Exercício 1 — O perceptron como modelo linear

Objetivos de aprendizagem

- Caracterizar o espaço de hipóteses de um neurônio único.
- Conectar o perceptron com identidade → regressão linear, e com sigmoid → regressão logística.
- Reconhecer que um neurônio só produz fronteiras lineares, motivando camadas ocultas.

Um **perceptron** (ou neurônio único) é a unidade computacional básica de uma rede neural. Dado um vetor de entrada $\mathbf{x} \in \mathbb{R}^p$, pesos $\mathbf{w} \in \mathbb{R}^p$, bias $b \in \mathbb{R}$ e uma função de ativação τ, o perceptron computa:

$$f(\mathbf{x}) \;=\; \tau\Big(\mathbf{w}^\top \mathbf{x} + b\Big) \;=\; \tau(w_1 x_1 + w_2 x_2 + \cdots + w_p x_p + b)\,.$$

(a) Escreva explicitamente o **espaço de hipóteses** $\mathcal{H}$ de um perceptron, em função dos parâmetros $\mathbf{w}, b$ e da função de ativação τ.

(b) Considere a escolha $\tau(z) = z$ (função identidade). Que modelo já visto no curso corresponde a esse perceptron? Qual a função de loss tipicamente usada para treiná-lo, e por quê?

(c) Considere agora a escolha $\tau(z) = s(z) = \frac{1}{1+e^{-z}}$ (função logística / sigmoid). Que modelo já visto no curso corresponde a esse perceptron? Qual a função de loss tipicamente usada para treiná-lo, e por quê?

(d) Mostre que, em ambos os casos (identidade e sigmoid), o **risco empírico** $\mathcal{R}_{\text{emp}}(\mathbf{w}, b)$ é uma função convexa de $(\mathbf{w}, b)$. *Dica:* cite o resultado padrão do curso ao invés de derivar de novo.

(e) Mostre que a **fronteira de decisão** de um perceptron com ativação sigmoid (i.e., o conjunto $\{\mathbf{x} : f(\mathbf{x}) = 0{,}5\}$) é **linear** em $\mathbf{x}$. Conclua: um único neurônio nunca consegue separar dados não-linearmente separáveis (e.g., dois círculos concêntricos), independentemente da sua ativação.

Exercício 2 — Forward pass à mão

Objetivos de aprendizagem

- Executar manualmente o forward pass de uma rede de camada oculta única.
- Aplicar ReLU e sigmoid componente a componente.
- Interpretar a saída de uma rede de classificação binária.

Considere uma rede neural com uma camada oculta de 3 neurônios e um neurônio de saída, treinada para classificação binária:

- Entrada: $\mathbf{x} \in \mathbb{R}^3$.
- Cada neurônio oculto $j \in \{1, 2, 3\}$ tem pesos $\mathbf{w}_j \in \mathbb{R}^3$ e bias b_j. Ativação da camada oculta: **ReLU**, $\sigma(v) = \max(0, v)$.
- Neurônio de saída tem pesos $\mathbf{u} \in \mathbb{R}^3$ e bias c. Ativação de saída: **sigmoid**, $\tau(v) = 1/(1 + e^{-v})$.

Os parâmetros aprendidos são:

$$\mathbf{w}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \quad \mathbf{w}_2 = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}, \quad \mathbf{w}_3 = \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ -2 \\ 0 \end{pmatrix},$$

$$\mathbf{u} = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}, \qquad c = -2.$$

Dada a entrada de teste $\mathbf{x}_* = (-1, 2, 1)^\top$:

(a) Calcule as três **pré-ativações** da camada oculta, $z_{\text{in},j} = \mathbf{w}_j^\top \mathbf{x}_* + b_j$ para $j = 1, 2, 3$. Apresente as três contas separadamente.

(b) Aplique a ativação ReLU em cada uma das pré-ativações e obtenha o vetor de **ativações** $\mathbf{z} = (z_1, z_2, z_3)^\top$ da camada oculta.

(c) Calcule a **pré-ativação de saída**, $f_{\text{in}} = \mathbf{u}^\top \mathbf{z} + c$.

(d) Aplique a sigmoid em f_{in} e reporte a saída final $f_{\text{out}} \in (0, 1)$. Como você interpreta esse valor no contexto da classificação binária? Qual classe a rede prevê para $\mathbf{x}_*$ se usarmos o threshold padrão de 0,5?

(e) **Mudança de ativação**: imagine que substituímos a ReLU da camada oculta pela função identidade ($\sigma(v) = v$), mas mantemos a sigmoid na saída. Refaça o cálculo de f_{in} para o mesmo $\mathbf{x}_*$. Compare com a sua resposta de (c). Por que os números são diferentes?

Exercício 3 — Por que precisamos de não-linearidade?

Objetivos de aprendizagem

- Demonstrar algebricamente que MLPs sem ativação não-linear colapsam a um modelo linear.
- Generalizar para redes com ℓ camadas ocultas.
- Justificar a centralidade da função de ativação (ReLU, sigmoid, tanh) no poder expressivo da rede.

