Computação Científica e Análise de Dados
Edição de Compartilhamento

João Victor Lopez Pereira
João Antonio Recio Paixão

Universidade Federal do Rio de Janeiro
Centro de Ciências da Matemática e da Natureza
Instituto de Computação

Rio de Janeiro - RJ
29 de abril de 2025

Atenção: Esta versão do documento está sendo disponibilizada pelo *Google Classroom* e pelo *Google Drive* para facilitar o acesso dos estudantes. Recomenda-se verificar regularmente se há uma versão mais atualizada disponível, garantindo assim um conteúdo mais completo e com menos inconsistências. As atualizações do material não serão necessariamente notificadas. Diferentes partes do documento estão em constante evolução e melhoria. Sinta-se a vontade para entrar em contato com João Victor Lopez Pereira para qualquer sugestão ou dúvida em relação ao conteúdo aqui presente.

Sumário

Lista de Figuras

Lista de Teoremas

Nota dos Autores

Esse documento contém o conteúdo da disciplina *Computação Científica e Análise de Dados*, conhecida informalmente como *CoCADa*, oferecida na Universidade Federal do Rio de Janeiro (UFRJ) como pré-requisito para a formação em Bacharelado em Ciência da Computação. O objetivo desse documento é fornecer um material de apoio para os estudantes da disciplina e interessados na área. O conteúdo aqui contido pode conter erros ou estar desatualizado. Caso encontre alguma imprecisão, incoerência ou tenha sugestões, sinta-se à vontade para entrar em contato com João Victor Lopez Pereira pelo *e-mail* `joaovlp@ic.ufrj.br`, para que assim, possamos corrigir rapidamente o problema e disponibilizar uma versão aprimorada.

Para compreender plenamente os conteúdos abordados neste documento, é essencial que o leitor tenha conhecimento prévio de cálculo e, sobretudo, de Álgebra Linear. Embora diversos teoremas sejam demonstrados ao longo do texto e nos apêndices, pressupomos familiaridade com a transição entre sistemas de equações e sua representação matricial, tratando esse processo de forma natural. Além disso, assumimos que o leitor já esteja confortável com definições, propriedades e nuances fundamentais da Álgebra Linear.

O documento está dividido em duas grandes partes. A primeira delas, nomeada *Problema, Motivação e Modelagem*, aborda a modelagem de 20 diferentes problemas que, no fim, resultam em 4 diferentes equações na qual sua solução — dada na segunda parte, nomeada *Truque, Troca de Variáveis e Fatoração* — resolve o problema inicial. Os 4 capítulos de cada parte e o respectivo problema resolvido é:

- Modelos Estáticos: encontrar x tal que $Ax = b$;
- Modelos Aproximados: encontrar x tal que $x \in \text{argmin}_x \ \|Ax - b\|$;
- Modelos Dinâmicos: encontrar x_k tal que $x_k = A^k x_0$;
- Modelos de Reconhecimento de Padrões: encontrar x tal que $x \in \text{argmax}_x \ \dfrac{\|Ax\|}{\|x\|}$.

Além disso, a forma como esses problemas são resolvidos nesse material é a partir de 4 diferentes fatorações provindas da Algebra Linear, sendo elas:

- Decomposição LU: $A = LU$;
- Decomposição QR: $A = QR$;
- Diagonalização: $A = VDV^{-1}$;

- Decomposição em Valores Singulares: $A = U\Sigma V^{\mathsf{T}}$.

A partir dessas *fatorações*, temos que os problemas se resumem a realizar uma *troca de variáveis* e utilizar *truques* simples de Álgebra Linear e Cálculo para serem resolvidos.

Este documento foi escrito utilizando o sistema de *typesetting* LaTeX. Grande parte dos gráficos e visualizações presentes no material foram gerados com a linguagem Julia.

Agradecimentos e Reconhecimentos

Agradecemos profundamente a Gustavo de Mendonça Freire por ter nos ajudado com a demonstração dos teoremas T1, T2, T18, T23, T29, T37, T38 e, principalmente, pelo teorema de sua *possível* autoria T10. Sua contribuição, mesmo que indireta, foi fundamental para garantir maior precisão matemática e clareza nas intuições sobre o significado das soluções apresentadas para os problemas aqui discutidos.

Tabelas

	Modelagem Resultante	Fatoração	Troca de Variáveis	Truque
Exatos	$Ax = b$	$A = LU$	$Ux = L^{-1}b$	Subst. Rev.
Aproximados	$\text{argmin}_x \ \|Ax - b\|$	$A = QR$	$\text{argmin}_x \ \left\|Rx - Q^{\mathsf{T}}b\right\|$	Subst. Rev.
Dinâmicos	$x_k = A^k x_0$	$A = VDV^{-1}$	$\bar{x}_k = D^k \bar{x}_0$	Recorrências
Padrões	$\text{argmax}_x \ \frac{\|Ax\|^2}{\|x\|^2}$	$A = U\Sigma V^{\mathsf{T}}$	$V\text{argmax}_x \ \frac{\|\Sigma x\|^2}{\|x\|^2}$	Maximização

Parte 1

Problema, Motivação e Modelagem

The purpose of abstraction is not to be vague, but to create a new semantic level in which one can be absolutely precise.

Edsger W. Dijkstra, *The Humble Programmer*

Capítulo 1.1

Modelagens Exatas

1.1.1 Interpolação Polinomial (Interpolação)

Um matemático foi desafiado a: dado um conjunto de n pontos (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) em $\mathbb{R}^2$, encontrar uma função polinomial $f : \mathbb{R} \to \mathbb{R}$ tal que, $\forall i \in \{1, \ldots, n\}$, $f(x_i) = y_i$. Em outras palavras, encontrar uma função que passe por todos os pontos de um conjunto.

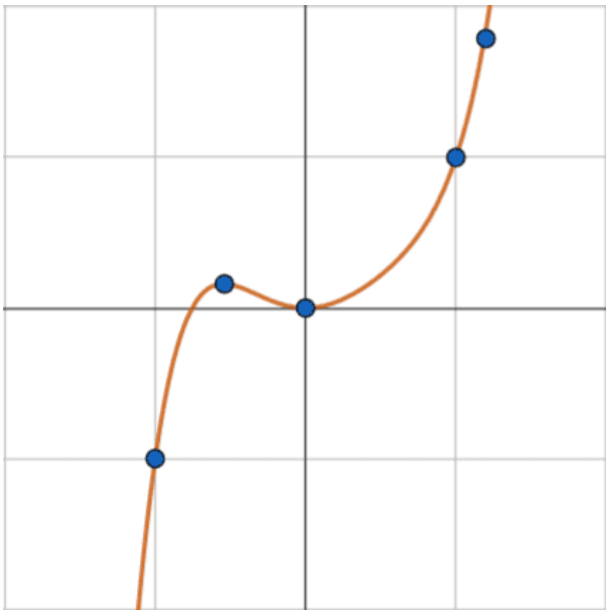


Figura F1: Exemplo de função resultante do processo de interpolação dados 5 pontos.

O processo de interpolação consiste em determinar uma função que passe exatamente por um determinado conjunto de pontos para aproximar, modelar ou prever o comportamento de um fenômeno.

Perceba que — para garantir a existência de uma função polinomial e, portanto, garantir que a interpolação polinomial seja possível — é preciso que necessariamente os pontos apresentem coordenada x diferentes visto que não é possível encontrar uma função polinomial que tenha 2

valores y associados a um mesmo x.[1]

Para que, dados n pontos, tenhamos a garantia de que a função passe por todos eles, é preciso necessariamente que a função polinomial tenha grau $n-1$ e, consequentemente, tenha n coeficientes.[2] Assim, temos que:

$$f(x) = c_0x^0 + c_1x^1 + \cdots + c_{n-1}x^{n-1}.$$

Como $f(x_i) = y_i$, temos que:

$$\begin{aligned}
y_1 &= c_0x_1^0 + \cdots + c_jx_1^j + \cdots + c_{n-1}x_1^{n-1},\\
&\vdots\\
y_i &= c_0x_i^0 + \cdots + c_jx_i^j + \cdots + c_{n-1}x_i^{n-1},\\
&\vdots\\
y_n &= c_0x_n^0 + \cdots + c_jx_n^j + \cdots + c_{n-1}x_n^{n-1},
\end{aligned}$$

$\forall i \in \{1, \ldots, n\}$, $j \in \{0, \ldots, n-1\}$.

Perceba que esse sistema de equações pode ser reescrito como sendo um produto matriz-vetor:

$$\underbrace{\begin{bmatrix}
x_1^0 & \ldots & x_1^j & \ldots & x_1^{n-1}\\
\vdots & \ddots & \vdots & \iddots & \vdots\\
x_i^0 & \ldots & x_i^j & \ldots & x_i^{n-1}\\
\vdots & \iddots & \vdots & \ddots & \vdots\\
x_n^0 & \ldots & x_n^j & \ldots & x_n^{n-1}
\end{bmatrix}}_{A}
\underbrace{\begin{bmatrix} c_0\\ \vdots\\ c_j\\ \vdots\\ c_{n-1} \end{bmatrix}}_{x}
=
\underbrace{\begin{bmatrix} y_1\\ \vdots\\ y_i\\ \vdots\\ y_n \end{bmatrix}}_{b}$$

E, dessa forma, para encontrar a função polinomial, nos resta encontrar os coeficientes da função f que compõem o vetor x tal que $Ax = b$.

Exercício E1:

Monte o sistema de equações e, em seguida, o sistema matriz-vetor no formato $Ax = b$ para o

[1] Visualmente, sabemos que não existe uma função que passe por 2 valores diferentes de y para um mesmo x. Ainda assim, é legal que esta nuance possa ser observada na própria fatoração que resolve o problema visto que 2 colunas da matriz de Vandermonde (matriz A) apresentarem os mesmos valores de x faz com que as colunas não sejam linearmente independentes e, consequentemente, a matriz não tenha posto completo, fazendo assim com que o sistema $Ax = b$ de fato não tenha solução. Mais detalhes sobre a fatoração podem ser encontrados na seção 2.1.3.

[2] Não estamos dizendo que a função resultante terá necessariamente grau $n-1$, pois os n pontos podem, por exemplo, estar alinhados, permitindo uma interpolação com uma função polinomial de grau menor. No entanto, em geral, para garantir a interpolação de n pontos distintos, consideramos uma função polinomial de grau no máximo $n-1$, pois assim asseguramos sua existência.

processo de interpolação dos pontos $(1,1)$, $(2,3)$, $(3,2)$, $(4,1)$ e $(5,0)$ por uma função polinomial de grau 4.

Exercício E2:

Monte o sistema de equações para o processo de interpolação dos pontos $(1,1)$, $(1,2)$ e $(2,3)$ para uma função polinomial f de grau 2. Tente resolver o sistema pelo método da substituição. O que pode ser concluído sobre a existência de solução? Como a escolha dos pontos reflete na existência de uma solução?

Exercício E3:

Explique o que o determinante de uma matriz A nos diz sobre a existência e a unicidade de soluções de um sistema linear $Ax = b$.

1.1.2 Barra de Calor (Diferenças Finitas)

Um engenheiro responsável pela construção de um prédio decidiu colocar uma barra de metal em um dos pisos mais instáveis para garantir maior estabilidade à construção. Essa barra passa por dentro de uma parede e apresenta ambos seus extremos como suas duas únicas regiões visíveis e que têm contato com o meio externo. O engenheiro deseja calcular a temperatura dos pontos internos da barra para garantir que o comprimento adicional obtido pela dilatação térmica do metal não danifique a construção.

Figura F2: Representação discretizada da barra.

O engenheiro mede a temperatura de ambos os extremos da barra, obtendo assim valores t_1 e t_n. Ele discretiza a barra em n nós equidistantes (incluindo os nós externos), conforme indicado na figura F2, e decide que uma boa aproximação para a temperatura de um dado ponto é a média da temperatura dos pontos vizinhos.[3] Em outras palavras:

$$t_i = \frac{t_{i-1} + t_{i+1}}{2},$$

$\forall i \in \{2, \ldots, n-1\}$, sendo t_i a temperatura do nó i.

[3]Tomar a temperatura em um ponto como sendo a média dos pontos adjacentes vem da aproximação da segunda derivada da equação do calor estacionária. A equação é $\frac{d^2t}{dx^2} = 0$, que pode ser aproximado por *diferenças finitas* como sendo $\frac{t_{i-1} - 2t_i + t_{i+1}}{\Delta x^2} = 0$, que, finalmente, resulta na equação $t_i = \frac{t_{i-1} + t_{i+1}}{2}$.

Essa equação pode ser reescrita como:

$$2t_i - t_{i-1} - t_{i+1} = 0.$$

Veja que, como essa equação vale $\forall i \in \{2, \ldots, n-1\}$, então temos o sistema:

$$\begin{aligned} 2t_2 - t_1 - t_3 &= 0, \\ &\vdots \\ 2t_i - t_{i-1} - t_{i+1} &= 0, \\ &\vdots \\ 2t_{n-1} - t_{n-2} - t_n &= 0. \end{aligned}$$

Mas como sabemos os valores de t_1 e t_n, então temos:

$$\begin{aligned} 2t_2 - t_3 &= t_1, \\ &\vdots \\ 2t_i - t_{i-1} - t_{i+1} &= 0, \\ &\vdots \\ 2t_{n-1} - t_{n-2} &= t_n, \end{aligned}$$

que pode ser reescrito na forma matriz-vetor como:

$$\underbrace{\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & & & \\ & & \ddots & & \\ & & & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} t_2 \\ \vdots \\ t_i \\ \vdots \\ t_{n-1} \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} t_1 \\ \vdots \\ 0 \\ \vdots \\ t_n \end{bmatrix}}_{b}$$

Dessa forma, temos que: descobrir as temperaturas internas da barra se resume em encontrar o vetor x que satisfaz $Ax = b$.

Exercício E4:

Dado que a temperatura nos nós externos t_1 e t_4 são 10 e 40, respectivamente, monte o sistema de equações e, em seguida, encontre o valor da temperatura nos nós internos t_2 e t_3 usando o método da substituição.

Exercício E5:

Dado $t_1 = 0$ e $t_5 = 100$, monte o sistema matriz-vetor $Ax = b$ onde $x = \begin{bmatrix} t_2 & t_3 & t_4 \end{bmatrix}^\mathsf{T}$.

1.1.3 Temperatura do Lago (Diferenças Finitas)

Um pesquisador tem como objetivo estudar o comportamento de um determinado lago em uma época específica do ano em que sua temperatura varia conforme o lugar por conta da incidência de raios solares. O problema é que o pesquisador não pode acessar as zonas internas do lago pois a presença de seu corpo ou de um veículo pode perturbar o equilíbrio térmico e ocasionar resultados enviesados em sua pesquisa. Dessa maneira, o pesquisador só tem acesso à zona costeira do lago, onde a temperatura pode ser facilmente medida por equipamentos pequenos o suficiente para não interferir no equilíbrio térmico da água. O pesquisador, então, tem a capacidade de mentir a temperatura das bordas do lago e decide que, de alguma maneira, deve calcular a temperatura interna do lago apenas a partir dos dados que tem acesso.

Para isso, o pesquisador decide primeiro discretizar o lago de tal forma que, a cada poucos metros caminhados pela borda, ele mede a temperatura e a anota em sua prancheta. Além disso, ele também faz um pequeno esboço do caminho que faz até que tenha dado uma volta inteira pelo lago. No fim, o pesquisador um desenho semelhante ao da figura F3.[4]

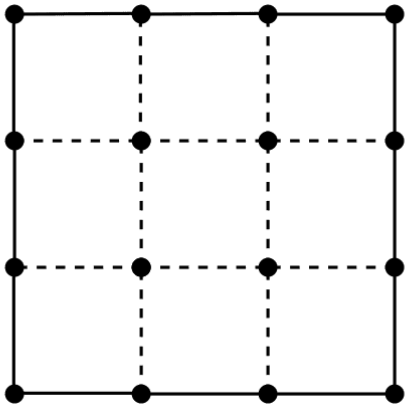


Figura F3: Representação discretizada do lago.

O pesquisador decide que uma boa aproximação para cada ponto interno do lago no qual ele não tem acesso à temperatura é a média dos 4 pontos em seu entorno.[5] Como cada ponto interno

[4]Perceba que não é necessário que o domínio (abstração do lago) apresente um formato geral quadrado ou retangular, apesar de que, com a fórmula de aproximação utilizada, é importante que a malha seja quadrada.

[5]Essa aproximação, de maneira bem semelhante à aproximação realizada em 1.1.2, vem da aproximação da segunda derivada da equação do calor bidimensional. A equação é

$$\frac{\partial^2 t}{\partial x^2} + \frac{\partial^2 t}{\partial y^2} = 0,$$

que pode ser aproximada por *diferenças finitas* como sendo

$$\frac{t_{i-1,j} - 2t_{i,j} + t_{i+1,j}}{\Delta x^2} + \frac{t_{i,j-1} - 2t_{i,j} + t_{i,j+1}}{\Delta y^2} = 0,$$

tem exatamente 4 pontos vizinhos, podemos dizer que a fórmula seguida pelo pesquisador é:

$$t_a = \frac{t_b + t_c + t_d + t_e}{4},$$

sendo t_a a temperatura do ponto de índice i para todo i internos à malha e t_b, t_c, t_d e t_e as temperaturas relativas aos pontos adjacentes.

Dessa forma, temos que, se o lago for discretizado em n pontos internos, temos que o sistema de equações

$$t_a = \frac{t_b + t_c + t_d + t_e}{4}, \quad \forall a \in \{1, \ldots, n\}$$

é válido. Além disso, veja que esse sistema pode ser reescrito como:

$$4t_a - t_b - t_c - t_d - t_e = 0, \quad \forall a \in \{1, \ldots, n\}.$$

Ao colocarmos as temperaturas t_a em um vetor x, podemos interpretar esse sistema de equações como um produto matriz-vetor

$$\underbrace{\begin{bmatrix} a_{11} & \ldots & a_{1i} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ a_{i1} & \ldots & a_{ii} & \ldots & a_{in} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{ni} & \ldots & a_{nn} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} t_1 \\ \vdots \\ t_i \\ \vdots \\ t_n \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} b_1 \\ \vdots \\ b_i \\ \vdots \\ b_n \end{bmatrix}}_{b}$$

onde A é a matriz que expressa a relação entre cada nó interno, x é o vetor das temperaturas dos nós internos e b é o vetor com as informações da condição de contorno e/ou 0, a depender do quão discretizado o lago foi. Nesse caso, os valores que os elementos a_{ii} da matriz A podem assumir só pode ser 4, -1 ou 0, visto que esses são os possíveis valores que multiplicam cada temperatura t_i na equação que antecipou o sistema na forma matriz-vetor.

Dessa forma, temos que encontrar o valor da temperatura dos nós internos do lago se resume em resolver o sistema linear $Ax = b$.

Exercício E6:

Desenho o lado e em seguida monte o sistema de equações e o sistema matriz-vetor $Ax = b$ para

---

que, se a malha for uniforme, resulta na equação

$$t_{ij} = \frac{t_{i-1,j} + t_{i+1,j} + t_{i,j-1} + t_{i,j+1}}{4}.$$

um lago (3×4) onde $t_{11} = t_{12} = t_{13} = t_{14} = 25$, $t_{31} = t_{32} = t_{33} = t_{34} = 15$ e $t_{21} = t_{24} = 20$, com $x = \begin{bmatrix} t_{22} & t_{23} \end{bmatrix}^{\mathsf{T}}$.

1.1.4 Nutrição

Um nutricionista está montando uma dieta para um de seus pacientes que envolve uma quantidade específica de diferentes macronutrientes. O paciente o informou que contém 2 tipos diferentes de cereais que apresentam concentrações diferentes de macronutrientes e calculou que, misturando esses dois cereais, consegue atingir a quantidade indicada pelo nutricionista. Porém, o paciente não informou ao nutricionista qual deve ser a quantidade de cada cereal que deve ser misturada para atingir a quantidade esperada de cada macronutriente.

Considerando que cada macronutriente seja independente entre si, temos que o primeiro cereal pode ser representado por um vetor em que cada coordenada simbolize o quanto daquele macronutriente aquele cereal possui. Considerando que o paciente pode variar na quantidade desse cereal (vetor), então temos uma espécie de reta que expressa a relação entre cada posição desse vetor (que é fixa).

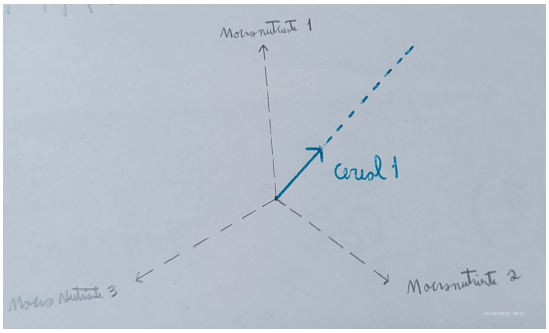


Figura F4: Representação do cereal 1 em $\mathbb{R}^3$.

Ao considerarmos o segundo cereal, visto que o paciente disse que suas concentrações são diferentes,[6] temos mais um vetor nesse espaço e temos então que a concentração almejada pelo nutricionista deve necessariamente estar no subespaço gerado pela combinação linear desses vetores (ou seja, deve ser mistura dos cereais).

[6]Caso as concentrações fossem iguais (ou múltiplas uma da outra), os vetores que representam esses cereais não seriam linearmente independentes e, consequentemente, o segundo cereal não ajudaria de forma alguma a alcançar a concentração almejada pelo nutricionista.

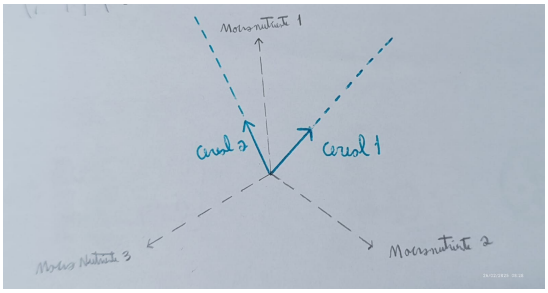


Figura F5: Representação dos cereais 1 e 2 em $\mathbb{R}^3$.

Dessa forma, temos então que, dados n macronutrientes:

$$c_1 = \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix} \quad \text{e} \quad c_2 = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix},$$

$\forall i \in \{1, \ldots, n\}$.

Como queremos uma combinação linear desses vetores, então temos um sistema:

$$q_1 \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix} + q_2 \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} z_1 \\ \vdots \\ z_i \\ \vdots \\ z_n \end{bmatrix},$$

que pode ser reescrito como

$$\underbrace{\begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_i & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} q_1 \\ q_2 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} z_1 \\ \vdots \\ z_i \\ \vdots \\ z_n \end{bmatrix}}_{b}.$$

Dessa forma, queremos descobrir o vetor x tal que $Ax = b$. Além disso, veja que possivelmente o vetor x encontrado apresentará coordenadas negativas, o que significa que de fato existe uma combinação linear dos cereais 1 e 2 que apresentam a concentração recomendada pelo nutricionista. Em outras palavras, b é um vetor que reside no plano gerado pelos vetores coluna de A.

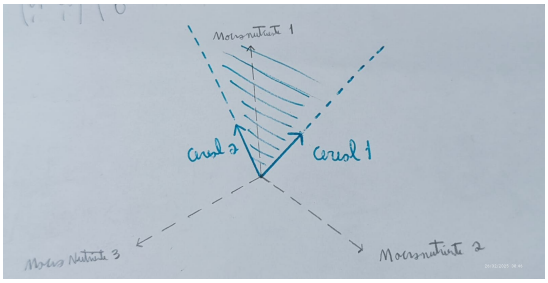


Figura F6: Representação do espaço gerado pela combinação linear com coeficientes positivos dos cereais 1 e 2 em $\mathbb{R}^3$.

Ainda assim, veja que, pela natureza do problema, não faz sentido x apresentar coordenadas negativas visto que não é possível colocar uma quantidade negativa de cereal em uma combinação de dois cereais. Dessa forma, dizemos que o cereal almejado só é obtido se existirem q_1 e q_2 não negativos tal que $Ax = b$.

