > 0$; aqui $\epsilon = 5$):



Descreva como a Huber loss trata os resíduos em comparação com L1 e L2. A partir do gráfico, proponha uma definição por partes para $L(y, f(x))$ em função do resíduo $r = y - f(x)$ e de ϵ.

Exercício 3 — Regressão polinomial

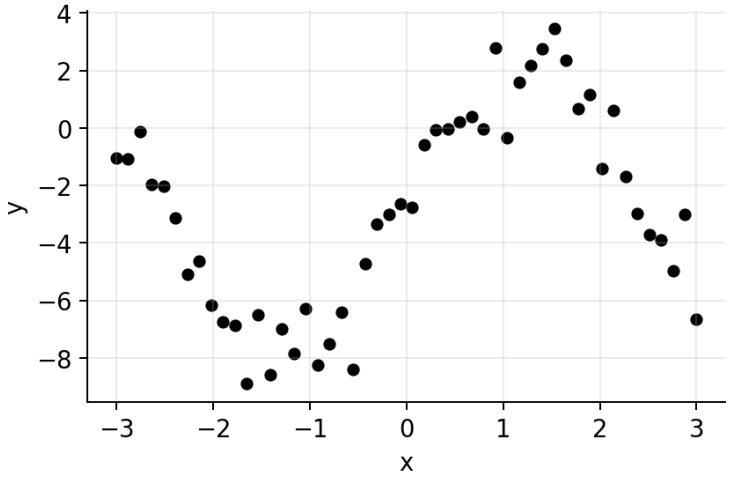
Objetivos de aprendizagem

- Expressar os componentes HRO de uma regressão polinomial.
- Derivar o gradiente do risco empírico e a atualização de gradiente descendente.
- Discutir o trade-off entre flexibilidade de modelo e generalização.

Temos um processo gerador de dados (com ruído):

$$y = -3 + 5 \cdot \sin(0.4\,\pi\,x) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 1),$$

do qual observamos $n = 50$ realizações, mostradas abaixo.



Resolvemos modelar esses dados por um polinômio cúbico (incluindo intercepto).

(a) Escreva o espaço de hipóteses $\mathcal{H}$ correspondente a esse polinômio cúbico.

(b) Escreva o risco empírico $R_{\text{emp}}(\boldsymbol{\theta})$ com respeito a $\boldsymbol{\theta}$ para um membro desse espaço de hipóteses, usando L2 loss. Seja o mais explícito possível.

(c) Podemos minimizar esse risco por gradiente descendente. Derive o gradiente $\nabla_{\boldsymbol{\theta}} R_{\text{emp}}(\boldsymbol{\theta})$. (Dica: é mais simples escrever o risco em forma matricial, $R_{\text{emp}}(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$, onde $\mathbf{X}$ é a matriz com colunas $1, x, x^2, x^3$.)

(d) Usando o gradiente do item anterior, descreva como atualizar o parâmetro atual $\boldsymbol{\theta}^{[t]}$ em um passo de gradiente descendente. Explique o que a fórmula significa intuitivamente.

(e) Com um polinômio cúbico dificilmente conseguiremos passar por todos os pontos. Discuta, em até dois parágrafos, as vantagens e desvantagens de escolher uma classe de modelos mais flexível. Você optaria por um modelo mais flexível nesse caso? Justifique.

Exercício 4 — Previsão no dataset `abalone`

Objetivos de aprendizagem

- Implementar um modelo de regressão em Python ponta a ponta.
- Analisar o ajuste básico de uma regressão linear.

Neste exercício você vai ajustar um modelo de regressão linear no dataset `abalone` do UCI Machine Learning Repository, para prever a idade (coluna `Rings`) de um molusco a partir do comprimento da concha mais longa (`LongestShell`) e do peso total (`WholeWeight`).

O arquivo é um CSV sem cabeçalho disponível em https://archive.ics.uci.edu/ml/machine-learning-datab abalone/abalone.data. Use o snippet abaixo para baixá-lo e colocá-lo em um `DataFrame`:

```
import pandas as pd

URL = "https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data"
cols = ["Sex", "LongestShell", "Diameter", "Height",
        "WholeWeight", "ShuckedWeight", "VisceraWeight",
        "ShellWeight", "Rings"]
df = pd.read_csv(URL, header=None, names=cols)
```

(a) Faça um scatter plot de `LongestShell` (eixo x) contra `WholeWeight` (eixo y), colorindo os pontos pela coluna `Rings`. Comente em uma ou duas frases o que o gráfico sugere sobre a relação entre as três variáveis.

(b) Usando `sklearn.linear_model.LinearRegression`, ajuste uma regressão linear que tenha `LongestShell` e `WholeWeight` como features e `Rings` como target. Reporte os coeficientes aprendidos e o intercepto.

(c) Para avaliar o ajuste, faça um scatter plot comparando os valores observados de `Rings` (y) e os valores previstos pelo modelo ($\hat{y}$). Inclua a reta diagonal $y = \hat{y}$ como referência. O que esse gráfico sugere sobre a qualidade do ajuste? A hipótese de linearidade parece apropriada?

(d) Calcule o **MAE** (*mean absolute error*) do modelo no próprio conjunto de treino usando `sklearn.metrics.mean_absolute_error`. Reporte o valor obtido e interprete-o na escala do problema (lembre-se de que `Rings` conta os anéis da concha, que aproximam a idade do animal em anos).

Dica

O MAE calculado aqui é o *erro de treino*. Ele subestima o erro que o modelo cometeria em dados novos — para uma avaliação honesta seria preciso separar um conjunto de teste, assunto das próximas listas.