Os slides afirmam que, sem ativação não-linear nas camadas ocultas, uma rede neural só consegue aprender fronteiras lineares — mesmo que tenha muitas camadas. Vamos demonstrar esse fato algebricamente.

Considere uma rede com **uma camada oculta** de m neurônios e **ativação identidade** ($\sigma(v) = v$), com saída escalar e ativação identidade também. Sejam:

- $W \in \mathbb{R}^{p \times m}$ a matriz de pesos da camada oculta (coluna j = pesos $\mathbf{w}_j$ do neurônio j);
- $\mathbf{b} \in \mathbb{R}^m$ o vetor de biases da camada oculta;
- $\mathbf{u} \in \mathbb{R}^m$ os pesos do neurônio de saída e $c \in \mathbb{R}$ seu bias.

(a) Escreva a saída $f(\mathbf{x})$ da rede como uma sequência de duas operações: primeiro a camada oculta produz $\mathbf{z} = W^\top \mathbf{x} + \mathbf{b}$, e em seguida o neurônio de saída produz $f(\mathbf{x}) = \mathbf{u}^\top \mathbf{z} + c$.

(b) Substitua a expressão de $\mathbf{z}$ no $f(\mathbf{x})$ e **simplifique algebricamente**, mostrando que f pode ser escrita na forma

$$f(\mathbf{x}) = \tilde{\mathbf{w}}^\top \mathbf{x} + \tilde{b},$$

com pesos efetivos $\tilde{\mathbf{w}}$ e bias efetivo $\tilde{b}$ explícitos. Dê as expressões fechadas para $\tilde{\mathbf{w}}$ e $\tilde{b}$ em termos de $W, \mathbf{b}, \mathbf{u}, c$.

(c) Generalize: considere uma rede com ℓ camadas ocultas, todas com ativação identidade. As camadas têm parâmetros $W^{(1)}, \mathbf{b}^{(1)}, W^{(2)}, \mathbf{b}^{(2)}, \ldots, W^{(\ell)}, \mathbf{b}^{(\ell)}$, e a saída final tem $\mathbf{u}, c$. Mostre que a função implementada pela rede ainda é da forma $f(\mathbf{x}) = \tilde{\mathbf{w}}^\top \mathbf{x} + \tilde{b}$, e dê a expressão de $\tilde{\mathbf{w}}$ como um produto matricial.

(d) Conclua, em até dois parágrafos, qual é o conteúdo prático desse resultado: *por que adicionar camadas a uma rede sem não-linearidades não acrescenta poder expressivo*? Conecte com a discussão do Exercício 1 sobre por que um único neurônio (modelo linear) falha em problemas como os círculos concêntricos. Por que, então, a escolha da ativação — ReLU, sigmoid, tanh — é tão central?

Exercício 4 — Representação e profundidade

Objetivos de aprendizagem

- Discutir representation learning vs. feature engineering manual.
- Entender, qualitativamente, como camadas ocultas resolvem problemas não-linearmente separáveis.
- Reconhecer o papel da composição de camadas em aumentar a expressividade.

Os exercícios anteriores cobriram a mecânica das redes neurais: um neurônio só vê fronteiras lineares (Ex 1), e redes sem ativação não-linear colapsam a modelos lineares (Ex 3). Este exercício é **conceitual**, e cobre os "porquês" que os slides discutem nas seções de introdução e motivação.

Responda cada item em até dois parágrafos.

(a) Representation learning vs. feature engineering. Antes de 2010, quando se queria treinar um classificador para imagens, era comum gastar semanas construindo *features* à mão (detectores de borda, descritores de textura, histogramas de cor). O que mudou com deep learning? Em outras palavras: o que significa dizer que uma rede neural "aprende a representação dos dados"? Por que essa automatização tende a ser uma vantagem em domínios como visão e linguagem?

(b) O exemplo dos círculos concêntricos. Considere o problema de classificar pontos do plano em duas classes: uma classe ocupa o disco $\{x_1^2 + x_2^2 \leq 1\}$ e a outra ocupa a coroa $\{1 < x_1^2 + x_2^2 \leq 4\}$. Como discutido no Exercício 1, um único neurônio (linear) não consegue separar essas classes. Descreva qualitativamente: como uma camada oculta com não-linearidade poderia resolver o problema? *Dica:* pense em o que cada neurônio oculto pode aprender a calcular — não precisa derivar pesos exatos, apenas discutir o tipo de transformação.

(c) Por que profundidade ajuda? Suponha que você tenha um orçamento fixo de, digamos, 1,000 parâmetros aprendíveis. Por que uma rede com vários poucos neurônios espalhados por várias camadas (rede "profunda") tende a ser mais expressiva do que uma rede com muitos neurônios numa única camada (rede "rasa")? Conecte sua resposta com a ideia de *representações cada vez mais abstratas* construídas camada a camada (slide do classificador do pássaro *kingfisher*).