1.1.5 Tomografia

Uma equipe composta por médicos e cientistas da computação busca desenvolver um equipamento capaz de calcular a densidade de diferentes partes do corpo humano. Nesse sistema, o paciente deita-se em uma esteira que o transporta por uma máquina equipada com tecnologia de raios-X, gerando imagens de alta precisão dos órgãos internos. Conforme a esteira avança, camadas finas do corpo são escaneadas, e o desafio do equipamento é determinar a densidade de cada região, levando em conta a composição dos diversos tecidos presentes.

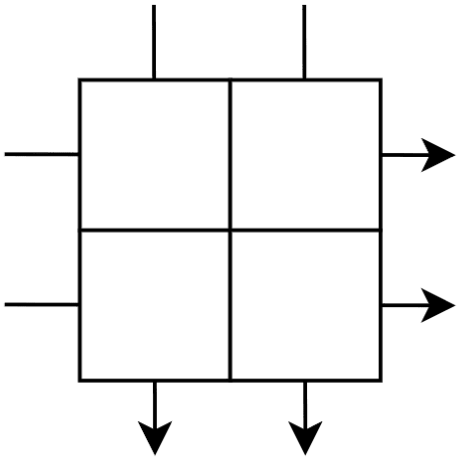


Figura F7: Representação abstrata do corpo de um paciente.

Devido à sobreposição de diferentes camadas de tecido, um único feixe de raio-X não é suficiente para determinar a densidade de uma região específica do corpo. Para contornar esse problema, o corpo do paciente é modelado como uma coleção de pequenas parcelas, e múltiplos feixes de raios-X são projetados sob diferentes ângulos. Com um número suficiente de medições, é possível reconstruir matematicamente a densidade de cada parcela individual, garantindo uma

análise precisa da estrutura interna do corpo.

Assumindo que o paciente tenha sido modelado de maneira semelhante à abstração da figura F7, chamamos cada uma das parcelas de seu corpo de x_i, $\forall i \in \{1, \ldots, n\}$, sendo n a quantidade de parcelas. Pelo resultado da emissão dos feixes de raios-X, temos alguns valores que representam a soma de diferentes tecidos.

Dessa forma, temos um sistema de equações onde cada equação representa uma combinação linear de x_i, $\forall i \in \{1, \ldots, n\}$, com coeficientes 0 ou 1. Em outras palavras, temos que, para cada equação, a densidade de uma região específica pode contribuir para determinada equação ou não, mas nunca contribuir parcialmente. Além disso, cada uma dessas combinações lineares resultará nos valores encontrados pela emissão de raio-X.

Dessa forma, temos um sistema $Ax = b$, em que A é a matriz formada composta pelos elementos 0 e 1, x é o vetor com as variáveis x_i e b é o vetor com o resultado das emissões de raio-X.

1.1.6 Sites por Importância (Ranking)

Uma equipe de desenvolvedores está criando um novo navegador que oferece ao usuário diversas funcionalidades. Um dos desenvolvedores desse grupo, responsável por fazer a ferramenta de busca, deseja ordenar os sites que aparecem pro usuário, ao realizar uma busca, por ordem de importância baseado em algum critério bem definido.

O desenvolvedor decide que para definir a importância de um site deve considerar a importância dos sites que o referenciam. Além disso, o desenvolvedor decide que um site que referência a muitos outros site deve ter sua importância distribuída dado que quanto mais sites são referenciados, menos importante cada um é.

Dessa maneira, dado n sites s_i, $i \in \{1, \ldots, n\}$, definimos que a importância do site s_i é

$$s_i = \sum_{j \in A_i} \frac{s_j}{q_j},$$

onde A_i é o conjunto dos sites que referenciam o site s_i e q_j a quantidade de sites que elemento desse conjunto referencia.

Por essa equação ser válida $\forall i \in \{1, \ldots, n\}$, temos:

$$
\begin{aligned}
s_1 &= \sum_{j \in A_1} \frac{s_j}{q_j}, \\
&\vdots \\
s_i &= \sum_{j \in A_i} \frac{s_j}{q_j}, \\
&\vdots \\
s_n &= \sum_{j \in A_n} \frac{s_j}{q_j},
\end{aligned}
$$

que, ao subtrair s_i de ambos os lados, pode ser reescrito como

$$
\begin{aligned}
-s_1 + \sum_{j \in A_1} \frac{s_j}{q_j} &= 0, \\
&\vdots \\
-s_i + \sum_{j \in A_i} \frac{s_j}{q_j} &= 0, \\
&\vdots \\
-s_n + \sum_{j \in A_n} \frac{s_j}{q_j} &= 0,
\end{aligned}
$$

que pode ser reescrito na forma matriz-vetor como

$$
\underbrace{\begin{bmatrix}
-1 & \dots & a_{1i} & \dots & a_{1n} \\
\vdots & \ddots & \vdots & \iddots & \vdots \\
a_{i1} & \dots & -1 & \dots & a_{in} \\
\vdots & \iddots & \vdots & \ddots & \vdots \\
a_{n1} & \dots & a_{ni} & \dots & -1
\end{bmatrix}}_{A}
\underbrace{\begin{bmatrix} s_1 \\ \vdots \\ s_i \\ \vdots \\ s_n \end{bmatrix}}_{x}
=
\underbrace{\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{b},
$$

onde A é a matriz que contém os coeficientes das equações que descrevem a importância de um dado filme, x é o vetor com a importância de cada filme e b é um vetor com zeros.

Dessa forma, temos que descobrir a importância dos filmes se resume em encontrar o vetor x que satisfaz $Ax = b$.

1.1.7 Separação de Ondas (Interpolação Senoidal)

Um músico responsável pela gravação dos instrumentos musicais em um estúdio percebeu que, ao gravar as faixas de determinada banda, a placa de áudio captou a microfonia dos instrumentos

por conta do mal posicionamento de um microfone específico. Após o músico ter arrumado o microfone, descobriu que os integrantes daquela banda já haviam ido embora e que não seria satisfatório chamá-los de volta para regravar todas suas faixas musicais. Dessa maneira, o músico decidiu que, de alguma maneira, removeria o chiado gerado pela microfonia diretamente nas faixas a partir de alguma técnica de separação de ondas.

Visto que as ondas captadas por aquela placa de áudio são sempre interpretadas como funções seno, temos então que a junção das ondas emitidas pelo instrumento e o chiado é uma combinação linear de suas ondas originais, em outras palavras, a onda o captada pelo equipamento é igual a uma combinação da onda n_1 do chiado com a onda n_2 do instrumento. Ou seja,

$$o = c_1 n_1 + c_2 n_2,$$

sendo n_1 e n_2 as intensidades de cada onda. Podemos visualizar uma função seno que representa duas notas $C0$ e $E1$ nos gráficos abaixo.

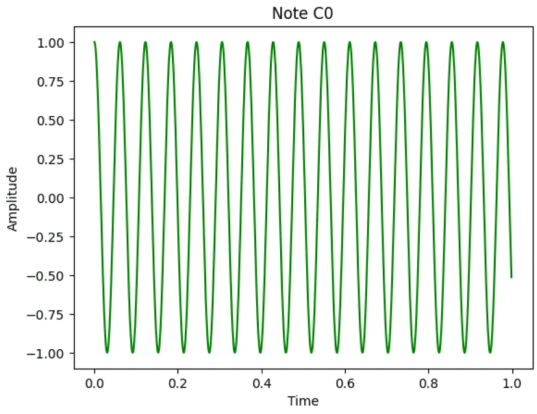


Figura F8: Representação da nota $C0$ que representa o chiado.

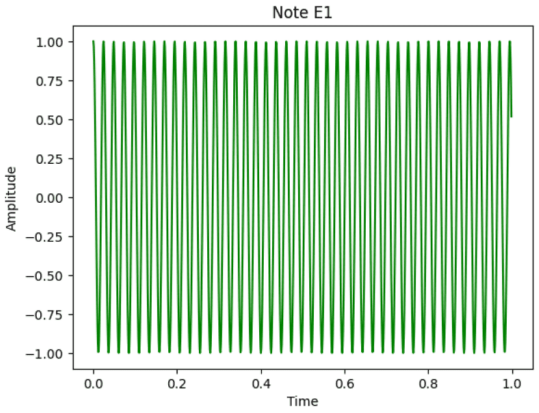


Figura F9: Representação da nota $E1$ emitida pelo instrumento.

Se juntarmos essas notas com intensidades (coeficientes) 1.2 e 0.4, por exemplo, temos a combinação:

$$o = 1.2\, C0 + 0.4\, E1,$$

que resulta na função de gráfico da figura F10.

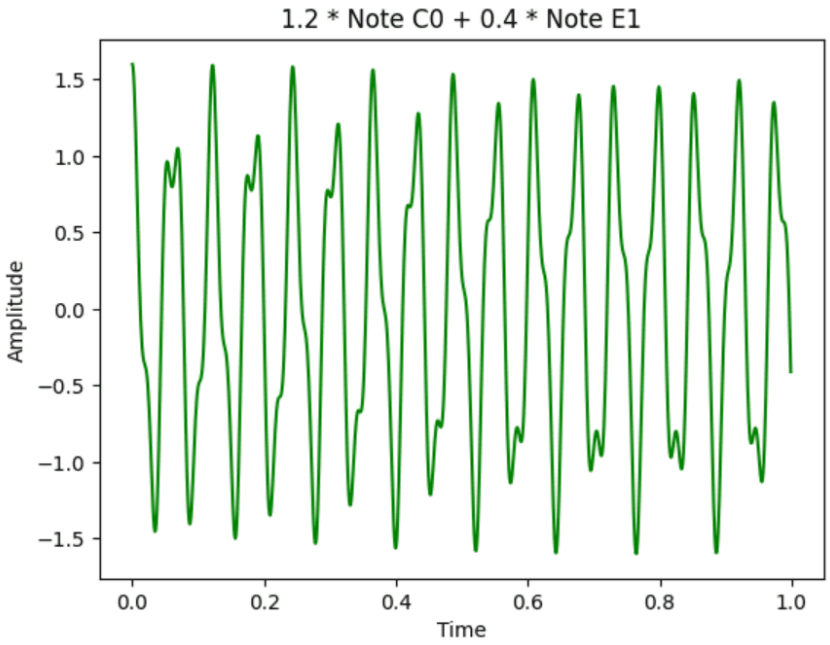


Figura F10: Representação da onda o resultante da combinação linear das notas $C0$ e $E1$.

Na seção 1.1.1, temos que resolver o sistema $Ax = b$ é equivalente a descobrir o quanto de cada vetor coluna da matriz A é necessário para representar o vetor b. Por exemplo, em uma interpolação cúbica:

$$\begin{bmatrix} x_1^0 & x_1^1 & x_1^2 & x_1^3 \\ x_2^0 & x_2^1 & x_2^2 & x_2^3 \\ x_3^0 & x_3^1 & x_3^2 & x_3^3 \\ x_4^0 & x_4^1 & x_4^2 & x_4^3 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

Que é o mesmo que:

$$c_1 \begin{bmatrix} x_1^0 \\ x_2^0 \\ x_3^0 \\ x_4^0 \end{bmatrix} + c_2 \begin{bmatrix} x_1^1 \\ x_2^1 \\ x_3^1 \\ x_4^1 \end{bmatrix} + c_3 \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_3^2 \\ x_4^2 \end{bmatrix} + c_4 \begin{bmatrix} x_1^3 \\ x_2^3 \\ x_3^3 \\ x_4^3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

Veja que, na representação do sistema como combinação linear dos vetores coluna de A, estamos tentando determinar um coeficiente que multiplica um componente que contém x_i de determinada potência. Em outras palavras, perceba que esse coeficiente determina o quanto daquele grau de polinômio é necessário para expressar cada y_i, mas perceba que cada vetor representa uma grau de polinômio:

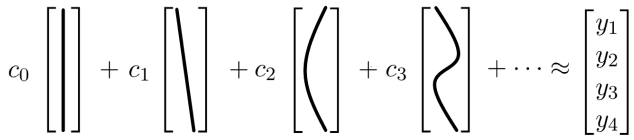


Figura F11: Perspectiva diferente da Matriz de Vandermonde.

Ou seja, se tivermos uma matriz onde cada vetor coluna representa uma frequência diferente de onda seno, temos que encontrar as frequências de n_1 e n_2 para poder separá-los é equivalente a encontrar o vetor x, que contem o quanto de cada frequência é preciso para representar a onda o, tal que $Ax = b$, sendo b o vetor que representa a onda o.

Capítulo 1.2

Modelagens Aproximadas

1.2.1 Peso (Regressão Polinomial de Grau 1)

Um fisiculturista recebeu orientação de sua equipe técnica para monitorar seu peso corporal ao longo do tempo, a fim de avaliar o impacto de sua dieta e rotina de treinamentos. Decidido a acompanhar sua evolução, ele se pesou quase todos os dias ao longo de aproximadamente três meses. Para cada dia em que registrou seu peso, ele marcou um ponto em um gráfico onde o eixo x representa a passagem do tempo em dias desde o primeiro dia de registro e o eixo y representa seu peso correspondente.

Após o período de 3 meses, o gráfico obtido foi semelhante ao ilustrado pela figura F12.

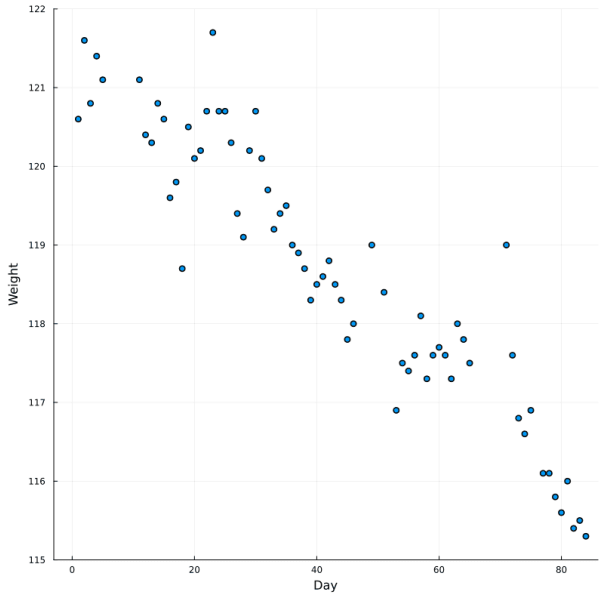


Figura F12: Gráfico que representa o peso do fisiculturista com o passar do tempo.

A equipe técnica do atleta percebeu que a variação de seu peso seguiu um padrão aproximadamente linear e, por isso, considerou interessante encontrar uma reta que *melhor* aproximasse essa tendência, com base em algum critério matemático. Dessa forma, poderiam prever com

maior precisão qual seria seu peso em determinado momento no futuro.

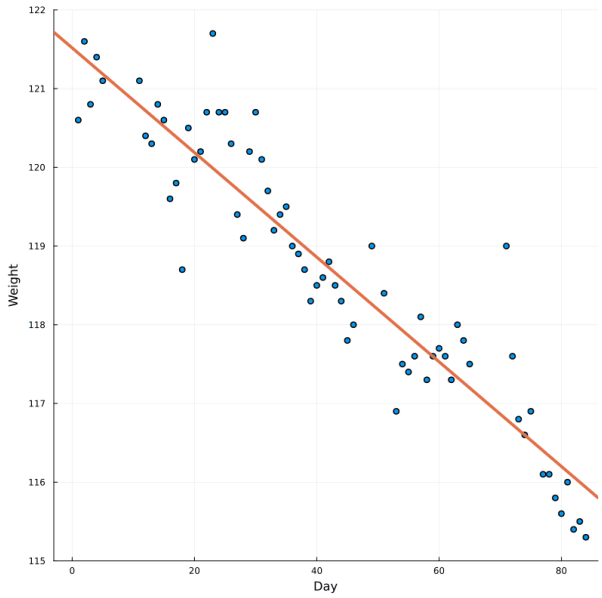


Figura F13: Gráfico que representa o peso do fisiculturista com o passar do tempo sobrescrito por uma reta que aproxima os pontos.

De maneira mais rigorosa, queremos obter uma função $f : \mathbb{R} \to \mathbb{R}$ tal que $f(x) = ax + b$ e que $\forall i \in \{1, \ldots, n\}$, $ax_i + b \approx y_i$, sendo n a quantidade de dias em que o fisiculturista se pesou, e x_i, y_i as coordenadas dos pontos no gráfico.

Dessa maneira, temos que:

$$\begin{aligned} ax_1 + b &\approx y_1, \\ &\vdots \\ ax_i + b &\approx y_i, \\ &\vdots \\ ax_n + b &\approx y_n, \end{aligned}$$

$\forall i \in \{1, \ldots, n\}$.

Podemos interpretar esse sistema na forma matriz-vetor:

$$\underbrace{\begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_i & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{x} \approx \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}}_{b}.$$

Dessa forma, só basta determinarmos algum critério de aproximação para podermos definir qual é o melhor vetor x que satisfaz $Ax \approx b$. Podemos definir o vetor que representa o erro da aproximação como sendo a diferença do vetor que usamos para aproximar pelo vetor que desejamos aproximar; em outras palavras, definimos o vetor do erro como sendo $Ax - b$ e como queremos minimizar o erro da nossa aproximação, definimos que o x que minimiza o erro deve satisfazer $x \in \operatorname{argmin}_x \ \|Ax - b\|$.

1.2.2 Previsão do Tempo (Regressão Polinomial)

Um meteorologista deseja prever o futuro comportamento atmosférico de determinada região para prever o tempo e as condições meteorológicas a partir de dados coletados acerca da temperatura daquela região com o passar do tempo. Durante um período de 11 meses, o meteorologista coletou amostras temperamentais e percebeu um certo padrão no formato dos dados. Ele decidiu que representaria cada mês a partir de uma enumeração de 1 a 11 e cada temperatura como seu próprio valor. Sendo assim, ele colocou 11 pontos em um gráfico, onde a coordenada x representa o mês e a coordenada y representa a temperatura, conforme ilustrado na figura F14.

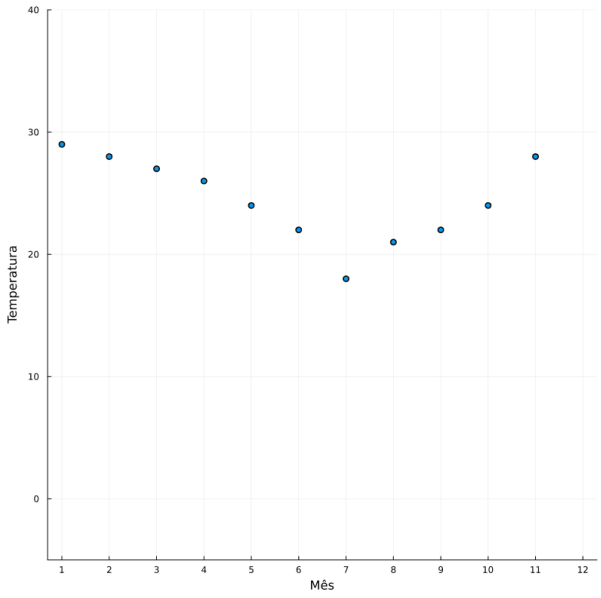


Figura F14: Gráfico que representa a temperatura a cada mês.

O meteorologista, entusiasta do campo de computação científica e análise de dados, decidiu fazer um processo de interpolação — exatamente como feito na seção 1.1.1 — para tentar prever a temperatura no mês de dezembro. Após o processo, ele encontrou uma função $f(x)$ e decidiu visualizá-la sobre o gráfico que já havia feito com os meses e as temperaturas. O resultado obtido está ilustrado na figura F15.

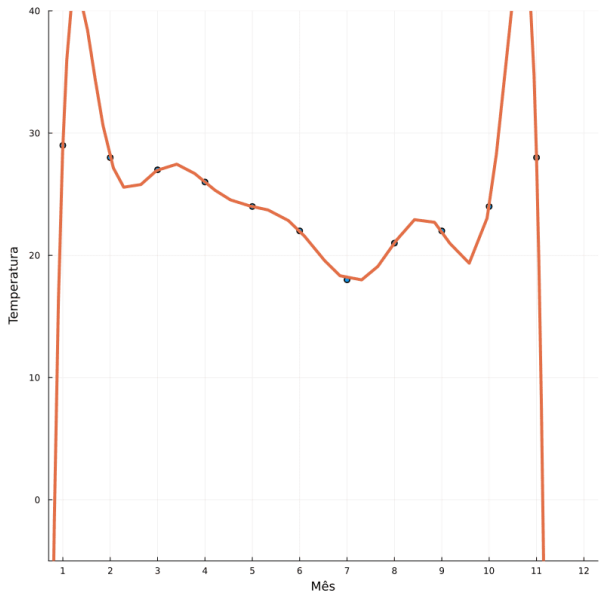


Figura F15: Gráfico que representa a temperatura a cada mês sobrescrito pela função resultante do processo de interpolação.

O meteorologista percebeu que o processo de interpolação funcionou visto que a função está de fato passando por todos os pontos representados no gráfico. Porém, ele percebeu que, se confiasse na previsão da função encontrada para aproximar a temperatura do mês de dezembro, ele diria que a temperatura seria de -1324 graus, o que não faz o menor sentido dado o contexto.

Após estudar mais a fundo o porquê da função de interpolação ter resultado em aproximações razoáveis para alguns períodos de tempo e uma péssima aproximação para outros, o meteorologista se deparou com o *fenômeno de Runge*, que é um problema de oscilação nas bordas do intervalo que ocorre sob certas circunstâncias no processo de interpolação. Por isso, o meteorologista decidiu que, ao invés de encontrar uma função polinomial que passasse exatamente por esses pontos, ele iria encontrar uma função polinomial de grau menor que aproximasse suficientemente bem esses pontos e previsse, com certa precisão, fenômenos futuros e que, por apresentar grau menor, dificilmente sofreria as consequências do *fenômeno de Runge*.

Visto que, ao invés de passar exatamente por todos os pontos, queremos aproximá-los, o objetivo passa a ser encontrar a *melhor* função polinomial $f : \mathbb{R} \to \mathbb{R}$ de grau m *baixo* tal que $f(x_i) = c_0 x_i^0 + \cdots + c_j x_i^j + \cdots + c_m x_i^m$ com $j \in \{0, \ldots, m\}$, e que, $\forall i \in \{1, \ldots, n\}$, $f(x_i) \approx y_i$. Esse problema é conhecido como *regressão polinomial*.[1]

[1] A *regressão polinomial* é um método utilizado para prever novos dados, modelar fenômenos e identificar relações entre variáveis, de maneira semelhante à interpolação polinomial. A principal diferença, no entanto, é que, na regressão, o objetivo não é ajustar o modelo para passar exatamente por todos os pontos de dados, mas sim para minimizar um erro, o que reduz problemas como o *fenômeno de Runge*, o *overfitting* e a complexidade excessiva do modelo, comuns na interpolação. Assim, a regressão é mais robusta e geralmente resulta em um modelo mais generalizável.

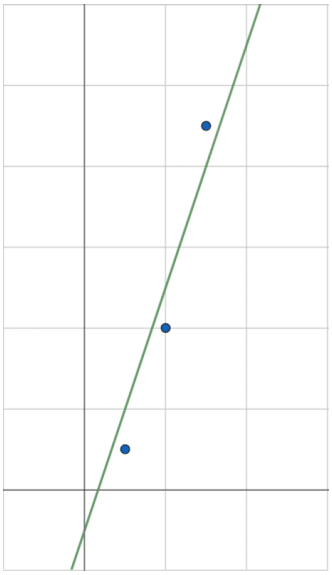


Figura F16: Regressão dos pontos $(1, 1)$, $(2, 4)$ e $(3, 9)$ por uma função de grau 1.

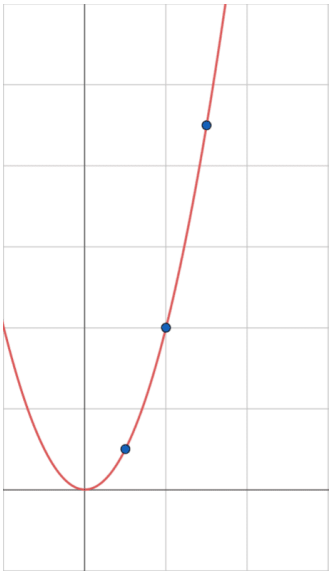


Figura F17: Interpolação dos pontos $(1, 1)$, $(2, 4)$ e $(3, 9)$.

O uso dos termos *grau baixo* e *grau alto* ao se referir ao grau das funções geradas pelo processo de interpolação pode ser enganoso visto que não existe um grau exato que podemos dizer ser alto ou baixo. Além disso, dado um conjunto de n pontos, não é correto afirmar que o polinômio interpolador terá idealmente grau $n-1$. Por exemplo, considere o caso em que n pontos são escolhidos aleatoriamente a partir de uma função quadrática. O polinômio ideal gerado pela interpolação tem grau 2, e não $n-1$. No entanto, o processo de interpolação polinomial geralmente encontrará um polinômio de grau $n-1$ por conta dos erros de ponto flutuante e ruído, mesmo que esse polinômio encontrado não represente com precisão a verdadeira identidade da função que modela o fenômeno que deu origem aos dados. Isso mostra que a interpolação pode introduzir uma complexidade indesejada quando o grau do polinômio resultante é maior que o necessário para capturar a estrutura real dos dados. Destacamos então a importância de manter o grau da função resultante do processo de regressão polinomial o mais baixo possível visto que o sobreajuste dos dados compromete a capacidade da função de generalizar para novos pontos, tornando-a inadequada para previsões. Assim, ao escolher o grau do polinômio, é fundamental encontrar um equilíbrio entre precisão e complexidade, garantindo uma boa representação dos dados sem comprometer sua interpretação prática.

Visto que f deve aproximar todos os pontos, temos o seguinte sistema de equações:

$$\begin{aligned}
y_1 &\approx c_0 x_1^0 + \cdots + c_j x_1^j + \cdots + c_m x_1^m \\
&\vdots \\
y_i &\approx c_0 x_i^0 + \cdots + c_j x_i^j + \cdots + c_m x_i^m \\
&\vdots \\
y_n &\approx c_0 x_n^0 + \cdots + c_j x_n^j + \cdots + c_m x_n^m
\end{aligned}$$

Veja que a escolha do grau m do polinômio não é objetiva e, dependendo de seu valor, a natureza do problema será diferente:

- Se $n < m + 1$, o sistema será indeterminado, com múltiplas soluções possíveis. Existem diversos polinômios de grau maior que n que passam por n pontos, mas essas soluções podem apresentar comportamento oscilatório e, em geral, não são adequadas para modelagem prática.
- Se $n = m + 1$ e os x_i forem distintos, o sistema poderá ser resolvido de forma exata visto que há exatamente um único polinômio de grau m que passará pelos n pontos. Esse é o problema de interpolação já testado pelo meteorologista e que resultou em uma péssima aproximação para futuras previsões.
- Se $n > m + 1$, teremos mais pontos do que graus no polinômio. É nesse caso que estamos mais interessados nessa seção. Existem inúmeros polinômios de grau m que aproximam os pontos. Queremos saber como encontrar o *melhor* deles.

Podemos reescrever esse sistema de equações como sendo um sistema matriz-vetor:

$$\underbrace{\begin{bmatrix} x_1^0 & \dots & x_1^j & \dots & x_1^m \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ x_i^0 & \dots & x_i^j & \dots & x_i^m \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ x_n^0 & \dots & x_n^j & \dots & x_n^m \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} c_0 \\ \vdots \\ c_j \\ \vdots \\ c_m \end{bmatrix}}_{x} \approx \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}}_{b},$$

$\forall i \in \{1, \dots, n\}$, $j \in \{0, \dots, m\}$.

Esse sistema não será exato para todo valor de m escolhido justamente pois a função polinomial não necessariamente passará por todos os pontos. Logo, não podemos resolver esse sistema pelos métodos utilizados no capítulo 2.1.

Podemos reescrever esse sistema matriz-vetor como uma combinação linear das colunas da matriz A:

$$c_0 \begin{bmatrix} x_1^0 \\ \vdots \\ x_i^0 \\ \vdots \\ x_n^0 \end{bmatrix} + \dots + c_j \begin{bmatrix} x_1^j \\ \vdots \\ x_i^j \\ \vdots \\ x_n^j \end{bmatrix} + \dots + c_m \begin{bmatrix} x_1^m \\ \vdots \\ x_i^m \\ \vdots \\ x_n^m \end{bmatrix} \approx \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}$$

É interessante observar que cada coluna dessa matriz está associada a uma base de funções polinomiais.

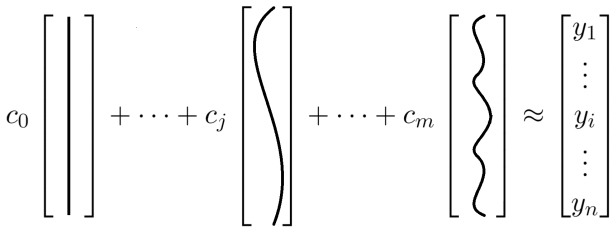


Figura F18: Representação alternativa para o sistema $Ax \approx b$.

A primeira coluna, por exemplo, corresponde a uma função constante. De forma geral, a coluna de índice j representa uma função polinomial de grau $j-1$. Neste caso, podemos interpretar o sistema como um subespaço vetorial gerado pelas colunas da matriz A. Os coeficientes c_j representam a contribuição de cada coluna para aproximar o vetor b. Isso significa que estamos buscando uma combinação linear das colunas de A que melhor aproxima b.

Veja que o vetor b não necessariamente estará contido no subespaço gerado pelas colunas da matriz A. Isso ocorre porque, em muitos casos, os dados podem conter ruídos e inconsistências ou apresentarem um padrão complexo que não pode ser exatamente representados por um polinômio de grau m. Nesse caso, a melhor aproximação possível para b consiste em encontrar uma combinação linear dos vetores coluna de A que minimiza a distância entre b e esse subespaço.

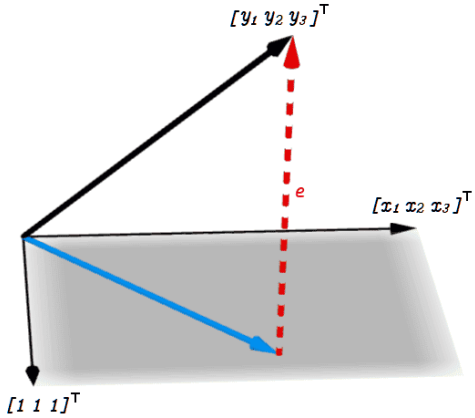


Figura F19: Visualização do problema para $m = 1$ e $n = 3$.

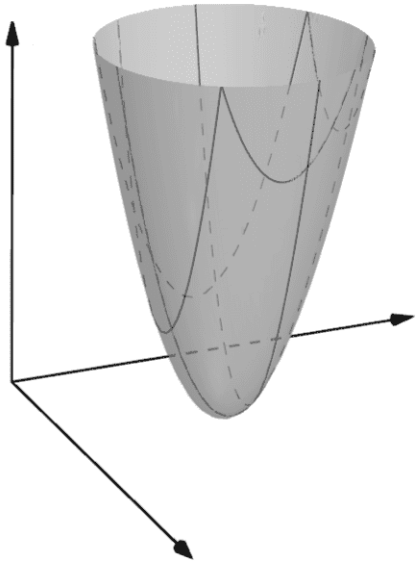


Figura F20: Paraboloide gerado pela função erro do caso da figura F19.

Definimos então o vetor do erro e como sendo

$$e = \underbrace{\begin{bmatrix} x_1^0 & \dots & x_1^j & \dots & x_1^m \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ x_i^0 & \dots & x_i^j & \dots & x_i^m \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ x_n^0 & \dots & x_n^j & \dots & x_n^m \end{bmatrix} \begin{bmatrix} c_0 \\ \vdots \\ c_j \\ \vdots \\ c_m \end{bmatrix}}_{Ax} - \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}}_{b}.$$

É previsível que, quantos mais graus de polinômios permitirmos que nossa função da regressão assuma, o erro diminuirá, visto que a função terá mais variáveis de forma a aproximar os dados pontos de um conjunto. Veja que, como mencionado anteriormente, utilizar uma função de grau próximo ao número de dados resulta em *overfitting* e em vários dos problemas da interpolação. Em contrapartida, utilizar uma função de grau extremamente baixo resulta em *underfitting*, apesar do modelo ser extremamente simples, a função resultante pode mal representar os dados. O alto erro gerado pelo *underfitting* e a baixa variação do erro gerado pelo *overfitting* podem ser visualizados nas figuras F21 e F22.

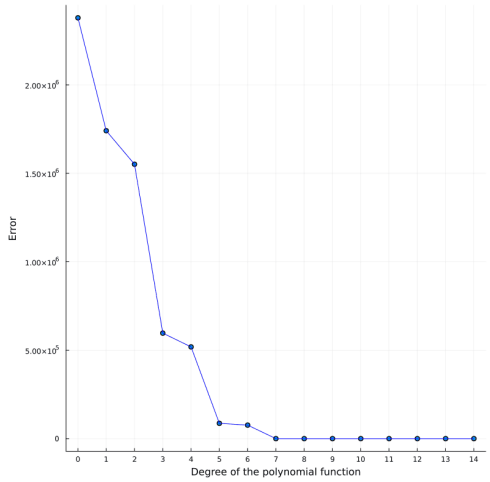


Figura F21: Gráfico arbitrário que relaciona o erro com o grau do polinômio escolhido.

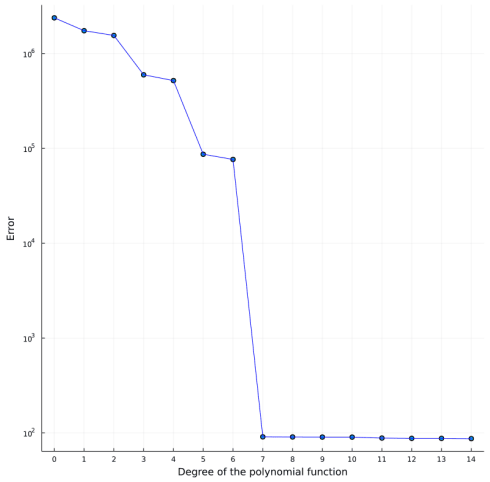


Figura F22: Gráfico arbitrário que relaciona o erro com o grau do polinômio escolhido com o eixo y em escala logarítmica.

A depender da tendência dos pontos, haverá um ponto no gráfico em que o erro cai bruscamente e parece não variar mais tanto conforme o grau do polinômio resultante do processo de regressão varia. Chamamos esse ponto de *ponto de inflexão* e consideramos que, na maioria dos casos, o valor do grau de polinômio desse ponto é o ideal para o modelo.

A respeito do gráfico da figura F21, por exemplo, a utilização de funções polinomiais de graus 0, 1 e 2 pode ser classificada como um caso de *underfitting*, enquanto a utilização de funções polinomiais de graus 12, 13 e 14 pode ser classificada como um caso de *overfitting*. Nesse caso específico, a escolha de um polinômio de grau 7 aparenta ser a mais adequada.

Por mais que, pelo gráfico F21, pareça que a escolha de grau 5, por exemplo, é excelente, a visualização do mesmo gráfico em escala logarítmica (figura F22) faz com que a disparidade entre os erros fique mais evidente e a determinação do grau apropriado para o modelo seja mais precisa.

Dessa forma, temos que o problema de regressão polinomial se resume em encontrar um vetor x que satisfaça $\text{argmin}_x \ \|Ax - b\|$.

1.2.3 Pratos par a par (Ranking)

Um cozinheiro — interessado em descobrir a comida preferida de um grupo de pessoas — decide fazer um questionário com k questões em que dois pratos são dados como opção e cada pessoa deve marcar qual dos dois ela prefere. Dessa forma, dado que m pessoas preencheram o formulário, dado n pratos diferente, o cineasta tinha um conjunto de dados com informações no formato

$$p_i : a \text{ x } b : p_j ,$$

sendo p_i e p_j pratos com $i, j \in \{1, \ldots, n\}$, $i \neq j$ e $a + b = m$, onde a é a quantidade de pessoas que prefere o prato p_i e b a quantidade de pessoas que prefere o prato p_j.

Dessa maneira, podemos representar esse resultado como sendo uma equação em que $p_i - p_j = a - b$. Ou seja, a diferença entre os pratos seria a diferença de suas pontuações, de forma que, se p_i for mais gostado do que p_j, então $p_i - p_j$ resulta em um número positivo.

Outro aspecto importante de se considerar é que o questionário não necessariamente cobre todas as combinações possíveis de 2 pratos e que a resposta das pessoas pode ser inconsistente. Por exemplo, podemos ter como resultado desse formulário que

$$\begin{aligned} p_1 - p_2 &= c_1 \\ p_1 - p_3 &= c_2 \\ p_2 - p_3 &= c_3 \end{aligned}$$

e, se tentarmos resolver esse sistema de equação de maneira exata, chegaremos em uma contradição.

Podemos escrever esse sistema de equações na forma matriz-vetor de tal modo que a matriz A seja composta pelos elementos -1, 0 e 1, que representam se determinado prato faz parte de uma determinada equação e se, se sim, se ele contribui adicionando ou subtraindo, o vetor x seja composto pela pontuação de cada prato e, por fim, o vetor b contenha os resultados das equações.

$$\underbrace{\begin{bmatrix} a_{11} & \ldots & a_{1i} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ a_{j1} & \ldots & a_{ji} & \ldots & a_{jn} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ a_{k1} & \ldots & a_{ki} & \ldots & a_{kn} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} p_1 \\ \vdots \\ p_i \\ \vdots \\ p_n \end{bmatrix}}_{x} \approx \underbrace{\begin{bmatrix} b_1 \\ \vdots \\ b_i \\ \vdots \\ b_n \end{bmatrix}}_{b}$$

sendo $a_{ji} = -1$, 0 ou 1 e b_i o resultado da equação $\forall i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, k\}$.

Dessa forma, precisamos definir algum critério de erro a ser minimizado de forma que possamos encontrar um valor para o vetor x que minimize o erro do sistema $Ax \approx b$. Diferentes definições podem ser dadas para o vetor erro e, consequentemente, resultados diferentes para o vetor x podem ser alcançados. Por fins de praticidade, simplicidade e conveniência, definiremos o vetor erro como sendo a norma da diferença de Ax por b. Dessa forma, como queremos minimizar o erro, gostaríamos de encontrar $x \in \text{argmin}_x \ \|Ax - b\|$.

1.2.4 Categoria de Filmes (Aproximação de Fatoração)

Um cineasta tem como objetivo classificar um conjunto de filmes de acordo com seu gênero apenas com as informações que ele tem em seu banco de dados. Um dos dados que o cineasta tem armazenado é uma tabela que cada posição no eixo x contém diferentes filmes e o eixo y contém os espectadores. Cada posição dessa tabela diz o quanto aquele usuário gostou daquele filme. Outra tabela tem que cada posição no eixo x contém diferentes gêneros e o eixo y contém os espectadores.

Ao abstrair essas tabelas como matrizes, percebemos que temos 2 matrizes:

$$\underbrace{\begin{bmatrix} \\ A \\ \\ \end{bmatrix}}_{A_{(\text{Usuários} \times \text{Filmes})}} \quad \text{e} \quad \underbrace{\begin{bmatrix} \\ C \\ \\ \end{bmatrix}}_{C_{(\text{Usuários} \times \text{Gêneros})}} .$$

Visto que queremos uma matriz $B_{(\text{Filmes} \times \text{Gêneros})}$, queremos então determinar B tal que $AB \approx C$.[2]

Por não termos a garantia de que de fato exista uma matriz B tal que $AB = C$, queremos então determinar a matriz B que *melhor* aproxima C ao ser multiplicada por A. Visto que nem sempre existe uma solução exata, queremos encontrar B que satisfaça alguma equação de minimização de erro de aproximação. Podemos definir a matriz do erro como sendo $AB - C$ e, consequentemente, temos então que queremos encontrar $B \in \text{argmin}_B \ \|AB - C\|_{\mathsf{F}}$.

Como B é uma matriz, podemos interpretar a matriz A ser aplicada em B como na verdade sendo a matriz A sendo aplicada nos vetores coluna de B para se aproximar dos vetores coluna de C. Desse modo, pelos teoremas [X] e [Y], temos que $\text{argmin}_B \ \|AB - C\|_{\mathsf{F}} = \text{argmin}_B \ \sum_{i=1}^{n} \|Ab_i - c_i\|^2$, sendo n a quantidade de vetores coluna de B e C.

[2]Interessante notar como a *tipagem* das matrizes nos diz muito sobre o que elas representam. Por exemplo, se a matriz A nos diz o quanto cada usuário gosta de determinados filmes e a matriz B nos diz quanto de cada gênero um filme é, então o produto dessas matrizes nos diz o quanto cada usuário gosta de cada gênero.

Capítulo 1.3

Modelagens Dinâmicas

1.3.1 Bactérias (Modelo de Leslie)

Um biólogo está estudando o crescimento populacional de determinada população de bactérias e deseja determinar o tamanho da população em determinado período de tempo. Para isso, o biólogo decide dividir a população em três faixas etárias: jovens, adultas e idosas. Os dados estatísticos de seu laboratório determinam diferentes probabilidades dessas bactérias permanecem ou mudarem de faixa etária a cada mês que passa:

- p_i: probabilidade de uma bactéria da faixa etária i *permanecer* na faixa etária i após 1 mês;
- e_i: probabilidade de uma bactéria da faixa etária i *envelhecer* para a faixa etária $i+1$ após 1 mês;
- f_i: probabilidade de uma bactéria da faixa etária i *fecundar* uma bactéria jovem após 1 mês.

Veja então que a quantidade de bactérias em determinada faixa etária e em determinado período de tempo pode ser escrita como função da quantidade de bactérias no mês anterior:

$$\begin{aligned} j(k) &= p_1 j(k-1) + f_2 a(k-1) + f_3 i(k-1), \\ a(k) &= e_1 j(k-1) + p_2 a(k-1) + \ 0i(k-1), \\ i(k) &= \ 0j(k-1) + e_2 a(k-1) + p_3 i(k-1), \end{aligned}$$

sendo $j(k)$, $a(k)$ e $i(k)$ as funções que determinam a quantidade de jovens, adultos e idosos, respectivamente, no mês k. Veja que essas funções são acopladas e recursivas por natureza: a quantidade de jovens em um mês k naturalmente depende da quantidade de adultos e idosos no mês passado, e a quantidade de adultos e idosos no mês passado dependerá da quantidade de jovens, adultos e idosos no mês retrasado e sucessivamente até que se chegue no mês inicial.

Podemos visualizar esse sistema de equações na forma matriz-vetor tal que:

$$\underbrace{\begin{bmatrix} j^k \\ a^k \\ i^k \end{bmatrix}}_{x_k} = \underbrace{\begin{bmatrix} p_1 & f_2 & f_3 \\ e_1 & p_2 & 0 \\ 0 & e_2 & p_3 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} j^{k-1} \\ a^{k-1} \\ i^{k-1} \end{bmatrix}}_{x_{k-1}},$$

sendo j^k, a^k e i^k a quantidade de jovens, adultos e idosos, respectivamente, no mês k, x_k o vetor que armazena essas informações e A a matriz de transição de um mês para o outro.[1]

Além disso, como A é uma matriz de transição que leva a população de bactérias de um mês k ao mês $k+1$, temos que aplicar a matriz k vezes no vetor que representa a população no mês 0 é o mesmo que calcular esse vetor no instante k:

$$\begin{bmatrix} j^k \\ a^k \\ i^k \end{bmatrix} = \begin{bmatrix} p_1 & f_2 & f_3 \\ e_1 & p_2 & 0 \\ 0 & e_2 & p_3 \end{bmatrix} \begin{bmatrix} j^{k-1} \\ a^{k-1} \\ i^{k-1} \end{bmatrix} = \begin{bmatrix} p_1 & f_2 & f_3 \\ e_1 & p_2 & 0 \\ 0 & e_2 & p_3 \end{bmatrix}^2 \begin{bmatrix} j^{k-2} \\ a^{k-2} \\ i^{k-2} \end{bmatrix} = \cdots = \begin{bmatrix} p_1 & f_2 & f_3 \\ e_1 & p_2 & 0 \\ 0 & e_2 & p_3 \end{bmatrix}^k \begin{bmatrix} j^0 \\ a^0 \\ i^0 \end{bmatrix}.$$

Assim, temos que $x_k = A^k x_0$.

Portanto, se o biólogo deseja descobrir a população de bactérias no mês k, é necessário que ele aplique a matriz de transição no vetor que representa a população inicial k vezes. Sendo essa, então, uma solução para seu problema mesmo que ineficiente. Dessa forma, gostaríamos de descobrir uma "fórmula fechada" para descobrir a quantidade de pessoas em cada faixa etária em um mês arbitrário. Em outras palavras, gostaríamos de resolver o sistema de recorrências.

1.3.2 Vampiros (Sistemas de EDOs)

Em um universo alternativo onde vampiros existam simultaneamente a humanos, um demógrafo responsável pelo levantamento de dados acerca de diferentes taxas acerca do crescimento populacional dessa espécie está conduzindo um estudo em que deseja descobrir se em algum momento futuro os vampiros entrarão em extinção, os humanos entrarão em extinção ou se ambos se encontrarão em equilíbrio.

O demógrafo percebeu que a função $x(t)$ que determina a população de humanos em um dado momento t depende diretamente da função $y(t)$ que determina a população de vampiros e vise-versa. Seja:

- α: taxa de crescimento populacional dos humanos,
- β: taxa de impacto dos vampiros sobre os humanos,
- γ: taxa de impacto dos humanos sobre os vampiros,

[1] Essa matriz A que contém as informações sobre taxas de natalidade e fecundidade de diferentes faixas etárias de uma população não é específica desse problema em específico. Mais informações a respeito dessa matriz e problemas parecidos podem ser encontrados pelo nome *Modelo de Leslie*.

- δ: taxa de crescimento populacional dos vampiros.

A taxa de crescimento populacional dos humanos α depende da natalidade e fatalidade natural dos humanos independentemente dos vampiros, e o mesmo para a taxa de crescimento δ dos vampiros. Isso é representado por $\alpha h(t)$ para humanos e $\delta v(t)$ para vampiros. A população de humanos também é afetada pela presença de vampiros (visto que os vampiros contaminam humanos) e a população de vampiros é afetada pela população de humanos (visto que humanos se defendem de vampiros). Isso é representado por $\beta v(t)$ para humanos e $\gamma h(t)$ para vampiros.

Além disso, a população inicial de humanos nesse universo é de χ e a de vampiros é υ.

Dessa forma, chegamos no sistema de equações diferenciais:

$$\begin{cases} h_t(t) = \alpha h(t) = \beta v(t) \\ v_t(t) = \gamma h(t) + \delta v(t) \\ h(0) = \chi \\ v(0) = \upsilon \end{cases}$$

Veja que esse sistema pode ser escrito na forma matriz-vetor como:

$$\underbrace{\begin{bmatrix} h_t \\ v_t \end{bmatrix}}_{x_t} = \underbrace{\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} h \\ v \end{bmatrix}}_{x}$$

Dessa forma, temos que o problema dos vampiros se resume em resolver o sistema de equações diferenciais $x_t = Ax$ de maneira a encontrar uma solução para $h(t)$ e $v(t)$ que não dependam uma da outra e nem de suas derivadas.

1.3.3 Coelhos de Fibonacci

Um cunicultor cria uma determinada população de coelhos, dividida em diferentes faixas etárias. A quantidade de indivíduos dessa população em determinado momento depende da quantidade de indivíduos nos momentos anteriores. Essa população de coelhos começou com apenas 2 coelhos, um macho e uma fêmea. Inicialmente, esses coelhos eram jovens e não se reproduziam mas, após 4 meses, começam a ter filhotes, que, por sua vez, também crescerão e também terão filhotes daqui a 4 meses.

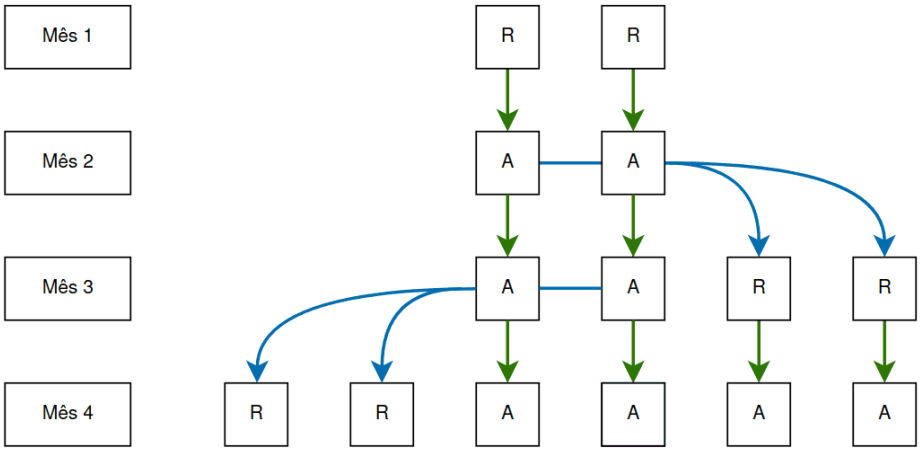


Figura F23: Representação da quantidade de coelhos em uma população com o passar dos meses. Onde R significa um coelho recém-nascido, A significa um coelho adulto, as setas retas/verdes representam o crescimento/mantimento dos coelhos e as setas curvas/azuis representam o nascimento de novos coelhos.

A cada 4 meses, um par de coelhos adultos terá um novo par de filhotes, e esses filhotes, após 4 meses, também começarão a se reproduzir. Assim, a cada 4 meses o número de coelhos na população aumenta de acordo com as gerações anteriores. O número total de coelhos a cada 4 meses segue uma sequência que depende da soma da quantidade de coelhos dos 8 meses anteriores.

Sendo assim, para determinar a quantidade de coelhos em um momento n arbitrário, é necessário saber a quantidade de coelhos no momento $n-1$ e no momento $n-2$, que, por sua vez, depende da quantidade de coelhos nos momentos $n-3$ e $n-4$, e assim por diante, até que a sequência chegue aos meses iniciais — nos quais sabemos que haviam apenas 2 coelhos.

Visto que a população de coelhos no instante n depende diretamente da população de coelhos nos instantes $n-1$ e $n-2$ e que, todos os coelhos do instante $n-1$ continuam ou se tornam adultos e que todos os que já eram adultos (os coelhos do instante $n-2$) terão a mesma quantidade de filhotes, temos que para um instante n arbitrário, a quantidade de coelhos é $f(n) = f(n-1) + f(n-2)$, sendo f a função que simboliza essa sequência. Essa sequência é conhecida como a sequência de Fibonacci, com a única diferença sendo o caso base que, ao invés de $f(0) = 2$ e $f(1) = 2$, temos $f(0) = 0$ e $f(1) = 1$.

De maneira mais precisa, seja $f(n)$, $f : \mathbb{N} \to \mathbb{N}$, a função que calcula o n° elemento da sequência de Fibonacci.

$$f(n) = \begin{cases} 0, & \text{se } n = 0, \\ 1, & \text{se } n = 1, \\ f(n-1) + f(n-2), & \text{caso contrário.} \end{cases}$$

Embora a definição recursiva de f seja elegantemente simples, ela apresenta um custo computacional elevado para entradas grandes. Isso ocorre porque cada chamada de f gera duas novas chamadas recursivas, resultando em um crescimento exponencial no número de cálculos

necessários. Consequentemente, f possui uma complexidade de tempo de $O(2^n)$, tornando-se impraticável para valores elevados de n.

Por isso, dada uma função recursiva e discreta $f(n)$, $f : \mathbb{N} \to \mathbb{R}$, queremos encontrar uma "fórmula fechada" para f.

Na sequência de Fibonacci, por exemplo, para calcular $f(n)$ no caso geral, é necessário conhecer os dois valores anteriores na sequência, $f(n-1)$ e $f(n-2)$. No entanto, se esses valores fossem armazenados, seria possível computar $f(n)$ com complexidade $O(1)$ visto que seria apenas somar os dois valores já pré-computados. Isso ocorre porque a sequência utiliza uma quantidade fixa de elementos anteriores para determinar o próximo valor. Assim, cada passo da sequência pode ser interpretado como uma nova computação baseada em valores já conhecidos. O padrão nos primeiros cálculos pode ser observado a seguir:

$$\begin{aligned} f(2) &= f(1) + f(0) \\ f(3) &= f(2) + f(1) \\ f(4) &= f(3) + f(2) \\ f(5) &= f(4) + f(3) \\ &\vdots \end{aligned}$$

Para calcular $f(6)$, é vantajoso já ter os valores de $f(5)$ e $f(4)$ previamente computados, evitando redundâncias e tornando o processo mais eficiente.

A equação para o cálculo de $f(6)$ pode ser interpretada da seguinte maneira:

$$\begin{cases} f(6) = f(5) + f(4), \\ \\ f(5) = f(5). \end{cases}$$

Embora a equação inferior pareça redundante, ela tem uma função importante. Ao calcular o próximo valor da sequência, precisaremos manipular apenas as "variáveis" $f(6)$ e $f(5)$, que já foram calculadas. Isso nos permite calcular os valores de $f(7)$ e $f(6)$ de maneira eficiente.

Generalizando esse sistema, temos

$$\begin{cases} f(n) = f(n-1) + f(n-2), \\ \\ f(n-1) = f(n-1). \end{cases}$$

Esse sistema pode ser reescrito de forma equivalente como:

$$\begin{cases} f(n) = 1f(n-1) + 1f(n-2), \\ \\ f(n-1) = 1f(n-1) + 0f(n-2). \end{cases}$$

Essa expressão deixa bem explícito a semelhança da sequência de Fibonacci com o produto matriz-vetor:

$$\underbrace{\begin{bmatrix} f(n) \\ f(n-1) \end{bmatrix}}_{x_n} = \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} f(n-1) \\ f(n-2) \end{bmatrix}}_{x_{n-1}}, \quad \text{sendo } x_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Perceba, então, que A é uma matriz de transição. Dada uma instância x_k da sequência, o próximo elemento será dado por $x_{k+1} = Ax_k$. Seguindo essa lógica, podemos concluir que o elemento k da sequência é o o mesmo que a aplicação da matriz k vezes no vetor inicial. Ou seja, $x_k = A^k x_0$. Dessa forma, podemos expressar o elemento x_k como sendo uma matriz aplicada nos elementos iniciais da sequência, de maneira análoga ao que fazemos em uma recorrência.

Visto a ineficiência desse método para o cálculo do elemento k da sequência de Fibonacci, gostaríamos de encontrar uma "fórmula fechada" que encontre o valor de x_k de maneira computacionalmente mais eficiente.

1.3.4 Navegando a Internet (Matriz Probabilística, Cadeia de Markov)

Uma empresa responsável por um mecanismo de pesquisa precisa desenvolver algum critério para indicar a um usuário qual site é o mais importante a partir de algum critério bem definido matematicamente.

Um estagiário dessa empresa percebeu que a relação entre sites e usuários pode ser representada a partir de um grafo direcional onde cada vértice representa um site e cada aresta representa a probabilidade do usuário ir de um site para o outro após um certo período de tempo, seja por redirecionamento ou *link*.

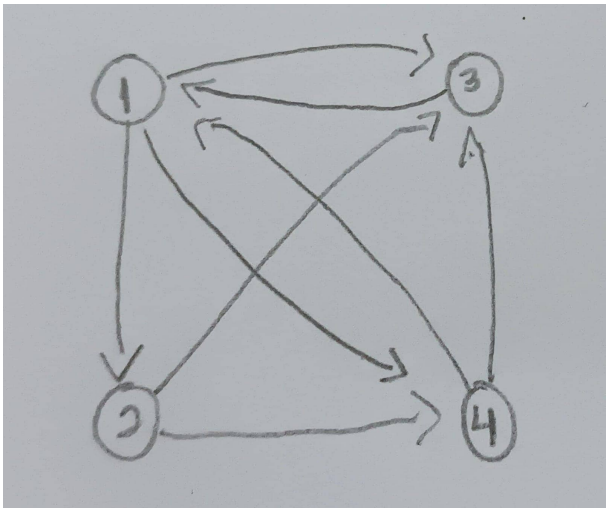


Figura F24: Representação dos sites e probabilidades de mudança de um site para outro.

Dessa forma, podemos medir a importância de um site com base no fluxo de visitantes que

ele recebe. A ideia principal é que um site é importante se ele recebe muitos visitantes e for muito provável do usuário ir para ele.

De maneira mais criteriosa, seja x_i^k a quantidade de pessoas no site i no instante k, n o número total de sites, p_{ji} a probabilidade de um usuário sair do site j para o site i.[2] Definimos a quantidade de usuários no site i no instante k como:

$$x_i^{k+1} = \sum_{j=1}^{n} p_{ij} x_j^k \quad , \quad \forall i, j \in \{1, \ldots, n\}.$$

Dessa forma, temos um modelo dinâmico de tal forma que a quantidade de usuários em determinado site no instante $k+1$ depende da quantidade de usuários nos outros sites no instante k.

Além disso, veja que, como essa expressão vale $\forall i \in \{1, \ldots, n\}$, temos então um sistema de equações que pode ser representado na forma matriz-vetor onde A é a matriz probabilística de transição que contém valores fracionários entre 0 a 1 e x_k é o vetor com os elementos x_i no instante k.

$$\underbrace{\begin{bmatrix} x_1^k \\ \vdots \\ x_i^k \\ \vdots \\ x_n^k \end{bmatrix}}_{x_k} = \underbrace{\begin{bmatrix} p_{11} & \cdots & p_{1j} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ p_{i1} & \cdots & p_{ij} & \cdots & p_{in} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nj} & \cdots & p_{nn} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1^{k-1} \\ \vdots \\ x_i^{k-1} \\ \vdots \\ x_n^{k-1} \end{bmatrix}}_{x_{k-1}}$$

Dessa forma, nosso problema se resume em descobrir qual posição do vetor x_k tem o maior valor após A ser aplicado várias vezes no vetor x_0, que representa a quantidade de pessoas no instante inicial.

1.3.5 Banco Imobiliário (Matriz Probabilística, Cadeia de Markov)

Um game designer percebeu ao jogar Banco Imobiliário que um de seus amigos ganha com uma frequência maior do que os outros e que esse amigo sempre compra residências em lugares repetidos em toda partida que joga. O game designer, então, decidiu que de alguma forma calcularia qual é a posição no jogo que é mais provável de se pisar e, consequentemente, a posição que mais rende pro jogador.

No caso de um tabuleiro padrão sem aspectos específicos do jogo como "prisão" e "volte

[2]Por p_{ji} ser a probabilidade de um usuário sair do site j para o site i e não é possível que 2 sites sejam acessados simultaneamente pelo usuário, então temos que cada coluna p_j da matriz tenha soma 1 visto que a soma das probabilidades precisa ser necessariamente 1, caracterizando a matriz A como *probabilística*.

algumas casas", teríamos que a probabilidade de se pisar em cada casa seria uniforme, visto que, se um dado for jogado em que o número resultante define quantas casas o jogador avançará, a probabilidade de se pisar em qualquer uma das próximas 6 casas é a mesma, visto que o dado apresenta a mesma probabilidade de cair um número de 1 a 6.

Nesse jogo específico de Banco Imobiliário, algumas das posições especiais do tabuleiro apresentam a mensagem "volte 2 casas", enquanto outras apresentam a mensagem "avance 1 casa", fazendo com que a probabilidade de terminar a rodada em uma das posição que é levada por uma posição especial é maior. Ainda assim, visto que esse tabuleiro em específico apresenta várias casas com essa dinâmica, não é claro qual delas é a que o jogador tem maior probabilidade de cair.

Desse modo, para as casas padrões do jogo, dizemos que a probabilidade do jogador terminar a rodada k na casa i é

$$p_i^k = \frac{p_{i-1}^{k-1}}{6} + \frac{p_{i-2}^{k-1}}{6} + \frac{p_{i-3}^{k-1}}{6} + \frac{p_{i-4}^{k-1}}{6} + \frac{p_{i-5}^{k-1}}{6} + \frac{p_{i-6}^{k-1}}{6}.$$

Em outras palavras, a probabilidade de na rodada k o jogador pisar na casa i é o somatório de $1/6$ da probabilidade de se estar em uma das 6 casas anteriores, visto que é só delas que é possível alcançar a casa i e que, pela distribuição do dado ser uniforme, a chance de em cada uma das 6 casas anteriores terminar a rodada na casa i é de $1/6$.

A complexidade do sistema surge no instante que as casas especiais são envolvidas. Supondo que, nesse momento, a casa $i-1$ seja uma casa com a mensagem "avance 1 casa". Nesse caso, a probabilidade de terminar a rodada k na casa i é

$$p_i^k = \frac{2p_{i-2}^{k-1}}{6} + \frac{2p_{i-3}^{k-1}}{6} + \frac{2p_{i-4}^{k-1}}{6} + \frac{2p_{i-5}^{k-1}}{6} + \frac{2p_{i-6}^{k-1}}{6} + \frac{p_{i-7}^{k-1}}{6}.$$

Dessa vez, a probabilidade de terminar a rodada em i estando em $i-1$ no round anterior é 0 visto que a probabilidade de terminar a rodada em $i-1$ é 0. Dessa vez, veja que a casa $i-7$ aparece na equação visto que é possível da $i-7$ pisar na $i-1$ e, consequentemente, terminar a rodada em i, e também a probabilidade de terminar a rodada em i estando nas casas $i-2$ a $i-6$ na rodada anterior é o dobro visto que é possível terminar em i ao, de fato, terminar em i ou ir para $i-1$ e avançar para i.

Dessa maneira, temos um sistema de equações extremamente subjetivo para cada tabuleiro mas que, da mesma maneira, pode ser escrito em um sistema dinâmico de tal forma que

$$\underbrace{\begin{bmatrix} p_1^k \\ \vdots \\ p_i^k \\ \vdots \\ p_n^k \end{bmatrix}}_{x_k} = \underbrace{\begin{bmatrix} a_{11} & \dots & a_{1i} & \dots & a_{1n} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ a_{i1} & \dots & a_{ii} & \dots & a_{in} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{ni} & \dots & a_{nn} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} p_1^{k-1} \\ \vdots \\ p_i^{k-1} \\ \vdots \\ p_n^{k-1} \end{bmatrix}}_{x_{k-1}}$$

x_k seja o vetor com as probabilidades x_i $\forall i \in \{1, \ldots, n\}$, sendo n o número de casas padrões e A seja a matriz de transição de uma rodada para a outra.

Dessa forma, nosso problema se resume em descobrir qual posição do vetor x_k tem o maior valor após A ser aplicado várias vezes no vetor x_0, que representa a distribuição da probabilidade no momento inicial do jogo.

1.3.6 Difusão do Calor

Um físico está realizando um experimento com o objetivo de estudar a difusão do calor em uma barra de metal que é aquecida continuamente por uma fonte em um de seus extremos e esfriada continuamente no outro, de modo que a temperatura nos extremos seja praticamente constante e igual à temperatura aplicada.

O físico acredita que, em algum momento inicial, a temperatura da barra fosse completamente uniforme, em torno de b° Celsius,e que, com o passar o tempo, o calor se espalhe dessa barra até que um equilíbrio seja atingido em algum momento.

A barra pode ser discretizada uniformemente em n nós de tal forma que a temperatura em um dado ponto p_i em um instante $k+1$ possa ser descrita como sendo a média das temperaturas dos nós vizinhos no instante anterior. Dessa forma, temos então que

$$t_i^{k+1} = \frac{t_{i-1}^k + t_{i+1}^k}{2}, \quad \forall i \in \{2, \ldots, n-1\},$$

onde t_1 e t_n são valores constantes conhecidos.

Por essa equação ser válida $\forall i \in \{2, \ldots, n-1\}$, temos então o sistema de equações

$$\begin{aligned}
t_2^{k+1} &= \frac{t_3^k}{2} + \frac{t_1}{2}, \\
&\vdots \\
t_i^{k+1} &= +\frac{t_{i-1}^k}{2} + \frac{t_{i+1}^k}{2}, \\
&\vdots \\
t_{n-1}^{k+1} &= \frac{t_{n-2}^k}{2} + \frac{t_n}{2},
\end{aligned}$$

que pode ser interpretado como o sistema dinâmico

$$\underbrace{\begin{bmatrix} t_2^{k+1} \\ \vdots \\ t_i^{k+1} \\ \vdots \\ t_{n-1}^{k+1} \end{bmatrix}}_{x_{k+1}} = \underbrace{\begin{bmatrix} 0 & 1/2 & & & \\ 1/2 & \ddots & & & \\ & & 0 & & \\ & & & \ddots & 1/2 \\ & & & 1/2 & 0 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} t_2^k \\ \vdots \\ t_i^k \\ \vdots \\ t_{n-1}^k \end{bmatrix}}_{x_k} + \underbrace{\begin{bmatrix} \frac{t_1}{2} \\ \vdots \\ 0 \\ \vdots \\ \frac{t_n}{2} \end{bmatrix}}_{c},$$

onde A é uma matriz de transação que marca a passagem do sistema de um estado k para um estado $k+1$, x_k é um vetor com as temperaturas dos nós internos no instante k, c sendo o vetor com as condições de fronteira, e x_0 sendo o vetor com os valores do chute inicial do físico.

A partir dessa modelagem, temos então que encontrar a temperatura nos nós internos em um instante k a partir de um chute inicial b se resume em encontrar x_{k+1} tal que $x_{k+1} = Ax_k$.

Capítulo 1.4

Modelagens de Reconhecimento de Padrões

1.4.1 Estrada para Bombeiros

O corpo de bombeiros junto de uma equipe de engenheiros estão planejando a construção de uma estrada reta que permita ao caminhão dos bombeiros chegar o mais próximo possível de todas as n casas distribuídas pelo território da cidade em caso de incêndio. O objetivo é encontrar a melhor orientação para essa estrada de forma estratégica, garantindo maior acessibilidade.

Para modelar esse problema, podemos representar cada casa como um ponto em um plano cartesiano em que o eixo x representa a longitude da casa e o eixo y representa a latitude. Dessa forma, a posição de cada casa pode ser descrita por sua coordenadas (x_i, y_i), onde i representa o índice da casa de 1 a n.

Nosso objetivo é encontrar uma linha reta que passe o mais próximo possível dessas casas, no sentido de minimizar a distância total entre as casas e a estrada. Para isso, precisamos escolher uma direção x, que abstrairemos como um vetor, ao longo da qual essa estrada será construída.

Se as casas estivessem perfeitamente alinhadas, a resposta seria simples: a estrada seguiria uma direção simples de determinar. No entanto, quando as casas estão distribuídas de maneira mais dispersa, precisamos de um critério melhor definido matematicamente para encontrar a melhor orientação.

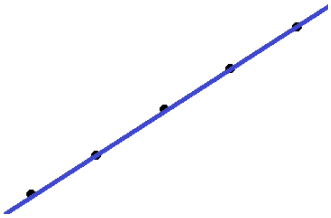


Figura F25: Representação de casas alinhadas e uma avenida que passa por elas.

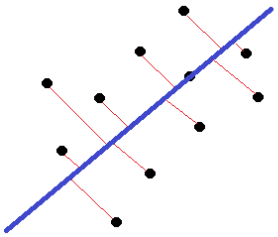


Figura F26: Representação de casas não-alinhadas e uma avenida que passa próximo a essas casas.

Definimos o erro de uma dada casa em relação à avenida como sendo a distância dessa casa à avenida ao quadrado; e o erro total do método como sendo o somatório dos quadrados das distâncias:

$$
\begin{aligned}
& \operatorname{argmin}_x \sum_{i=1}^{n} \operatorname{dist}(a_i, x)^2 \\
= \quad & \{ \text{ Teorema T25 } \} \\
& \operatorname{argmin}_x \sum_{i=1}^{n} \sqrt{\|a_i\|^2 - \left(\frac{a_i^\mathsf{T} x}{\|x\|}\right)^2}^2 \\
= \quad & \{ \text{ Simplificação } \} \\
& \operatorname{argmin}_x \sum_{i=1}^{n} \|a_i\|^2 - \left(\frac{a_i^\mathsf{T} x}{\|x\|}\right)^2 \\
= \quad & \{ \text{ Teorema T32 } \} \\
& \operatorname{argmin}_x \sum_{i=1}^{n} - \left(\frac{a_i^\mathsf{T} x}{\|x\|}\right)^2 \\
= \quad & \{ \text{ Teorema T35 } \} \\
& \operatorname{argmax}_x \sum_{i=1}^{n} \left(\frac{a_i^\mathsf{T} x}{\|x\|}\right)^2 \\
= \quad & \{ \text{ Remoção do termo constante do somatório } \} \\
& \operatorname{argmax}_x \frac{1}{\|x\|^2} \sum_{i=1}^{n} (a_i^\mathsf{T} x)^2 \\
= \quad & \{ \text{ Definição } \} \\
& \operatorname{argmax}_x \frac{1}{\|x\|^2} \left\| \begin{bmatrix} a_1^\mathsf{T} x \\ \vdots \\ a_n^\mathsf{T} x \end{bmatrix} \right\|^2 \\
= \quad & \{ \text{ Reescrita matricial do vetor } \} \\
& \operatorname{argmax}_x \frac{1}{\|x\|^2} \left\| \begin{bmatrix} — & a_1^\mathsf{T} & — \\ & \vdots & \\ — & a_n^\mathsf{T} & — \end{bmatrix} \begin{bmatrix} | \\ x \\ | \end{bmatrix} \right\|^2 \\
= \quad & \{ \text{ Reescrita do produto matriz-vetor } \} \\
& \operatorname{argmax}_x \frac{1}{\|x\|^2} \|Ax\|^2 \\
= \quad & \{ \ \frac{1}{b} a = \frac{a}{b} \ \} \\
& \operatorname{argmax}_x \frac{\|Ax\|^2}{\|x\|^2}
\end{aligned}
$$

Dessa forma, temos que o problema dos bombeiros se resume em encontrar o vetor x tal que $x \in \operatorname{argmax}_x \frac{\|Ax\|^2}{\|x\|^2}$.

Exercício E7:

Na seção 1.4.1 estamos calculando

$$\text{argmin}_x \sum_{i=1}^{n} \text{dist}(a_i, x)^2,$$

ou seja, calculando o conjunto dos valores de x que minimizam o somatório dos quadrados das distâncias dos pontos até x. Ainda assim, perceba que, para esse conjunto existir, é necessário que essa função apresente valor mínimo.

Prove que $\sum_{i=1}^{n} \text{dist}(a_i, x)^2$ apresenta ponto de mínimo.

Dica: Expanda a definição da função de distância.

1.4.2 Um(s) e Zero(s)

O gerente de um banco que ainda trabalha com cheques assinados decide inventar um método automático de determinar se o dígito escrito por um cliente é um dígito 0 ou um dígito 1. O dígito é escrito em um capturador de assinatura digital e, internamente, é representado como uma matriz de dimensão (3×3). Algumas representações possíveis para os dígitos 0 e 1 são:

- 0: $\begin{bmatrix} \blacksquare & \blacksquare & \blacksquare \\ \blacksquare & \square & \blacksquare \\ \blacksquare & \blacksquare & \blacksquare \end{bmatrix}$, $\begin{bmatrix} \square & \blacksquare & \square \\ \blacksquare & \square & \blacksquare \\ \square & \blacksquare & \square \end{bmatrix}$, $\begin{bmatrix} \square & \blacksquare & \blacksquare \\ \blacksquare & \square & \blacksquare \\ \blacksquare & \blacksquare & \square \end{bmatrix}$, ou então, $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$, $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$;
- 1: $\begin{bmatrix} \square & \blacksquare & \square \\ \square & \blacksquare & \square \\ \square & \blacksquare & \square \end{bmatrix}$, $\begin{bmatrix} \blacksquare & \blacksquare & \square \\ \square & \blacksquare & \square \\ \blacksquare & \blacksquare & \blacksquare \end{bmatrix}$, $\begin{bmatrix} \blacksquare & \blacksquare & \blacksquare \\ \square & \blacksquare & \square \\ \blacksquare & \blacksquare & \blacksquare \end{bmatrix}$, ou então, $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$.

Diferentes critérios podem ser definidos pelo próprio gerente do banco, como, por exemplo, representações mais simétricas e largas apresentam maior probabilidade de serem o dígito 0 do que o dígito 1. Ainda assim, o gerente decide que deve buscar algum critério matemático mais fundamentado.

Podemos representar essas matrizes como pontos em $\mathbb{R}^9$ e, de alguma maneira, visualizar em alguma dimensão menor o quão semelhantes são essas diferentes representações para que, dessa forma, dado uma nova representação estranha, como

$$\begin{bmatrix} \square & \blacksquare & \blacksquare \\ \square & \blacksquare & \blacksquare \\ \square & \blacksquare & \blacksquare \end{bmatrix} \text{ ou } \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix},$$

saibamos matematicamente definir se ele se assemelha mais com um 0 ou com um 1.

Visto que não conseguimos visualizar dados em dimensão superior a 3, parece uma escolha razoável escolher um valor entre 1 e 3 para a dimensão no qual gostaríamos de visualizar esses dados. Por conveniência, escolheremos um plano de 2 dimensões no qual represente da melhor maneira possível esses pontos em $\mathbb{R}^9$.

Esse plano será formado necessariamente por 2 vetores e então, como consequência, teremos que cada representação dos pontos em $\mathbb{R}^2$ será uma combinação linear desses vetores.

De maneira mais precisa, podemos dizer que queremos reduzir a dimensão dos dados de $\mathbb{R}^9$ para $\mathbb{R}^2$ de modo a preservar da melhor maneira possível a natureza entre os dados. Para isso, definimos que queremos encontrar um plano em $\mathbb{R}^9$ que minimiza a distância ao quadrado aos pontos que devem ser representados. Dessa forma, temos que, sendo a_i o ponto i, $\forall i \in \{1, \ldots, n\}$, n a quantidade de pontos e x a matriz com seus vetores-coluna sendo os vetores que geram o plano.

$$\begin{aligned} & \operatorname{argmin}_x \sum_{i=1}^{n} \operatorname{dist}(a_i, x)^2 \\ = \quad & \{ \text{ Demonstração feita na seção 1.4.1 } \} \\ & \operatorname{argmax}_x \frac{\|Ax\|^2}{\|x\|^2}. \end{aligned}$$

Ou seja, o problema de encontrar o melhor plano que representa esses dados se resume em encontrar $x \in \operatorname{argmax}_x \frac{\|Ax\|^2}{\|x\|^2}$.

1.4.3 Compressão de Dados

Um contador, que utiliza planilhas para armazenamento de diferentes informações numéricas, percebeu que suas planilhas contém uma quantidade enorme de redundância de dados. Por exemplo, a coluna que contém os ganhos g da empresa conforme os diferentes 12 meses dos ano j apresenta valores com um certo padrão repetido inúmeras vezes a cada ano.

O contador percebeu então que poderia reescrever essa matriz de maneira aproximada como sendo:

$$\begin{bmatrix} g_{1,1} & \cdots & g_{1,j} & \cdots & g_{1,12} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ g_{i,1} & \cdots & g_{i,j} & \cdots & g_{i,12} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ g_{n,1} & \cdots & g_{n,j} & \cdots & g_{n,12} \end{bmatrix} \approx \begin{bmatrix} g'_1 \\ \vdots \\ g'_i \\ \vdots \\ g'_n \end{bmatrix} \begin{bmatrix} m_1 & \ldots & m_j & \ldots & m_{12} \end{bmatrix}$$

$\forall i \in \{1, \ldots, n\}, j \in \{1, \ldots, 12\}$, sendo n a quantidade de anos que a empresa tem seus dados de ganho armazenados, $g_{i,j}$ sendo o ganho no ano i no mês j, g'_i sendo o ganho total que a

empresa teve no ano i e m_j sendo um fator que diz a distribuição do ganho total daquele ano conforme os meses.

Dessa forma, o contador percebeu que a matriz que antes ocupava um total de $n \times 12$ espaços, agora ocupa somente $n + 12$ espaços mas com a perda de certa precisão na informação.

Dessa forma, gostaríamos de encontrar um método matematicamente preciso de definir como melhor aproximar uma matriz como sendo o produto de duas matrizes de maneira que o máximo de informação possível seja mantida e que haja uma eficiente compressão de dados.

Se decompormos uma matriz $A_{n\times m}$ como sendo o produto de duas matrizes $B_{n\times k}$ e $C_{k\times m}$, precisamos que, para que haja compressão de dados, a soma do tamanho de ambas as matrizes B e C seja menor do que de A. Em outras palavras, $nk + km < nm$, ou seja, $k < \dfrac{nm}{n+m}$. Caso contrário, estaríamos armazenando uma matriz utilizando 2 matrizes ainda maiores que aparentemente não nos trás qualquer vantagem.

Nesse caso, gostaríamos de encontrar k vetores coluna para a matriz B e k vetores linha para a matriz C de tal forma que BC seja a melhor aproximação possível de posto k para a matriz A. Temos que, necessariamente, essa aproximação terá posto k pois não faz sentido colocar mais uma coluna em B para representar novos dados se esses dados já poderiam ser representados no subespaço gerado pelas colunas de B.

Além disso, a aproximação utilizada pelo contador trás um significado para cada posição dos vetores utilizados, enquanto cada posição das matrizes B e C não apresentam um significado tão claro, apesar de matematicamente ser o melhor critério possível utilizado de forma a aproximar o produto BC da matriz A.[1]

Podemos interpretar esse produto BC como sendo a projeção dos valores de matriz A no subespaço gerado pelos vetores coluna de B, onde C armazena os valores dessas projeções. Dessa forma, queremos determinar que vetores coluna x são esses que minimizam o somatório das distâncias ao quadrado de A até x.

$$\begin{aligned} & \operatorname{argmin}_x \sum_{i=1}^{n} \operatorname{dist}(a_i, x)^2 \\ = \quad & \{ \text{ Demonstração feita na seção 1.4.1 } \} \\ & \operatorname{argmax}_x \frac{\|Ax\|^2}{\|x\|^2}. \end{aligned}$$

Ou seja, o problema de encontrar as melhores matrizes B e C tal que seu produto melhor que representa esses dados se resume em encontrar $x \in \operatorname{argmax}_x \dfrac{\|Ax\|^2}{\|x\|^2}$.

[1]Interpretar as matrizes B e C não é trivial. Esse tipo de dificuldade faz parte da *análise de componentes principais*. O processo de obtenção dessas matrizes e a estrutura de seus elementos são discutidos no capítulo 2.4. No entanto, a interpretação de seus significados pode variar conforme o problema, não havendo uma solução geral para essa questão.

Parte 2

Truque, Troca de Variáveis e Fatoração

[...] It's in words that the magic is — Abracadabra, Open Sesame, and the rest — but the magic words in one story aren't magical in the next. The real magic is to understand which words work, and when, and for what; the trick is to learn the trick. [...] And those words are made from the letters of our alphabet: a couple-dozen squiggles we can draw with the pen. This is the key! And the treasure, too, if we can only get our hands on it! It's as if — as if the key to the treasure is the treasure!

John Barth, *Chimera*

Capítulo 2.1

Modelos Exatos

2.1.1 Truque

Queremos encontrar x tal que $Ax = b$. No entanto, quando A é uma matriz densa, essa tarefa pode ser complexa pois todas (ou quase todas) as posições do vetor x estão diretamente relacionadas com todas (ou quase todas) as posição do vetor b.

Veja que, se A fosse uma matriz triangular superior, o sistema poderia ser resolvido de forma simples e eficiente por meio da *substituição reversa*. Por exemplo, no caso de A apresentar dimensão (3×3):

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix},$$

perceba que é fácil escolher o valor de x_3 que faz com que $a_{33}x_3 = b_3$. De maneira análoga, após ter um valor para x_3, é simples escolher um valor para x_2 que satisfaça $a_{22}x_2 + a_{23}x_3 = b_2$ visto que é apenas resolver uma equação de uma variável (dado que já temos um valor para x_3) e assim sucessivamente.

Por exemplo, em um caso onde:

$$\begin{bmatrix} 7 & 4 & 1 \\ 0 & 2 & 3 \\ 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 21 \\ 12 \\ 10 \end{bmatrix},$$

É possível escolher $x_3 = 2$ para que, de fato, a igualdade da última linha do sistema seja satisfeita. De maneira recursiva, podemos agora determinar $x_2 = 3$ de tal forma que $2x_2 + 3x_3 = 12$ e, por fim, $x_3 = 1$ para que, de fato, $7x_1 + 4x_2 + 1x_3 = 21$. Assim, temos que é fácil encontrar o vetor x tal que $Ax = b$ quando A é uma matriz triangular superior.

2.1.2 Troca de Variáveis

Por meio da Eliminação Gaussiana, temos a garantia de que toda matriz A pode ser decomposta na formato $A = LU$ desde que seus vetores coluna sejam linearmente independentes, onde L é uma matriz triangular inferior e U é uma matriz triangular superior, ambas com diagonal não-nula.

Essa decomposição faz com que o cálculo de x tal que $Ax = b$ seja muito mais simples.

Teorema T1 (Inversibilidade de Matriz Triangular Inferior com Diagonal Não Nula)**.** *Se L é uma matriz triangular inferior de ordem n com diagonal não-nula então $\exists L^{-1}$ tal que $LL^{-1} = I$.*

Demonstração.

Caso Base: $n = 0$: A matriz L é vazia, que pode ser trivialmente inversível. Logo, o caso base é válido.

Hipótese de Indução: $\forall n \in N$ tal que $n < k$, $\exists L^{-1}$ tal que $LL^{-1} = I$.

Passo Indutivo: $n = k$;

$$\begin{bmatrix} a & 0 \\ \bar{u} & M \end{bmatrix} \begin{bmatrix} x_0 \\ \bar{x} \end{bmatrix} = \begin{bmatrix} y_0 \\ \bar{y} \end{bmatrix}$$

$\iff$ { Mudando a representação do sistema para *point-full* }

$$\begin{cases} ax_0 = y_0 \\ x_0\bar{u} + M\bar{x} = \bar{y} \end{cases}$$

$\iff$ { $a \neq 0$ pois a diagonal é não-nula e subtraindo $x_0 u$ na segunda equação }

$$\begin{cases} x_0 = \dfrac{y_0}{a} \\ M\bar{x} = \bar{y} - x_0\bar{u} \end{cases}$$

$\iff$ { Substituindo o valor de x_0 na segunda equação }

$$\begin{cases} x_0 = \dfrac{y_0}{a} \\ M\bar{x} = \bar{y} - \dfrac{y_0}{a}u \end{cases}$$

$\iff$ { Hipótese de Indução na segunda equação }

$$\begin{cases} x_0 = \dfrac{y_0}{a} \\ M^{-1}M\bar{x} = M^{-1}(\bar{y} - \dfrac{y_0}{a}u) \end{cases}$$

$\iff$ { Distributiva e $M^{-1}M = I$ }

$$\begin{cases} x_0 = \dfrac{y_0}{a} \\ \bar{x} = M^{-1}\bar{y} - \dfrac{y_0}{a}M^{-1}u \end{cases}$$

$\iff$ { Mudando a representação do sistema para *point-wise* }

$$\begin{bmatrix} x_0 \\ \bar{x} \end{bmatrix} = \begin{bmatrix} 1/a & 0 \\ \dfrac{M^{-1}\bar{u}}{a} & M^{-1} \end{bmatrix} \begin{bmatrix} y_0 \\ \bar{y} \end{bmatrix}$$

Veja que começamos com um sistema

$$\underbrace{\begin{bmatrix} a & 0 \\ \bar{u} & M \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} x_0 \\ \bar{x} \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} y_0 \\ \bar{y} \end{bmatrix}}_{y}$$

e terminamos com

$$\underbrace{\begin{bmatrix} x_0 \\ \bar{x} \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 1/a & 0 \\ \dfrac{M^{-1}\bar{u}}{a} & M^{-1} \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} y_0 \\ \bar{y} \end{bmatrix}}_{y},$$

de forma que $B = L^{-1}$. Ou seja, temos que

$$\begin{bmatrix} a & 0 \\ \bar{u} & M \end{bmatrix}^{-1} = \begin{bmatrix} 1/a & 0 \\ \dfrac{M^{-1}\bar{u}}{a} & M^{-1} \end{bmatrix},$$

assim, mostrando não só que existe uma inversa para a matriz L mas também como encontrar essa inversa. □

Como corolário desse teorema, temos também que é possível encontrarmos uma inversa para toda matriz U triangular superior com diagonal não nula.

Teorema T2 (Inversibilidade de Matriz Triangular Superior com Diagonal Não Nula)**.** *Se U é uma matriz triangular superior de ordem n com diagonal não-nula, então $\exists U^{-1}$ tal que $UU^{-1} = I$.*

Demonstração.

Pelo teorema T1, sabemos que U^{T} é inversível. Ou seja:

$$\begin{aligned} & U^{\mathsf{T}^{-1}} U^{\mathsf{T}} = I \\ = \quad & \{ \text{ Teorema T20 } \} \\ & (UU^{-1})^{\mathsf{T}} = I^{\mathsf{T}} \\ = \quad & \{ \ I^{\mathsf{T}} = I \ \} \\ & (U^{-1})^{\mathsf{T}} U^{\mathsf{T}} = I \\ = \quad & \{ \text{ Definição de inversa } \} \\ & U^{\mathsf{T}^{-1}} = (U^{-1})^{\mathsf{T}} \\ = \quad & \{ \text{ Tomando a transposta de ambos os lados } \} \\ & U^{-1} = (U^{\mathsf{T}^{-1}})^{\mathsf{T}} \end{aligned}$$

□

Se L e U são uma matriz triangulares com diagonal não nula, temos a garantia de que suas inversas existem, consequentemente, podemos, pelo teorema T3, reescrever o sistema $Ax = b$.

Teorema T3 (Troca de Variáveis de Modelos Exatos)**.** *Seja A uma matriz tal que $A = LU$, onde L é uma matriz triangular inferior e U é uma matriz triangular superior, ambas com diagonal não nula. Sejam x, b e c vetores tais que $c = L^{-1}b$. Se $Ax = b$, então $Ux = c$.*

Demonstração.

$$
\begin{aligned}
& Ax = b \\
\iff & \quad \{\ A = LU\ \} \\
& LUx = b \\
\iff & \quad \{\ \text{Teorema T1 e multiplicando por } L^{-1}\ \} \\
& L^{-1}LUx = L^{-1}b \\
\iff & \quad \{\ L^{-1}L = I\ \} \\
& Ux = L^{-1}b \\
\iff & \quad \{\ c = L^{-1}b\ \} \\
& Ux = c
\end{aligned}
$$

□

Dessa forma, temos que a solução x para o sistema $Ax = b$ é a mesma solução x do sistema $Ux = c$, sendo U uma matriz triangular superior e $c = L^{-1}b$.

2.1.3 Fatoração LU

2.1.3.1 Objetivo e Ideia da Fatoração LU

A decomposição LU consiste em fatorar uma matriz A em um produto de matrizes L e U, onde L é uma matriz triangular inferior e U é uma matriz triangular superior.

A vantagem dessa fatoração se evidencia no instante que se torna necessária a resolução de um sistema na forma $Ax = b$, bem como evidenciado na seção 2.1.2, onde o x que satisfaz $Ax = b$, pelo teorema T3, também satisfaz $Ux = L^{-1}b$.

A fatoração LU nos permite identificar o posto[1] de uma matriz com certa facilidade. Perceba que se A for decomposta como sendo o produto de duas matrizes triangulares sem quaisquer elementos nulos nas diagonais, então nenhuma dimensão da imagem dessa matriz está sendo perdida em relação ao domínio. Caso o contrário fosse verdadeiro, teríamos então que uma ou mais dimensões estão sendo perdidas, fazendo com que a matriz apresente determinante 0 e, portanto, não tenha inversa.

[1] O posto de uma matriz A é definido como sendo o número máximo de linhas ou colunas linearmente independentes de A. Em outras palavras, o posto pode ser interpretado como sendo a dimensão do espaço vetorial gerado pelas colunas da matriz A.

Uma peculiaridade dessa fatoração — que é exatamente o porquê de a utilizarmos — é o fato de que, se L for uma matriz de posto completo, então sua inversa necessariamente existe, o que nos garante a existência de L^{-1} e, se U tiver posto completo, temos a garantia de solução para o sistema $Ux = L^{-1}b$.

Ainda assim, mesmo com essas propriedades positivas acerca dessa decomposição, nos resta encontrar uma maneira de, dado uma matriz A, fatorá-la como sendo o produto de L e U.

2.1.3.2 Algoritmo: Eliminação Gaussiana

A decomposição LU pode ser obtida como resultado de um algoritmo chamado de Eliminação Gaussiana, que consiste em realizar diversas operações de linha em uma matriz de modo a eliminar os elementos abaixo dos pivôs da matriz.

Veja que, no caso de uma matriz A de dimensão (2×2) e de posto completo, ao multiplicar A por uma matriz L^{-1} que modifica a segunda linha de A usando a primeira, temos que:

$$\underbrace{\begin{bmatrix} 1 & 0 \\ \left(\dfrac{-a_{21}}{a_{11}}\right) & 1 \end{bmatrix}}_{L^{-1}} \underbrace{\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}}_{A} = \begin{bmatrix} a_{11} & a_{12} \\ \left(\dfrac{-a_{21}a_{11}}{a_{11}} + a_{21}\right) & \left(\dfrac{-a_{21}}{a_{11}}a_{12} + a_{22}\right) \end{bmatrix},$$

que, como $\left(\dfrac{-a_{21}a_{11}}{a_{11}} + a_{21}\right) = 0$, é o mesmo que:

$$\underbrace{\begin{bmatrix} 1 & 0 \\ \left(\dfrac{-a_{21}}{a_{11}}\right) & 1 \end{bmatrix}}_{L^{-1}} \underbrace{\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} a_{11} & a_{12} \\ 0 & \left(\dfrac{-a_{21}}{a_{11}}a_{12} + a_{22}\right) \end{bmatrix}}_{U}.$$

Por L^{-1} ser uma matriz triangular inferior com diagonal não nula, temos a garantia de que, pelo teorema T1, sua inversa L existe e, consequentemente, podemos fatorar A como:

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 1 & 0 \\ \left(\dfrac{a_{21}}{a_{11}}\right) & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} a_{11} & a_{12} \\ 0 & \left(\dfrac{-a_{21}}{a_{11}}a_{12} + a_{22}\right) \end{bmatrix}}_{U}.$$

No caso geral de uma matriz A de posto completo, temos:

$$\underbrace{\begin{bmatrix} a_{11} & \dots & a_{1i} & \dots & a_{1n} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ a_{i1} & \dots & a_{ii} & \dots & a_{in} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{ni} & \dots & a_{nn} \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 1 & & & & \\ \vdots & \ddots & & & \\ a_{i1} & \dots & 1 & & \\ \vdots & \iddots & \vdots & \ddots & \\ a_{n1} & \dots & a_{ni} & \dots & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} u_{11} & \dots & u_{1i} & \dots & u_{1n} \\ & \ddots & \vdots & \iddots & \vdots \\ & & u_{ii} & \dots & u_{in} \\ & & & \ddots & \vdots \\ & & & & u_{nn} \end{bmatrix}}_{U}.$$

2.1.4 Exercícios de Modelos Exatos

Exercício E8:

Use alguns dados da tabela abaixo para calcular cos(40) usando interpolação.

α	$30°$	$45°$	$60°$
sen α	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$
cos α	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$
tg α	$\frac{\sqrt{3}}{3}$	1	$\sqrt{3}$

Exercício E9:

Determine uma boa aproximação para $\log(3)_2$ usando interpolação por uma parábola.

Exercício E10:

Dado o PVC $y''(x) = 4x$ com $y(0) = 5$ e $y(10) = 20$

1. Monte o sistema linear que aproxima pelo método de diferenças finitas com $n = 4$ intervalos na discretização.
2. Monte o sistema linear que aproxima pelo método de diferenças finitas com $n = 10$ intervalos na discretização.
3. Resolva o sistema linear obtido no item anterior no Julia.
4. Use interpolação polinomial (pode ser com grau = 2 ou grau = 3) para descobrir $y(3.2345)$.
5. Descreva o pseudo-código dessa questão.

Exercício E11:

A caixa de um cereal para o café da manhã apresenta o número de calorias e as quantidades de proteínas, carboidratos e gordura contidos em uma porção do cereal. As quantidades para dois cereais conhecidos são dadas a seguir: uma porção do cereal 1 contém 50 calorias, 20g de carboidratos e 2g de gordura e uma porção do cereal 2 contém 100 calorias, 15g de carboidratos e 1g de gordura.

Determine se é possível misturar esses cereais de modo a encontrar uma mistura que apresente 175 calorias, 45g de carboidratos e 4g de gordura. Se for possível, determine quantas porções de cada cereal são necessárias para atingir esses valores.

Exercício E12:

[Idades] Sabemos relações da idade entre 5 amigos, Alberto, Bernardo, Carol, Denise e Ernesto. Sabemos que Bernardo é 5 anos mais velho que Alberto, 15 anos mais velho que Denise e 3 anos mais velho que Carol. Denise é 7 anos mais velha que Ernesto e 10 anos mais nova que Alberto.

1. Represente a situação acima desenhando um grafo.
2. Represente a situação acima com uma tabela (ou matriz).

Capítulo 2.2

Modelos Aproximados

2.2.1 Truque

Queremos encontrar $\text{argmin}_x \ \|Ax - b\|$. No entanto, quando A é uma matriz densa, essa tarefa pode ser complexa pois a escolha de x que aproxima Ax de b não é direta.

Caso A seja uma matriz triangular superior com linhas inferiores nulas, o sistema pode ser resolvido de forma mais eficiente via substituição reversa. Por exemplo, no caso de um sistema com uma matriz A de dimensão (5×3), a minimização do erro envolve a seguinte norma:

$$\left\| \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} - \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} \right\|$$

Note que, independentemente dos valores escolhidos para x_1, x_2 e x_3, a diferença $Ax - b$ sempre conterá os valores de b_4 e b_5, pois não há graus de liberdade no sistema que permitam anulá-los.

Por outro lado, os valores de b_1, b_2 e b_3 podem ser minimizados ao escolher x_1, x_2 e x_3 de maneira adequada via substituição reversa. Assim, a minimização da norma do erro ocorre nos primeiros três elementos do vetor residual, enquanto os últimos dois não podem ser ajustados pela escolha de x.

Por exemplo, em um caso onde:

$$\left\| \begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} - \begin{bmatrix} 8 \\ 12 \\ 9 \\ 2 \\ 4 \end{bmatrix} \right\|,$$

temos que, ao realizar a substituição reversa na parcela superior do sistema, o vetor $x = \begin{bmatrix}1 & 2 & 1\end{bmatrix}^\mathsf{T}$ faz com que a diferença $Ax - b$ deixe o vetor $\begin{bmatrix}0 & 0 & 0 & 2 & 4\end{bmatrix}^\mathsf{T}$ como resíduo, sendo esse, então, o erro da aproximação.

2.2.2 Troca de Variáveis

Para simplificar o problema $\text{argmin}_x \ \|Ax - b\|$, buscamos uma matriz Q^T que modifique A e b de maneira a tornar a determinação de x mais simples.

Queremos garantir que a norma de $Ax - b$ permaneça inalterada ao multiplicá-la por Q^T. Uma maneira de assegurar essa propriedade é escolher Q^T como uma matriz ortogonal. Pelo teorema T4, a multiplicação de um vetor por uma matriz ortogonal faz com que a sua norma seja preservada, nesse caso, garantindo que a norma do vetor resíduo $(Ax - b)$ não se altere.

Por meio de algoritmos como Gram-Schmidt, reflexões de Householder ou rotações de Givens, temos a garantia de que qualquer matriz A pode ser decomposta no formato $A = QR$, onde Q é uma matriz ortogonal e R é uma matriz composta por uma matriz triangular superior $\bar{R}$ e uma matriz 0 com linhas zeradas.

Essa decomposição simplifica significativamente o problema: enquanto o cálculo de $\text{argmin}_x \ \|Ax - b\|$ tende a ser complicado, a equação equivalente $\text{argmin}_x \ \left\|Rx - Q^\mathsf{T}b\right\|$, devido à estrutura triangular de R, é muito mais fácil de se resolver.

Teorema T4 (Preservação da Norma por Matrizes Ortogonais)**.** *Seja Q uma matriz e v um vetor. Se Q é uma matriz ortogonal então $\|Qv\| = \|v\|$.*

Demonstração.

$$
\begin{aligned}
& \|Qv\| \\
= \quad & \{ \text{ Definição } \} \\
& \sqrt{(Qv)^\mathsf{T} Qv} \\
= \quad & \{ \text{ Teorema T20 } \} \\
& \sqrt{v^\mathsf{T} Q^\mathsf{T} Qv} \\
= \quad & \{ \text{ Definição } \} \\
& \sqrt{v^\mathsf{T} Iv} \\
= \quad & \{ \ Iv = v \ \} \\
& \sqrt{v^\mathsf{T} v} \\
= \quad & \{ \text{ Definição } \} \\
& \|v\|
\end{aligned}
$$

Portanto, provamos que, se Q é uma matriz ortogonal, então $\|Qv\| = \|v\|$. □

Teorema T5 (Troca de Variáveis de Modelos Aproximados)**.** *Seja A uma matriz tal que $A = QR$, sendo Q uma matriz ortogonal e R uma matriz composta por uma matriz $\bar{R}$ triangular*

superior e por uma matriz 0 *composta de zeros. Sejam* x*,* b *e* c *vetores tais que* $c = Q^\mathsf{T} b$*, onde* c *é composto por vetores* c_1 *e* c_2*. Então:*

- $argmin_x \ \|Ax - b\| = argmin_x \ \left\|\bar{R}x - c_1\right\|$.
- $min_x \ \|Ax - b\| = \|c_2\|$.

Demonstração.

$$
\begin{aligned}
& \|Ax - b\| \\
= & \quad \{ \text{ Teorema T4 } \} \\
& \left\|Q^\mathsf{T}(Ax - b)\right\| \\
= & \quad \{ \text{ Distributiva } \} \\
& \left\|Q^\mathsf{T}Ax - Q^\mathsf{T}b\right\| \\
= & \quad \{ \ R = Q^\mathsf{T}A \text{ e } c = Q^\mathsf{T}b \ \} \\
& \|Rx - c\|
\end{aligned}
$$

Calculando $\text{argmin}_x \ \|Ax - b\|$:

$$
\begin{aligned}
& \text{argmin}_x \ \|Ax - b\| \\
= & \quad \{ \ \|Ax - b\| = \|Rx - c\| \ \} \\
& \text{argmin}_x \ \|Rx - c\| \\
= & \quad \{ \text{ Teorema T31 } \} \\
& \text{argmin}_x \ \|Rx - c\|^2 \\
= & \quad \{ \text{ Propriedade } \} \\
& \text{argmin}_x \ \left\|\bar{R}x - c_1\right\|^2 + \|0x - c_2\|^2 \\
= & \quad \{ \text{ Teorema T32 } \} \\
& \text{argmin}_x \ \left\|\bar{R}x - c_1\right\|^2 \\
= & \quad \{ \text{ Teorema T31 } \} \\
& \text{argmin}_x \ \left\|\bar{R}x - c_1\right\|
\end{aligned}
$$

Calculando $\text{min}_x \ \|Ax - b\|$:

$$
\begin{aligned}
& \text{min}_x \ \|Ax - b\| \\
= & \quad \{ \ \|Ax - b\| = \|Rx - c\| \ \} \\
& \text{min}_x \ \|Rx - c\| \\
= & \quad \{ \ \|Rx - c\| = \sqrt{\|Rx - c\|^2} \text{ pois } \|Rx - c\| \geq 0 \ \} \\
& \text{min}_x \ \sqrt{\|Rx - c\|^2} \\
= & \quad \{ \text{ Propriedade } \}
\end{aligned}
$$

$$
\begin{aligned}
& \min_x \sqrt{\left\|\bar{R}x - c_1\right\|^2 + \|0x - c_2\|^2} \\
= \quad & \{\ \bar{R}x - c_1 \text{ tem solução em } x \text{ pelo teorema T2 } \} \\
& \min_x \sqrt{\|c_2\|^2} \\
= \quad & \{\ \sqrt{\|c_2\|^2} = \|c_2\| \text{ pois } \|c_2\| \geq 0 \ \} \\
& \min_x \ \|c_2\| \\
= \quad & \{\ \text{Mínimo de uma constante é a própria constante } \} \\
& \|c_2\|
\end{aligned}
$$

□

Visto que A apresenta dimensão $(n \times k)$, temos que Q e R apresentam dimensões $(n \times n)$ e $(n \times k)$, respectivamente, e como R é triangular superior e $n \geq k$, temos que, R apresenta $n - k$ linhas inteiras com o valor 0.

Como $\bar{R}$ é uma matriz triangular superior e, pelo teorema T2, $\bar{R}$ é necessariamente invertível e, portanto, como apresenta posto completo, é possível encontrar um vetor x tal que $\bar{R}x = c_1$.

Por fim, descobrimos como encontrar o vetor x que minimiza $\|Ax - b\|$ e também a determinar o erro absoluto da aproximação.

2.2.3 Fatoração QR

2.2.3.1 Ideia e Objetivo da Fatoração QR

A decomposição QR consiste em fatorar uma matriz A em um produto de matrizes Q e R, sendo Q uma matriz ortogonal e R uma matriz triangular superior.

A vantagem da fatoração se dá no fato da matriz Q ser ortogonal, fazendo com que, pelo teorema T4, a norma de qualquer vetor x aplicado em A seja igual a norma de R aplicado em x.

Dessa maneira, temos que, como provado em T3, encontrar x que satisfaz $x \in \operatorname{argmin}_x \|Ax - b\|^2$ também satisfaz $x \in \operatorname{argmin}_x \left\|Rx - Q^\mathsf{T} b\right\|^2$.

Além disso, perceber que fatorar uma matriz A como o produto de uma matriz ortogonal Q e uma matriz triangular superior R é equivalente a dizer que toda matriz A pode ser expressa como o resultado de uma transformação ortogonal (uma rotação ou reflexão) seguida de uma transformação triangular superior.

Como Q é uma matriz ortogonal, Q é necessariamente uma matriz quadrada e define uma transformação linear bijetora. Ou seja, a transformação associada a Q preserva a dimensão do espaço vetorial e não colapsa vetores distintos na mesma imagem. Dessa forma, escrever A como QR é equivalente a dizer que as colunas de A são combinações lineares dos vetores ortonormais que compõem as colunas de Q com os coeficientes dados pela matriz R.

Contudo, há casos onde a matriz R apresenta valores nulos em sua diagonal principal. Isso ocorre quando a matriz A apresenta colunas linearmente dependentes. Intuitivamente, se uma coluna de A puder ser escrita como combinação linear das outras, então essa solução não acrescenta nova direção no espaço vetorial gerado, fazendo com que seu coeficiente em R seja nulo.

Ou seja, a decomposição QR também reflete o fato a matriz A apresentar posto completo ou não. No caso de A não apresentar posto completo e R conter linhas zeradas, bem como ilustrado em 2.2.1, isso evidencia que parte do espaço vetorial gerado pelas colunas de A é redundante — isto é, algumas colunas não contribuem com novas direções no subespaço gerado.

Nesse contexto, a decomposição QR não apenas facilita a resolução do modelo de aproximação, como também oferece uma interpretação geométrica e estrutural da matriz A: as colunas ortonormais de Q representam as direções relevantes do espaço gerado, enquanto os elementos não nulos de R indicam a contribuição efetiva de cada uma dessas direções.

2.2.3.2 Algoritmo de Gram-Schmidt

O algoritmo de Gram-Schmidt consiste em encontrar uma base ortonormal para um dado conjunto de vetores. De maneira mais precisa, dado um conjunto de vetores linearmente independentes $A = \{a_1, \ldots, a_n\}$, o algoritmo gera um conjunto de vetores ortogonais $U = \{u_1, \ldots, u_n\}$ que gera o mesmo subespaço n-dimensional.

A aplicação do algoritmo de Gram-Schmidt no caso de matrizes resulta na decomposição QR. Essa decomposição pode ser vista como uma simples reescrita dos vetores coluna de uma matriz como combinação linear de vetores linearmente independentes do mesmo subespaço.

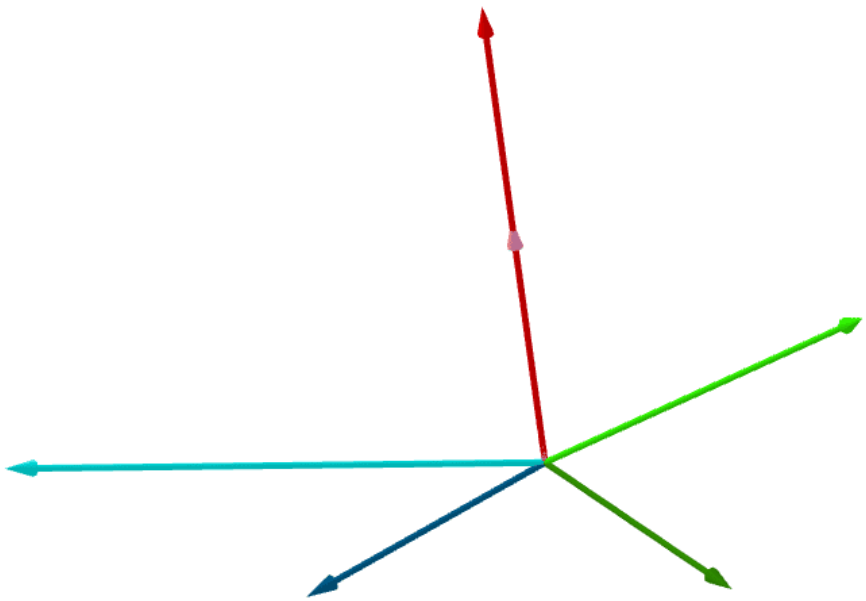


Figura F27: Exemplo de 3 vetores em $\mathbb{R}^3$ gerados pelo processo de Gram-Schmidt.

Nesse caso, a matriz Q contém esses vetores ortogonais (normalizados), e a matriz R contém os coeficientes das combinações lineares dos vetores coluna de Q que formam os vetores coluna da matriz original.

A matriz R necessariamente será triangular superior. O primeiro vetor coluna de A é utilizado para formar o primeiro vetor coluna de Q, que é simplesmente uma normalização do primeiro vetor de A.

O segundo vetor coluna de Q deve ser ortogonal ao primeiro vetor coluna de Q, então ele será obtido pela projeção do segundo vetor de A sobre o primeiro vetor de Q, e removendo essa projeção para, assim, garantirmos que o componente restante seja perpendicular ao primeiro vetor de Q, seguido por uma normalização. Este processo continua para os demais vetores coluna de A, cada um sendo ortogonalizado em relação aos vetores já formados de Q.

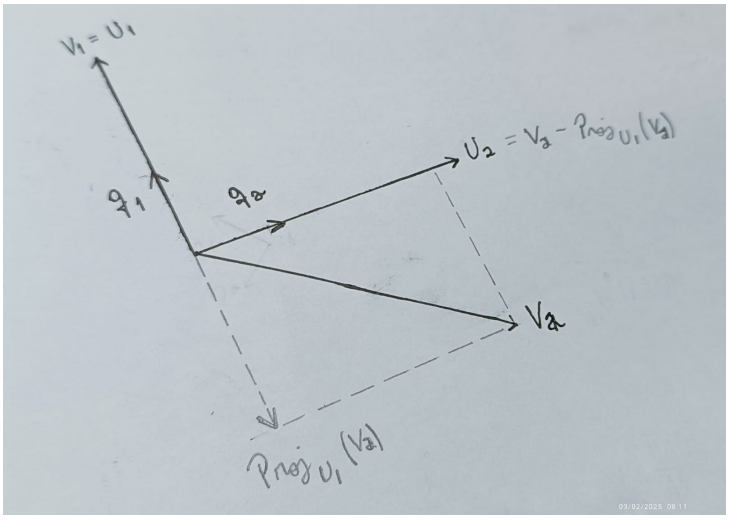


Figura F28: Exemplo de 2 vetores em $\mathbb{R}^2$ gerados pelo processo de Gram-Schmidt.

Veja que, no caso de 2 vetores em $\mathbb{R}^2$, é possível visualizar bem e intuitivamente encontrar a fatoração QR para uma matriz A. Nesse caso, temos que:

$$\begin{cases} u_1 = a_1, \\ \\ u_2 = a_2 - \dfrac{\langle a_2 \,|\, u_1 \rangle \, u_1}{\|u_1\|}, \end{cases}$$

Mas veja que u_1 e u_2 não apresentam norma 1. Portanto, declaramos:

$$q_i = \frac{u_i}{\|u_i\|} \quad , \quad \forall i \in \{1, 2\}.$$

Dessa forma, temos

$$\begin{cases} q_1 = \dfrac{u_1}{\|u_1\|}, \\ \\ q_2 = \dfrac{u_2}{\|u_2\|}, \end{cases}$$

que resulta em

$$\begin{cases} q_1 = \dfrac{u_1}{\|u_1\|}, \\ \\ q_2 = \dfrac{u_2}{\|u_2\|} = \dfrac{a_2 - \langle u_1 \,|\, a_2 \rangle \, q_1}{\|u_2\|}, \end{cases}$$

que é o mesmo que

$$\begin{cases} a_1 = \|u_1\| \, q_1, \\ \\ a_2 = \|u_2\| \, q_2 + \langle u_1 \,|\, v_2 \rangle \, q_1 \end{cases}$$

que, finalmente, pode ser reescrito como

$$\begin{bmatrix} | & | \\ a_1 & a_2 \\ | & | \end{bmatrix} = \begin{bmatrix} | & | \\ q_1 & q_2 \\ | & | \end{bmatrix} \begin{bmatrix} \|u_1\| & \langle u_1 \,|\, v_2 \rangle \\ 0 & \|u_2\| \end{bmatrix}.$$

No caso geral, temos:

$$\underbrace{\begin{bmatrix} | & & | & & | \\ a_1 & \dots & a_i & \dots & a_n \\ | & & | & & | \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} | & & | & & | \\ q_1 & \dots & q_i & \dots & q_n \\ | & & | & & | \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} \|u_1\| & \dots & \langle u_1 \,|\, v_i \rangle & \dots & \langle u_1 \,|\, v_n \rangle \\ & \ddots & \vdots & \ddots & \vdots \\ & & \|u_i\| & \dots & \langle u_i \,|\, v_n \rangle \\ & & & \ddots & \vdots \\ & & & & \|u_n\| \end{bmatrix}}_{R}.$$

2.2.4 Extensões

2.2.4.1 Resolução de Modelos Aproximados por Derivação

Na seção 1.2.2 vimos que uma possível interpretação para um sistema no formato $Ax \approx b$ é que Ax é uma combinação linear com coeficientes x_i de cada vetor coluna fr índice i da matriz

A. Também vimos que uma possível visualização para essa interpretação em $\mathbb{R}^3$ é representada pelas figuras F29 e F30.

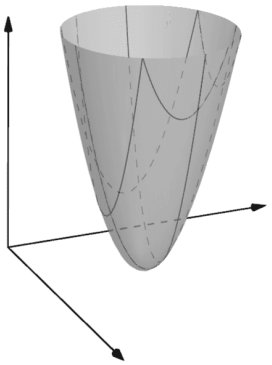


Figura F29: Representação do erro do sistema sistema $Ax \approx b$.

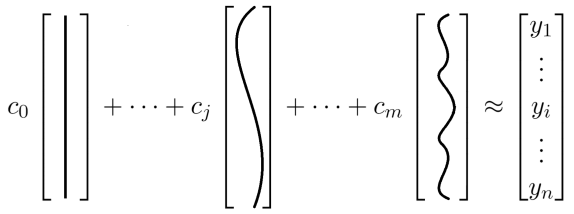


Figura F30: Representação do sistema $Ax \approx b$.

Por meio dessas imagens podemos intuitivamente acreditar que uma possível forma de encontrar o vetor Ax no espaço gerado pelos vetores coluna de A que melhor aproxima o vetor b é por meio de uma projeção ortogonal do vetor b nesse espaço. Além disso, por essa projeção ser ortogonal, temos então que o vetor do erro definido como $Ax - b$ é perpendicular à esse espaço. Dessa maneira, temos que o produto interno de cada vetor coluna de A com o vetor $Ax - b$ é igual a zero. Com isso, temos que

$$
\begin{aligned}
& \langle A_i \,|\, Ax - b \rangle = 0 \ \forall i \in \{1, \ldots, n\} \\
\iff & \quad \{ \text{ Reescrevendo como produto matriz-vetor } \} \\
& A^\mathsf{T}(Ax - b) = 0 \\
\iff & \quad \{ \text{ Distributiva } \} \\
& A^\mathsf{T}Ax - A^\mathsf{T}b = 0 \\
\iff & \quad \{ \text{ Somando } A^\mathsf{T}b \text{ de ambos os lados } \} \\
& A^\mathsf{T}Ax = A^\mathsf{T}b
\end{aligned}
$$

Ou seja, teríamos que $x \in \operatorname{argmin}_x \ \|Ax - b\|$ satisfaz $x \in \{x \mid A^\mathsf{T}Ax = A^\mathsf{T}b\}$. Ainda assim, isso não conta como prova matemática e, portanto, uma maneira formal de chegar nesse resultado seria a partir do cálculo do valor de x que minimize o erro $\|Ax - b\|$.

Teorema T6. *$x \in argmin_x \ \|Ax - b\| \iff x$ satisfaz $A^\mathsf{T}Ax = A^\mathsf{T}b$.*

Demonstração.

$$
\begin{aligned}
& x \in \operatorname{argmin}_x \ \|Ax - b\| \\
\iff & \quad \{ \text{ Teorema T31 } \}
\end{aligned}
$$

$$
\begin{aligned}
& x \in \operatorname{argmin}_x \ \|Ax - b\|^2 \\
\iff & \quad \{ \text{ Definição } \} \\
& x \in \operatorname{argmin}_x \ (Ax - b)^\mathsf{T}(Ax - b) \\
\iff & \quad \{ \text{ Distributiva } \} \\
& x \in \operatorname{argmin}_x \ (Ax)^\mathsf{T}Ax - (Ax)^\mathsf{T}b - (Ax)^\mathsf{T}b - b^\mathsf{T}b \\
\iff & \quad \{ \text{ Somando termos iguais } \} \\
& x \in \operatorname{argmin}_x \ (Ax)^\mathsf{T}Ax - 2(Ax)^\mathsf{T}b - b^\mathsf{T}b \\
\iff & \quad \{ \text{ Teorema T32 } \} \\
& x \in \operatorname{argmin}_x \ (Ax)^\mathsf{T}Ax - 2(Ax)^\mathsf{T}b \\
\iff & \quad \{ \text{ Teorema T20 } \} \\
& x \in \operatorname{argmin}_x \ x^\mathsf{T}A^\mathsf{T}Ax - 2x^\mathsf{T}A^\mathsf{T}b \\
\iff & \quad \{ \text{ Derivando e igualando a 0 para encontrar o ponto de mínimo } \} \\
& \nabla_x \ x^\mathsf{T}A^\mathsf{T}Ax - 2x^\mathsf{T}A^\mathsf{T}b = 0 \\
\iff & \quad \{ \text{ Teoremas T39 e T40 } \} \\
& 2A^\mathsf{T}Ax - 2A^\mathsf{T}b = 0 \\
\iff & \quad \{ \text{ Somando } 2A^\mathsf{T}b \text{ de ambos os lados } \} \\
& 2A^\mathsf{T}Ax = 2A^\mathsf{T}b \\
\iff & \quad \{ \text{ Divindo por 2 de ambos os lados } \} \\
& A^\mathsf{T}Ax = A^\mathsf{T}b
\end{aligned}
$$

□

Dessa forma, temos que, ao derivar a fórmula do erro a qual queremos minimizar é igualar a 0, achamos que o mínimo da função ocorre quando $A^\mathsf{T}Ax = A^\mathsf{T}b$.

2.2.4.2 Regressão Não-linear

Vimos nas seções 1.2.2 e 2.2.2 que resolver uma regressão polinomial equivale a encontrar um vetor x tal que

$$x \in \operatorname{argmin}_x \ \|Ax - b\|,$$

onde A é uma matriz de Vandermonde, b é um vetor contendo as coordenadas no eixo y dos n pontos do processo, e x é o vetor cujas entradas representam os coeficientes do polinômio resultante da regressão.

Caso quiséssemos aumentar o grau da função resultante do processo, era apenas inserir mais uma coluna na matriz A e um coeficiente no vetor x. Para o caso de funções não lineares como, por exemplo,

$$f(x) = e^{c_2 x},$$
$$g(x) = c_1 e^{c_2 x}, \text{ ou}$$
$$h(x) = c_1 e^{c_2 x} + c_3,$$

não existe uma forma genérica de tratar visto que essas funções apresentam diferentes formatos sem um padrão evidente.

Para que possamos realizar o processo de regressão em um conjunto de dados com um modelo não linear, é preciso que possamos manipular o sistema de equações resultante do problema com o intuito que poder expressar o modelo que contém funções como e^x ou $\log(()\,x)$ na forma matriz-vetor.

Por exemplo, escolhemos um modelo genérico como

$$f(x) = c_1 e^{c_2 x} + c$$

e, dados n pontos (x_1, y_1), ..., (x_i, y_i), ..., (x_n, y_n), $\forall i \in \{1, \ldots, n\}$, temos que $f(x_i) \approx y_i$.

Veja então que temos o sistema de equações:

$$\begin{aligned} c_1 e^{c_2 x_1} + c &\approx y_1, \\ &\vdots \\ c_1 e^{c_2 x_i} + c &\approx y_i, \\ &\vdots \\ c_1 e^{c_2 x_n} + c &\approx y_n. \end{aligned}$$

Como c é uma constante, podemos subtraí-la de ambos os lados, assim obtendo o sistema

$$\begin{aligned} c_1 e^{c_2 x_1} &\approx y_1 - c, \\ &\vdots \\ c_1 e^{c_2 x_i} &\approx y_i - c, \\ &\vdots \\ c_1 e^{c_2 x_n} &\approx y_n - c. \end{aligned}$$

Veja então que é conveniente computar o ln de ambos os lados visto que na esquerda das equações temos uma função exponencial e na direita das equações temos números. Dessa forma, obtendo o sistema

$$\begin{aligned} \ln\left(c_1 e^{c_2 x_1}\right) &\approx \ln\left(y_1 - c\right), \\ &\vdots \\ \ln\left(c_1 e^{c_2 x_i}\right) &\approx \ln\left(y_i - c\right), \\ &\vdots \\ \ln\left(c_1 e^{c_2 x_n}\right) &\approx \ln\left(y_n - c\right). \end{aligned}$$

Que pelas propriedades do logaritmo $\ln(ab) = \ln(a) + \ln(b)$ e $\ln(e^x) = x$, pode ser reescrito como

$$\begin{aligned}
\ln(c_1) + c_2 x_1 &\approx \ln(y_1 - c), \\
&\vdots \\
\ln(c_1) + c_2 x_i &\approx \ln(y_i - c), \\
&\vdots \\
\ln(c_1) + c_2 x_n &\approx \ln(y_n - c).
\end{aligned}$$

Se considerarmos a troca de variáveis $\bar{c_1} = \ln(c_1)$, então temos um sistema que aparenta ser linear

$$\begin{aligned}
\bar{c_1} + c_2 x_1 &\approx \ln(y_1 - c), \\
&\vdots \\
\bar{c_1} + c_2 x_i &\approx \ln(y_i - c), \\
&\vdots \\
\bar{c_1} + c_2 x_n &\approx \ln(y_n - c),
\end{aligned}$$

que, finalmente, pode ser reescrito como

$$\underbrace{\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} \bar{c_1} \\ c_2 \end{bmatrix}}_{x} \approx \underbrace{\begin{bmatrix} \ln(y_1 - c) \\ \vdots \\ \ln(y_i - c) \\ \vdots \\ \ln(y_n - c) \end{bmatrix}}_{b}.$$

Podemos definir o erro da aproximação como sendo $\|Ax - b\|$ e, consequentemente, resolver o sistema exatamente como fizemos na seção 2.2.2. Ainda assim, é importante levar duas coisas em consideração:

- o valor de $\bar{c_1}$ encontrado não corresponde ao valor de c_1 desejado na função $f(x) = c_1 e^{c_2 x} + c$. Para encontrarmos c_1, nesse caso, bastaria apenas considerar que $c_1 = e^{\bar{c_1}}$ e, dessa forma, finalmente obteríamos os coeficientes corretos da função f.
- Por mais que os valores de $\bar{c_1}$ e c_2 encontrados sejam os componentes do vetor x que minimizam o erro $\|Ax - b\|$, não temos qualquer garantia de que c_1 e c_2 sejam, de fato, os coeficientes que minimizem a soma do quadrado das distâncias verticais dos pontos e da função resultante.

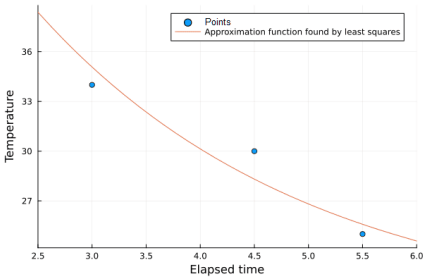


Figura F31: Exemplo de aproximação para dados pontos com uma função exponencial.

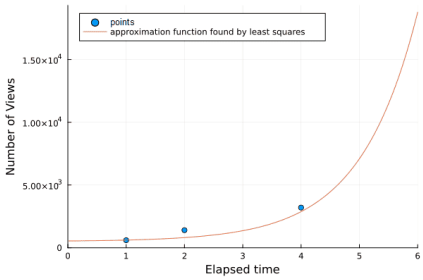


Figura F32: Outro exemplo de aproximação para dados pontos com uma função exponencial.

A ideia por trás dessa resolução para regressões não lineares é modificar o sistema de equações de forma que o sistema possa ser representado por um sistema matriz-vetor, como se fosse linear, e, consequentemente, podemos resolvê-lo como nos casos lineares. Após ter o sistema resolvido, modificamos a solução do sistema de forma com que as modificações feitas na função para que ela pudesse "se tornar linear" sejam desfeitas.

2.2.4.3 Regressão Logística

Seja Ω um conjunto de dado tais que Ω é bi-particionado em 2 subconjuntos A e B. Dado uma nova informação, gostaríamos de estimar em qual subconjunto é mais provável que ela esteja.

Podemos reduzir a dimensão desses dados para $\mathbb{R}^1$, por exemplo, conforme visto em 2.4.2, e adicionar uma segunda coordenada que serve como uma espécie de indicativo de qual subconjunto aquela informação pertence. Por exemplo, podemos colocar como segunda coordenada dos elementos de A o valor 0 e de B o valor 1 para que, dessa forma, possamos, por meio de uma regressão, encontrar uma curva que aproxime os pontos.

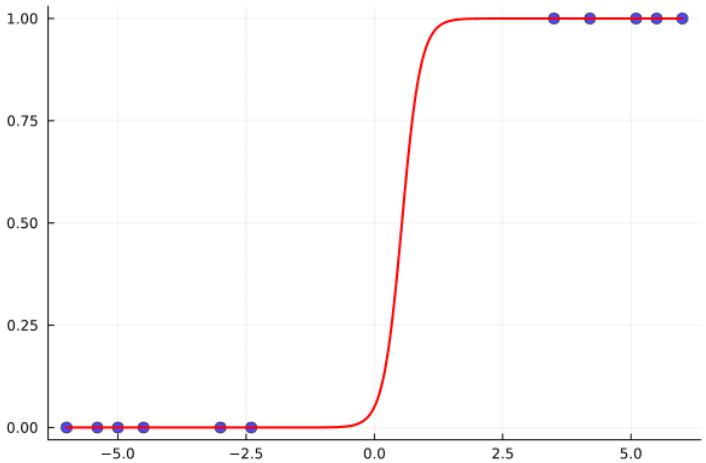


Figura F33: Exemplo da curva gerada pelo processo de regressão logística.

Veja que uma função capaz de aproximar pontos divididos em alturas 0 e 1 é

$$p(x) = \frac{e^{c_0+c_1x}}{1+e^{c_0+c_1x}}$$

visto que

$$\lim_{x\to-\infty} \frac{e^{c_0+c_1x}}{1+e^{c_0+c_1x}} = \begin{cases} 1 & \text{se } c_1 < 0, \\ \dfrac{e^{c_0}}{1+e^{c_0}} & \text{se } c_1 = 0, \\ 0 & \text{se } c_1 > 0, \end{cases}$$

e

$$\lim_{x\to\infty} \frac{e^{c_0+c_1x}}{1+e^{c_0+c_1x}} = \begin{cases} 0 & \text{se } c_1 < 0, \\ \dfrac{e^{c_0}}{1+e^{c_0}} & \text{se } c_1 = 0, \\ 1 & \text{se } c_1 > 0. \end{cases}$$

Indiferentemente de A apresentar segunda coordenada 0 ou 1, por estarmos utilizando uma regressão, os valores de c_0 e c_1 se ajustarão conforme o formato da função para aproximar os pontos de A e B. Sendo assim, não precisamos nos preocupar com restrição no valor de c_1.

O problema de definir o erro como sendo $\|Ax - b\|$ ou então

$$\sqrt{(f(x_1)-y_1)^2 + \cdots + (f(x_i)-y_i)^2 + \cdots + (f(x_n)-y_n)^2},$$

que é equivalente, é que, por mais que esse erro seja o que gostaríamos de minimizar, a manipulação algébrica até chegar no resultado desejado pode não fazer sentido visto que ao derivar a equação e igualar a 0 teremos uma derivada não linear por conta da função $p(x)$ e também não parece fazer muito sentido o cálculo da probabilidade ao quadrado. Sendo assim, faremos uma estimativa para esse procedimento por máxima verossimilhança.

Queremos que os valores de alguns $p(x_i)$ sejam altos a depender de qual conjunto atribuímos os valores de 0 e 1, tal que (1 ou 0) seja escolhido de acordo com a segunda coordenada que atribuímos ao ponto x_i. Sendo assim, temos:

$$\begin{aligned}
& \operatorname{argmax}_{c_0,c_1} \prod_{y_i=1} p(x_i) \cdot \prod_{y_i=0} (1-p(x_i)) \\
= \quad & \{ \text{ Teorema T34 } \} \\
& \operatorname{argmax}_{c_0,c_1} \log\left(\prod_{y_i=1} p(x_i) \cdot \prod_{y_i=0} (1-p(x_i)) \right) \\
= \quad & \{ \log(ab) = \log(a) + \log(b) \ \} \\
& \operatorname{argmax}_{c_0,c_1} \sum_{y_i=1} \log(p(x_i)) + \sum_{y_i=0} \log(1-p(x_i)) \\
= \quad & \{ \text{ Derivando em } c_1 \ \}
\end{aligned}$$

$$
\begin{aligned}
& \frac{\partial}{\partial c_1}\text{argmax}_{c_0,c_1} \sum_{y_i=1} \log\left(p(x_i)\right) + \sum_{y_i=0} \log\left(1-p(x_i)\right) \\
= \quad & \{ \text{ Cálculo da derivada } \} \\
& \sum_{i=0}^{n} \frac{1}{p(x_i)} p_{c_1}(x_i) x_i - \sum_{i=0}^{n} \frac{1}{1-p(x_i)} p_{c_i}(x_i) x_i
\end{aligned}
$$

Mas veja que $p_{c_1}(x_i)$ é

$$
\begin{aligned}
& p_{c_1}(x_i) = \left(\frac{e^{c_0+c_1x}}{1+e^{c_0+c_1x}}\right)_{c_i} \\
= \quad & \{ \text{ Cálculo da derivada } \} \\
& \frac{e^{c_0+c_1x}(1+e^{c_0+c_1x}) - e^{c_0+c_1x}e^{c_0+c_1x}}{(e^{c_0+c_1x})^2} \\
= \quad & \{ \text{ Fatoração do termo comum } e^{c_0+c_1x} \ \} \\
& \frac{e^{c_0+c_1x}(1+e^{c_0+c_1x} - e^{c_0+c_1x})}{(1+e^{c_0+c_1x})^2} \\
= \quad & \{ \text{ Simplificação } \} \\
& \frac{e^{c_0+c_1x}}{(1+e^{c_0+c_1x})^2} \\
= \quad & \{ \text{ Reescrita do denominador } \} \\
& \frac{e^{c_0+c_1x}}{(1+e^{c_0+c_1x})} \frac{1}{(1+e^{c_0+c_1x})} \\
= \quad & \{ \text{ Definição de } p(x) \ \} \\
& p(x_i)(1-p(x_i))
\end{aligned}
$$

logo:

$$
\begin{aligned}
& \sum_{y_i=1} \frac{1}{p(x_i)} p(x_i)(1-p(x_i)) x_i - \sum_{y_i=0} \frac{1}{1-p(x_i)} p(x_i)(1-p(x_i)) x_i \\
= \quad & \{ \text{ Simplificando os fatores comuns } \} \\
& \sum_{y_i=1} (1-p(x_i)) x_i - \sum_{y_i=0} p(x_i) x_i \\
= \quad & \{ \text{ Reescrevendo as somas como somatórios sobre todos os dados } \} \\
& \sum_{i=0}^{n} y_i (1-p(x_i)) x_i - \sum_{i=0}^{n} (1-y_i) p(x_i) x_i \\
= \quad & \{ \text{ Distribuindo os termos dentro da soma } \} \\
& \sum_{i=0}^{n} y_i (1-p(x_i)) x_i - (1-y_i) p(x_i) x_i \\
= \quad & \{ \text{ Expansão dos produtos dentro da soma } \}
\end{aligned}
$$

$$
\begin{aligned}
& \sum_{i=0}^{n}(y_i - y_i p(x_i) - p(x_i) + y_i p(x_i))x_i \\
= & \quad \{ \text{ Simplificação } \} \\
& \sum_{i=0}^{n}(y_i - p(x_i))x_i
\end{aligned}
$$

e, para o cálculo do gradiente em c_0:

$$
\begin{aligned}
& \frac{\partial}{\partial c_0}\left(\sum_{y_i=1}\log\left(p(x_i)\right) + \sum_{y_i=0}\log\left(1-p(x_i)\right)\right) \\
= & \quad \{ \text{ Cálculo da derivada } \} \\
& \sum_{y_i=1}\frac{1}{p(x_i)}\frac{\partial p(x_i)}{\partial c_0} + \sum_{y_i=0}\frac{1}{1-p(x_i)}\frac{\partial(1-p(x_i))}{\partial c_0} \\
= & \quad \{ \text{ Usando que } \tfrac{\partial}{\partial c_0}p(x_i) = p(x_i)(1-p(x_i)) \ \} \\
& \sum_{y_i=1}\frac{1}{p(x_i)}p(x_i)(1-p(x_i)) + \sum_{y_i=0}\frac{1}{1-p(x_i)}(-p(x_i)(1-p(x_i))) \\
= & \quad \{ \text{ Simplificando } \} \\
& \sum_{y_i=1}(1-p(x_i)) - \sum_{y_i=0}p(x_i) \\
= & \quad \{ \text{ Reescrevendo as somas como somatórios sobre todos os dados } \} \\
& \sum_{i=0}^{n} y_i(1-p(x_i)) - \sum_{i=0}^{n}(1-y_i)p(x_i) \\
= & \quad \{ \text{ Distribuindo os termos dentro da soma } \} \\
& \sum_{i=0}^{n} y_i(1-p(x_i)) - (1-y_i)p(x_i) \\
= & \quad \{ \text{ Expansão dos produtos dentro da soma } \} \\
& \sum_{i=0}^{n}(y_i - y_i p(x_i) - p(x_i) + y_i p(x_i)) \\
= & \quad \{ \text{ Simplificando } \} \\
& \sum_{i=0}^{n}(y_i - p(x_i))
\end{aligned}
$$

Dessa forma, temos ambos os componentes do gradiente do cálculo do erro que gostaríamos de minimizar e, portanto, podemos encontrar um mínimo local para essa função utilizando o método do gradiente descendente mostrado na subseção 2.3.4.4, assim, possibilitando a obtenção de valores para c_0 e c_1 tais que $p(x_i) \approx y_i$.

A regressão logística se prova extremamente importante em cenários onde um conjunto de dados pré-definidos contém dados o suficiente de tal forma que a natureza de determinado problema possa ser observada de maneira clara ao observar esses dados em uma dimensão inferior.

Dessa forma, fazendo com que, ao encontrar uma função probabilística p que, dado um novo dado, possamos estimar suficientemente bem em qual categoria ele pertence.

2.2.4.4 Regressão de Ridge

Conforme abordado nas modelagens 1.2.1 e 1.2.2, problemas de regressão podem ser reduzidos à minimização de $\|Ax - b\|^2$, em que A é uma matriz de Vandermonde formada pelas potências dos valores da primeira coordenada dos pontos, b é o vetor com os valores da segunda coordenada, e x representa o vetor de coeficientes da função polinomial resultante do processo de regressão. Ainda que o vetor x obtido minimize de fato o erro quadrático $\|Ax - b\|^2$, a função resultante pode sofrer de *overfitting*, especialmente quando o grau do polinômio é escolhido de forma inadequada.

Nesse contexto, buscamos não apenas minimizar o erro de ajuste, mas também controlar a complexidade da função ajustada, o que pode ser alcançado restringindo a norma de x. Isso nos leva ao conceito de regressão de Ridge, também conhecida como regularização L^2, em que introduzimos um termo adicional de penalização à função de custo, visando obter coeficientes de menor magnitude e, assim, reduzir a sensibilidade do modelo a pequenas variações nos dados.

É razoável assumir que os pontos utilizados no processo de regressão originam de alguma função polinomial que representa perfeitamente os dados reais mas que nossas observações dos dados estão sujeitas a um termo de ruído gaussiano.

Dessa forma, definimos $P_c[A]$ como sendo a probabilidade de se obter a atual distribuição de um dado conjunto de pontos dado um modelo específico de coeficientes c. Por cada ponto representar um evento/dado diferente, temos que todo ponto é independente e, portanto

$$P_c[A] = \prod_{i=1}^{n} P_c[A_i],$$

sendo n a quantidade de pontos.

Dessa forma, encontrar os coeficientes c mais prováveis (no caso, coeficientes do polinômio que deu origem aos dados) para o problema de regressão é equivalente a calcular $\text{argmax}_c\ P_c[A]$.

Outra assunção razoável a ser feita é que os coeficientes c da função real que deu origem aos pontos os quais estamos aplicando o processo de regressão apresenta alguma espécie de padrão e ruído em seus coeficientes, de maneira que uma solução seja naturalmente mais provável de ocorrer do que outra.

Dessa forma, ao invés de maximizarmos $P_c[A]$, buscamos maximizar $P[A, c]$ de forma que os coeficientes c encontrados tanto expliquem bem os dados de entrada quanto façam algum sentido em relação ao padrão que acreditamos que eles seguem.

Teorema T7. *Suponha que o vetor de coeficientes x de um polinômio que deu origem a um conjunto de dados b segue uma distribuição gaussiana a priori com média zero e variância τ^2, e que os dados gerados por esse polinômio são todos independentes e apresentam ruido gaussiano com variância σ^2. Então $x \in argmax_c\ P[A, c] \implies (A^\mathsf{T}A + \lambda I)x = A^\mathsf{T}b$, sendo A a matriz de Vandermonde.*

Demonstração.

$$
\begin{aligned}
& x \in \operatorname{argmax}_c \; P[A, c] \\
\iff & \quad \{ \text{ Regra da Probabilidade Condicional } \} \\
& x \in \operatorname{argmax}_c \; P_c[A] \cdot P[c] \\
\iff & \quad \{ \; P_c[A] \text{ e } P[c] \text{ apresentam ruído em distribuição gaussiana } \} \\
& x \in \operatorname{argmax}_c \; \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - c^\mathsf{T} x_i)^2}{2\sigma^2}} \cdot \prod_{k=1}^{d} \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{c_k^2}{2\tau^2}} \\
\iff & \quad \{ \text{ Teorema T34 } \} \\
& x \in \operatorname{argmax}_c \; \ln\left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - c^\mathsf{T} x_i)^2}{2\sigma^2}} \cdot \prod_{k=1}^{d} \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{c_k^2}{2\tau^2}} \right) \\
\iff & \quad \{ \; \ln(ab) = \ln(a) + \ln(b) \; \} \\
& x \in \operatorname{argmax}_c \; \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}} \right) + \sum_{i=1}^{n} \ln\left(e^{-\frac{(y_i - c^\mathsf{T} x_i)^2}{2\sigma^2}} \right) + \sum_{k=1}^{d} \ln\left(\frac{1}{\tau\sqrt{2\pi}} \right) + \sum_{k=1}^{d} \ln\left(e^{-\frac{c_k^2}{2\tau^2}} \right) \\
\iff & \quad \{ \text{ Teorema T36 } \} \\
& x \in \operatorname{argmax}_c \; \sum_{i=1}^{n} \ln\left(e^{-\frac{(y_i - c^\mathsf{T} x_i)^2}{2\sigma^2}} \right) + \sum_{k=1}^{d} \ln\left(e^{-\frac{c_k^2}{2\tau^2}} \right) \\
\iff & \quad \{ \; \ln(e^x) = x \; \} \\
& x \in \operatorname{argmax}_c \; \sum_{i=1}^{n} -\frac{(y_i - c^\mathsf{T} x_i)^2}{2\sigma^2} + \sum_{k=1}^{d} -\frac{c_k^2}{2\tau^2} \\
\iff & \quad \{ \text{ Evidenciando termos } \} \\
& x \in \operatorname{argmax}_c \; \frac{1}{-2\sigma^2} \left(\sum_{i=1}^{n} (y_i - c^\mathsf{T} x_i)^2 + \sum_{k=1}^{d} \frac{2\sigma^2 c_k^2}{2\tau^2} \right) \\
\iff & \quad \{ \text{ Teoremas T36 e T35 } \} \\
& x \in \operatorname{argmin}_c \; \sum_{i=1}^{n} (y_i - c^\mathsf{T} x_i)^2 + \sum_{k=1}^{d} \frac{\sigma^2}{\tau^2} c_k^2 \\
\iff & \quad \{ \; \lambda = \frac{\sigma^2}{\tau^2} \; \} \\
& x \in \operatorname{argmin}_c \; \sum_{i=1}^{n} (y_i - c^\mathsf{T} x_i)^2 + \lambda \sum_{k=1}^{d} c_k^2 \\
\iff & \quad \{ \text{ Representação matricial } \} \\
& x \in \operatorname{argmin}_x \; \|b - Ax\|^2 + \lambda \|x\|^2 \\
\iff & \quad \{ \; \|b - Ax\|^2 = \|Ax - b\|^2 \; \} \\
& x \in \operatorname{argmin}_x \; \|Ax - b\|^2 + \lambda \|x\|^2 \\
\Downarrow & \quad \{ \text{ Derivando e igualando a 0 para encontrar o ponto de mínimo } \}
\end{aligned}
$$

$$
\begin{aligned}
& \nabla_x \ \|Ax-b\|^2 + \lambda \|x\|^2 = 0 \\
\iff & \quad \{ \text{ Definição } \} \\
& \nabla_x \ (Ax-b)^\mathsf{T}(Ax-b) + \lambda \|x\|^2 = 0 \\
\iff & \quad \{ \text{ Distributiva } \} \\
& \nabla_x \ (Ax)^\mathsf{T}Ax - (Ax)^\mathsf{T}b - (Ax)^\mathsf{T}b - b^\mathsf{T}b + \lambda \|x\|^2 = 0 \\
\iff & \quad \{ \text{ Somando termos iguais } \} \\
& \nabla_x \ (Ax)^\mathsf{T}Ax - (Ax)^\mathsf{T}b - 2(Ax)^\mathsf{T}b - b^\mathsf{T}b + \lambda \|x\|^2 = 0 \\
\iff & \quad \{ \text{ Teorema T20 } \} \\
& \nabla_x \ x^\mathsf{T}A^\mathsf{T}Ax - (Ax)^\mathsf{T}b - 2(Ax)^\mathsf{T}b - b^\mathsf{T}b + \lambda \|x\|^2 = 0 \\
\iff & \quad \{ \ \nabla_x \ -b^\mathsf{T}b = 0 \ \} \\
& \nabla_x \ x^\mathsf{T}A^\mathsf{T}Ax - (Ax)^\mathsf{T}b - 2(Ax)^\mathsf{T}b + \lambda \|x\|^2 = 0 \\
\iff & \quad \{ \text{ Teoremas T39, T40 e T41 } \} \\
& 2A^\mathsf{T}Ax - 2A^\mathsf{T}b + 2\lambda x = 0 \\
\iff & \quad \{ \text{ Somando } 2A^\mathsf{T}b \text{ de ambos os lados } \} \\
& 2A^\mathsf{T}Ax + \lambda 2x = 2A^\mathsf{T}b \\
\iff & \quad \{ \text{ Divindo por 2 dos dois lados } \} \\
& A^\mathsf{T}Ax + \lambda x = A^\mathsf{T}b \\
\iff & \quad \{ \ x = Ix \ \} \\
& A^\mathsf{T}Ax + \lambda Ix = A^\mathsf{T}b \\
\iff & \quad \{ \text{ Distributiva } \} \\
& (A^\mathsf{T}A + \lambda I)x = A^\mathsf{T}b
\end{aligned}
$$

□

Além disso, veja que se desconsiderarmos o termo $P[c]$ do cálculo probabilístico, ou seja, o ruído a priore dos coeficientes da função que deu origem aos dados, a solução x de maior probabilidade de ocorrência é aquela que minimiza o erro dos mínimos quadrados ($\|Ax-b\|^2$). Ao considerarmos esse termo, temos a introdução de uma nova parcela na equação $A^\mathsf{T}Ax = A^\mathsf{T}b$, que é o termo λIx, resultando em $(A^\mathsf{T}A + \lambda I)x = A^\mathsf{T}b$.

Ao olharmos cuidadosamente para o conjunto $\text{argmin}_x \ \|Ax-b\|^2 + \lambda \|x\|^2$, podemos interpretar que o x que resolve esse sistema é o vetor com os coeficientes da regressão que aproximam o sistema $Ax \approx b$ mas que também apresenta complexidade baixa (coeficientes pequenos) por conta da penalidade expressa pela parcela $\lambda \|x\|^2$, que é a soma dos quadrados dos pesos, penalizando soluções que apresentem o vetor x muito grande em comparação com soluções mais simples. Dessa maneira, podemos entender que λ é o parâmetro que controla a intensidade da penalidade que queremos atribuir para essa parcela.

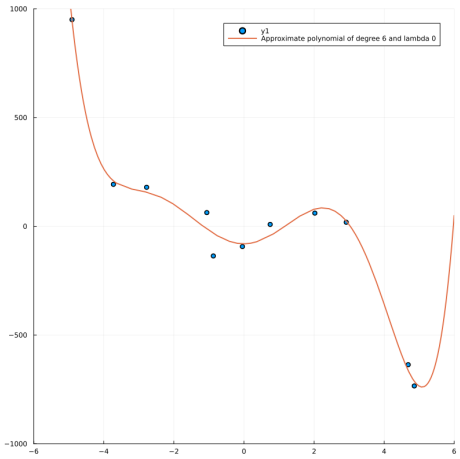


Figura F34: Exemplo de regressão usando $\lambda = 0$.

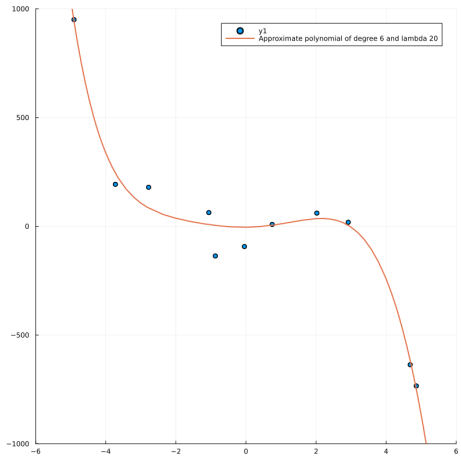


Figura F35: Exemplo de regressão usando $\lambda = 20$.

Pela definição de λ como sendo a razão entre σ^2 e τ^2, que são a variância do ruído dos dados e a variância a priori dos coeficientes. Dessa forma, temos que o valor de λ varia a respeito do quão confiável nossos dados são. Se os dados forem confiáveis e acreditarmos que os coeficientes apresentam bastante variação, então o valor de λ será pequeno. Nesse caso, preferimos que o foco da regressão seja no ajuste dos dados ao invés de forçar os coeficientes a serem pequenos. Em contrapartida, se os dados apresentarem um alto ruído ou acreditarmos que os coeficientes deveriam apresentar pequena variação, λ será grande. Nesse caso, preferimos um modelo mais simples com coeficientes pequenos e pouco *overfitting*.

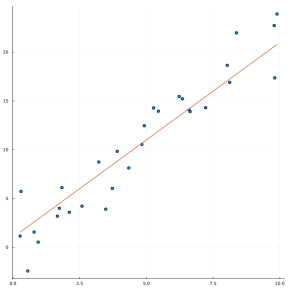


Figura F36: Exemplo de dados com ruído com distribuição gaussiana e a função original que deu origem a eles.

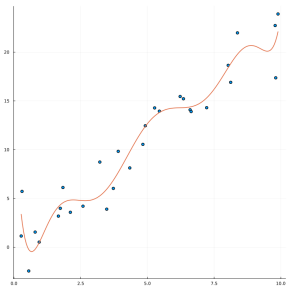


Figura F37: Exemplo de regressão dos dados da figura F36 usando $\lambda = 0$.

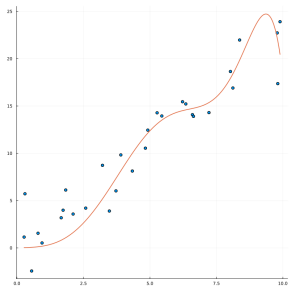


Figura F38: Exemplo de regressão dos dados da figura F36 usando $\lambda = 100$.

2.2.5 Exercícios de Modelos Aproximados

Exercício E13:

Sabemos que $Q_{(3\times3)}$ é uma matriz ortogonal e que $A = Q \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$

e que $b = Q \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$.

1. Prove algebricamente que $\|Qv\| = \|v\|$, para qualquer $v \in \mathbb{R}^3$.
2. Determine o x que minimiza $\|Ax - b\|$.
3. Determine o cosseno do ângulo entre o b e o plano gerado pelas colunas de A, se possível.

Exercício E14:

Queremos resolver o sistema $Ax = b$ que tem 3 equações e 2 variáveis. O vetor b e a matriz A foram concatenadas em uma matriz só $\begin{bmatrix} b|A \end{bmatrix}$ (b está na primeira coluna e A no restante). Sabemos também que

$$\begin{bmatrix} b|A \end{bmatrix} = Q \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & 2 \\ 3 & 0 & 0 \end{bmatrix},$$

sendo Q uma matriz ortogonal. Somente usando essas informações, determine se o sistema $Ax = b$ tem solução e, se sim, determine a solução do sistema. Caso contrário, aproxime o sistema por mínimos quadrados e determine o erro (resposta numérica).

Exercício E15:

A tabela abaixo foi obtida como resultado de um experimento relativo ao valor da temperatura T (em graus Celsius) com a posição x (em centímetros):

x	1	2	4
T(x)	8	6	12

Determine a curva da forma $T(x) = 2^{xc} + 4$ que melhor se ajusta aos dados da tabela com o método de mínimos quadrados com coeficientes não-lineares e use o modelo para calcular $T(0)$.

Exercício E16:

Seja $(c - 300)^2 + (c + m - 400)^2 + (m - 700)^2$ o custo mensal em reais de uma fábrica dado que ela produz c cadeiras e m mesas. Determine quantas cadeiras e mesas a fábrica deve produzir para minimizar seu custo usando mínimos quadrados (resposta numérica).

Exercício E17:

Dados os pontos $(2, 4)$, $(4, 7)$ e $(6, 14)$, escreva um pseudo-código para determinar uma aproximação para o valor da função em $x = 5$, sabendo que o modelo é $y = c_1 x^{c_2} + 3$.

Exercício E18:

As tabelas abaixo foram obtidas como resultado de um experimento relativo ao valor da temperatura T (em graus Celsius) com a posição x (em centímetros). Determine a curva da forma $T(x)$ que melhor se ajusta aos dados da tabela com o método de mínimos quadrados com coeficientes não-lineares e use o modelo para calcular $T(K)$.

1. Usando o modelo $T(x) = c_0 x^{c_1}$, aproxime $T(0.3)$.

x	0.1	0.2	0.4	0.8	0.9
T(x)	22	43	84	210	320

2. Usando o modelo $T(x) = 2^{xc} + 4$, aproxime $T(8)$.

x	1	2	4
T(x)	8	6	12

Exercício E19:

(Escolhendo o polinômio correto):

1. Gere aleatoriamente 30 pontos de um polinômio de grau 5.
2. Faça regressão polinomial com polinômios de grau 0 até 29.
3. É possível fazer a regressão com um polinômio de grau maior que 29? Justifique.
4. Faça o plot do Erro total (eixo y) por grau (eixo x). O que se pode dizer desse gráfico conforme o grau aumenta? Era o que você esperava? Por quê?

Exercício E20:

1. Determine o vetor na reta gerada por $\begin{bmatrix}2\\3\end{bmatrix}$ mais próximo do vetor $\begin{bmatrix}7\\5\end{bmatrix}$ usando cálculo.
2. Dado dois vetores a e $b \in \mathbb{R}^n$, determine o algoritmo geral para achar o vetor na reta gerada por a, mais próximo do vetor b usando cálculo.

Exercício E21:

Suponha que fizemos medições de uma dada quantidade em 21 momentos, igualmente espaçados, entre $x = -10$ e $x = 10$. Todas estas medições têm valor nulo, exceto a que foi feita em $x = 0$, que vale 1. Use cálculo para descobrir a melhor reta que se adapta a esses pontos.

Exercício E22:

Determine o melhor ponto que aproxima os pontos $(3,4)$, $(5,7)$, e $(22,10)$ por cálculo.

Exercício E23:

Entre todos os vetores que são combinações lineares de $\begin{bmatrix}1\\2\\3\end{bmatrix}$ e $\begin{bmatrix}2\\2\\2\end{bmatrix}$, determine qual é o mais próximo de $\begin{bmatrix}3\\4\\6\end{bmatrix}$. Use cálculo 2.

Exercício E24:

Seja

$$\begin{cases} x_1 + 3x_2 = 7 \\ 3x_2 = 10 \\ 4x_2 = 5 \end{cases}$$

um sistema linear.

Seja $a_1 = \begin{bmatrix}1 & 0 & 0\end{bmatrix}^\mathsf{T}$, $a_2 = \begin{bmatrix}3 & 3 & 4\end{bmatrix}^\mathsf{T}$, $a_3 = \begin{bmatrix}7 & 10 & 5\end{bmatrix}^\mathsf{T}$ e $A = \begin{bmatrix} | & | & | \\ a_1 & a_2 & a_3 \\ | & | & | \end{bmatrix}$. Repare a ligação **forte** do sistema linear com os vetores a_1, a_2 e a_3.

1. Determine a solução aproximada por mínimos quadrados do sistema por Householder.
2. Entre todos os vetores que são combinações lineares de a_1 e a_2, determine o mais próximo de a_3.
3. Determine o R do QR de A.
4. Seja θ o ângulo entre o vetor a_3 e o plano gerado por a_1 e a_2. Determine $\text{sen}(\theta)$.
5. Determine o Q do QR de A por Householder.
6. Seja v um vetor em $\mathbb{R}^n$ com norma igual à 1. Prove que a transformação $I - 2vv^\mathsf{T}$ é ortogonal.

Exercício E25:

(Lei de Moore) Em 1982 os computadores já tinham 1 milhão de transistores, em 1988 os computadores já tinham 4 milhões de transistores, e em 2000 já tinham 8 milhões de transistores. Usando o modelo exponencial com a base que você quiser ($y = 2^x$, $y = 3^x$, $y = 4^x$, ...),

1. Determine, usando mínimos quadrados, quantos anos demora para dobrar o número de transistores. Faça um esboço do modelo exponencial que melhor aproxima os dados. Nesse problema use derivadas de cálculo para resolver o problema de mínimos quadrados.

2. Qual é o erro absoluto do problema? Mostre o erro no esboço do item anterior.

Exercício E26:

Filmes Seja U a matriz com a preferência de 4 usuários por 5 filmes levando em consideração somente o nível de comédia:

$$U = \begin{bmatrix} 2 & 10 & 20 & 200 & -10 \\ -3 & -15 & -30 & -300 & 15 \\ 5 & 25 & 50 & 500 & -25 \\ 7 & 35 & 70 & 700 & -35 \end{bmatrix}$$

Determine uma possível solução para quando cada usuário gosta (ou não gosta) de comédia e quanto cada filme é (ou não é) de comédia.

Exercício E27:

Um estudante fez uma pesquisa com 13 alunos de uma turma de Computação Científica e Análise de Dados e descobriu certas preferências quando perguntou para eles escolherem entre dois filmes:

1. Toy story 12 x 1 Rocky
2. De volta pro futuro 8 x 5 Curtindo a vida adoidado
3. Os incríveis 10 x 3 Duna
4. Batman begins 7 x 5 Harry Potter 1
5. Shrek 11 x 2 Duna
6. Harry Potter 10 x 3 Rocky
7. Toy story 9 x 4 De volta para o futuro
8. Os incríveis 9 x 4 Harry potter 1
9. Curtindo a vida adoidado 7 x 5 Duna
10. De volta para o futuro 7 x 5 Duna
11. Shrek 12 x 1 Rocky
12. Os incríveis 9 x 4 Batman Begins
13. Toy story 8 x 5 Batman Begins
14. Os incríveis 10 x 3 Curtindo a vida adoidado

Determine um ranking dos filmes preferidos dos 13 alunos usando mínimos quadrados.

Exercício E28:

Um entusiasta da área de Computação Científica e Análise de Dados decidiu se pesar durante 1 ano no primeiro dia de cada mês, obtendo os seguintes dados a cada pesagem:

Mês	Jan	Fev	Mar	Abr	Mai	Jun
Peso (kg)	121.2	120.5	119.8	120.0	117.8	118.2

Mês	Jul	Ago	Set	Out	Nov	Dez
Peso (kg)	117.3	115.2	116.0	114.5	115.3	113.7

Utilize regressão para estimar em quantos meses o entusiasta terá 110.0 kg. Argumente o porquê da sua escolha de função para a regressão ser uma boa escolha.

Exercício E29:

Uma notícia falsa está se espalhando rapidamente por uma rede social. Às 15:35 da tarde já havia 600 publicações com a notícia, às 15:36 já havia 1400 publicações com a notícia, e às 15:38 já havia 3200 publicações com a notícia.

Uma pesquisadora acredita que o modelo que relaciona o tempo e o número de publicações é dado por:

$$\text{número de publicações} = 3^{ct} + 500$$

para alguma constante c. Determine a que horas o número de publicações chegará a 24700.

Exercício E30:

A polícia chega ao local de um assassinato às 15h. Eles imediatamente medem e registram a temperatura do corpo, que está em 34°C. e inspecionam minuciosamente a área. Quando terminam a inspeção, às 16h30, eles medem novamente a temperatura do corpo, que caiu para 30°C. Eles esperam mais 1 hora e medem a temperatura novamente, que caiu para 25°C.

A temperatura no local do crime permaneceu estável em 20°C, e a temperatura normal do corpo é de 37°C.

Sabendo que a temperatura do corpo obedece à Lei de Resfriamento de Newton, use regressão com coeficientes não lineares para estimar o horário em que a pessoa foi assassinada.

A Lei de Resfriamento de Newton tem a forma:

$$T(t) = T_{\text{final}} + (T_{\text{inicial}} - T_{\text{final}})e^{-kt},$$

onde $T(t)$ é a temperatura do corpo em função do tempo, T_{inicial} é a temperatura inicial (no

momento da morte), T_{final} é a temperatura final que o corpo atingiu, e k é a constante de resfriamento.

Exercício E31:

Seja P o plano gerado por $\begin{bmatrix}2 & 0 & 1\end{bmatrix}^\mathsf{T}$ e $\begin{bmatrix}1 & 2 & 2\end{bmatrix}^\mathsf{T}$

1. Determine um vetor perpendicular ao plano P.
2. Determine o vetor que é a projeção ortogonal do vetor $\begin{bmatrix}1 & 1 & 1\end{bmatrix}^\mathsf{T}$ no plano.
3. Determine a distância do ponto $\begin{bmatrix}1 & 1 & 1\end{bmatrix}^\mathsf{T}$ ao plano.
4. Determine o vetor que é a projeção ortogonal do vetor $\begin{bmatrix}3 & 2 & 3\end{bmatrix}^\mathsf{T}$ no plano.
5. Determine a distância do $\begin{bmatrix}3 & 2 & 3\end{bmatrix}^\mathsf{T}$ ao plano.

Exercício E32:

Determine se as afirmações abaixo são verdadeiras ou falsas. Prove se verdadeira, ou exiba um contra-exemplo caso falsa.

1. Se A e B são matrizes ortogonais, então AB é uma matriz ortogonal.
2. Se $\det(A) = 1$ ou -1, então A é uma matriz ortogonal.
3. Se A é uma matriz ortogonal, então $\det(A) = 1$ ou -1.

Exercício E33:

Escreva as duas definições diferentes de uma matriz ortogonal e explique por que elas são equivalentes.

Exercício E34:

Sejam A e B matrizes. Prove *algebricamente* que se Q é uma matriz ortogonal ($Q^\mathsf{T}Q = I$), então $\text{dist}(QA, QB) = \text{dist}(A, B)$.

Exercício E35:

Seja $AZ = LLQU$ aonde Q e Z são matrizes ortogonais, L é uma matriz triangular inferior com diagonal não-nula e U é uma matriz triangular superior com diagonal não-nula. Todas as matrizes tem dimensão $(n \times n)$. Dado U, L, Z, Q, e b, escreva um pseudo-código eficiente (dica: o seu algoritmo deve ser $O(n^2)$) que determine x tal que $Ax = Lb$. Você pode usar as funções de substituição reversa e direta no seu código (não precisa escrever-las).

Exercício E36:

Determine a solução para o sistema

$$\begin{cases} x_1 + 2x_2 &= 5 \\ 3x_2 &= 1 \\ 4x_2 &= 3 \end{cases}$$

que minimize o erro $\|Ax - b\|$.

Exercício E37:

Usando a informação que

$$\begin{bmatrix} | & | \\ a_1 & a_2 \\ | & | \end{bmatrix} = \begin{bmatrix} | & | \\ q_1 & q_2 \\ | & | \end{bmatrix} \begin{bmatrix} 5 & 9 \\ 0 & 4 \end{bmatrix}$$

e que q_1 e q_2 tem norma igual à 1 e são perpendiculares entre si. (Dica: primeiro faça um desenho.)

1. Determine a distância do vetor a_2 para o reta gerada por a_1.
2. Determine o tamanho de a_1.
3. Determine o tamanho da projeção de a_1 na reta gerada por q_1.
4. Determine o tamanho da projeção de a_1 na reta gerada por q_2.
5. Determine o tamanho da projeção de a_2 na reta gerada por q_1.
6. Determine o tamanho da projeção de a_2 na reta gerada por q_2.
7. Determine o tamanho de a_2.

Exercício E38:

Seja

$$\begin{bmatrix} | & | & | & | \\ a_1 & a_2 & a_3 & a_4 \\ | & | & | & | \end{bmatrix} = \begin{bmatrix} | & | & | & | \\ v_1 & v_2 & v_3 & v_4 \\ | & | & | & | \end{bmatrix} \begin{bmatrix} 3 & 4 & 2 & 4 \\ 0 & 7 & 0 & 5 \\ 0 & 0 & 3 & 6 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

e v_1, v_2, v_3 e $v_4 \in \mathbb{R}^4$ tem norma igual à 1 e são perpendiculares entre si.

1. Usando as propriedades algébricas do produto interno, determine o cosseno entre a_3 e a_1.
2. Determine c_1 e $c_2 \in \mathbb{R}$ tal que para todo d_1 e $d_2 \in \mathbb{R}$, $\text{dist}(c_1v_1 + c_2v_2, a_3) \leq \text{dist}(d_1v_1 + d_2v_2, a_3)$.
3. Usando as propriedades algébricas do produto interno, determine $z \in \mathbb{R}$ não-nulo tal que $zv_1 = v_2$, se possível. Justifique algebricamente se não for possível.

Exercício E39:

Seja n um vetor unitário em $\mathbb{R}^2$ e A uma matriz tal que $A = I - 2nn^\mathsf{T}$.

1. Verifique que A é uma matriz 2×2.
2. O vetor n é um autovetor de A? Qual é o autovalor associado?
3. Um vetor perpendicular a n é um autovetor de A? Qual é o autovalor associado?
4. Explique em palavras e com um desenho o que essa transformação faz.

Exercício E40:

Um usuário quer resolver o sistema $Ax = b$ que tem 3 equações e 2 variáveis. Ele juntou a matriz A com o vetor b em uma matriz só, $\big[A|b\big]$ (anexou b na última coluna), e rodou o QR nessa matriz tal que

$$\big[A|b\big] = Q \begin{bmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 3 \end{bmatrix}$$

Somente usando as informações da matriz R, determine a solução do sistema linear $Ax = b$, se o sistema tem solução. Caso contrário aproxime o sistema por mínimos quadrados e determine o erro, somente com as informações da matriz R.

Exercício E41:

Sabemos que $Q_{3\times3}$ é uma matriz ortogonal e que

$$A = Q \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$$

e que

$$b = Q \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

1. Determine o x que minimiza $\|Ax - b\|$.
2. É possível determinar a distância do b para o plano gerado pelas colunas de A?

Exercício E42:

Calcule a fatoração QR da matriz

$$A = \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}.$$

Exercício E43:

Usando a informação que a decomposição QR de A é

$$\begin{bmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{bmatrix} = Q \begin{bmatrix} 3 & 4 & 5 \\ 0 & 7 & 6 \\ 0 & 0 & 9 \end{bmatrix}.$$

Determine:

1. a solução de mínimos quadrados de $Ax \approx b$
2. o erro absoluto de mínimo quadrados.
3. o erro relativo ($cos(\theta)$) de mínimo quadrados.

Exercício E44:

Determine a melhor aproximação por mínimos quadrados do sistema

$$\begin{cases} x_1 & = 1 \\ 2x_2 & = 1 \\ 3x_2 & = 1 \end{cases}$$

utilizando a decomposição QR. Qual é o erro do sistema?

Exercício E45:

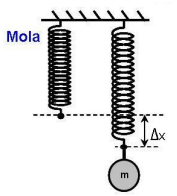
Problemas de interpolação e regressão

Quais desses problemas podem ser resolvidos exatamente ou precisa ser resolvido aproximadamente por mínimos quadrados? Para cada item, escreva um pseudo-código que resolva o problema.

1. Existe uma reta $y = ax + b$ que passa pelos pontos $(1, 3)$ e $(4, 5)$?
2. Existe uma reta $y = ax + b$ que passa pelos pontos $(1, 3)$, $(4, 5)$ e $(2, 4)$?
3. Existe uma parábola $y = ax^2 + bx + c$ que passa pelos pontos $(1, 3)$ e $(4, 5)$?
4. Existe uma parábola $y = ax^2 + bx + c$ que passa pelos pontos $(1, 3)$, $(4, 5)$, $(2, 4)$ e $(3, 5)$?

Exercício E46:

A Lei de Hooke é dada pela equação $mg = kx$.



$m\ (kg)$	5	6	7	8
$x\ (m)$	1	2	3	4

Neste caso, tem-se uma massa m, a gravidade g $(10m/s^2)$, deslocamento Δx da posição original, e a constante da mola k, na qual a constante k tem um valor diferente para cada mola. A lei de Hooke diz que a força exercida por uma mola é diretamente proporcional à sua deformação. No laboratório, prendemos diferentes massas na mesma mola e medimos o deslocamento para cada massa. Use a tabela de medições acima para determinar uma aproximação para a constante da mola utilizada por mínimos quadrados.

Exercício E47:

A tabela abaixo foi obtida como resultado de um experimento relativo ao valor da temperatura T (em graus Celsius) com a posição x (em centímetros):

x	0.1	0.2	0.4	0.8	0.9
$T(x)$	22	43	84	210	320

Qual é uma boa aproximação para $x = 1.6$?

Exercício E48:

Use o método dos mínimos quadrado para encontrar a melhor *função constante* $(y = c_0)$ que se adapta aos pontos:

1. $(1000, 4)$, $(300, 6)$ e $(6000, 11)$.

2. (x_0, y_0), ..., (x_n, y_n) (Faça algebricamente).

(Dica para 2b) Resolva esse problema de cálculo 1: Determine o mínimo global de $h(c) = (y_0 - c)^2 + \ldots + (y_n - c)^2$ aonde y_0, ..., y_n são valores dados.

Exercício E49:

Dado pontos $(2, 4)$, $(4, 7)$ e $(6, 14)$ escreva um pseudo-código para determinar uma aproximação para o valor da função em $x = 5$, sabendo que o modelo é $y = c_1 x^{c_2} + 3$.

Exercício E50:

A tabela abaixo foi obtida como resultado de um experimento relativo ao valor da temperatura T (em graus Celsius) com a posição x (em centímetros):

x	-1	0	1
T(x)	0	1	0

1. Determine a curva da forma $T(x) = c_1 x^3 + c_2$ que melhor se ajusta aos dados da tabela por regressão.
2. Determine o erro absoluto.
3. Determine o cosseno do erro relativo.

Exercício E51:

Para qualquer base a, o seu computador não sabe calcular $\log_a$. Determine um *pseudo-código* para rodar no seu computador para achar uma boa aproximação para $\log(200)_3$ usando interpolação por uma cúbica.

Exercício E52:

A tabela abaixo foi obtida como resultado de um experimento relativo ao valor da temperatura T (em graus Celsius) com a posição x (em centímetros):

x	1	10	1000
T(x)	0.1	1	100

Determine a curva da forma $T(x) = x^c$ que melhor se ajusta aos dados da tabela com o método de mínimos quadrados com coeficientes não-lineares e use o modelo para calcular $T(2)$.

Exercício E53:

(Não unicidade de mínimos quadrados) Um aluno fez uma pesquisa com 3 alunos de uma turma de Computação Científica e Análise de Dados e descobriu certas preferências quando perguntou para eles escolherem entre dois filmes:

1. Batman begins 18 x 5 Os incríveis
2. Os incríveis 5 x 1 Harry Potter 1
3. Batman begins 2 x 5 Harry Potter 1

Qual é o filme preferido dos 3 alunos? A solução de mínimos quadrados é única?

Exercício E54:

A tabela abaixo foi obtida como resultado de um experimento relativo ao valor da temperatura T (em graus Celsius) com a posição x (em centímetros):

x	1	2	4	5	6
T(x)	8	6	12	15	20

Determine a curva da forma $h(x) = c_0x^2 + c_1 + c_2cos(x) + c_3x$ (modelo) que melhor se ajusta a todos os dados da tabela com o método de mínimos quadrados e use para calcular o valor da temperatura na posição $3cm$.

Exercício E55:

Quando rodamos o algoritmo QR nos vetores a_1, a_2, $a_3 \in \mathbb{R}^3$, na terceira iteração, houve uma divisão por zero. Determine quais afirmações são verdadeiras e justifique.

1. a_2 e a_3 são sempre colineares
2. a_1, a_2 e a_3 são sempre coplanares.
3. a_1 e a_3 são sempre colineares.
4. a_2 e a_1 são sempre colineares.

Exercício E56:

Seja

$$\begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \approx \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} x \end{bmatrix}.$$

Determine o x que melhor aproxima por mínimos quadrados.

Exercício E57:

Seja

$$\begin{bmatrix} 2 & 2 \\ 5 & 7 \end{bmatrix} \approx \begin{bmatrix} 2 \\ 6 \end{bmatrix} \begin{bmatrix} x & y \end{bmatrix}.$$

Determine o x e y que melhor aproximam por mínimos quadrados.

Exercício E58:

(Filmes) Eu trabalho na Netflix e gostaria de saber um pouco sobre os gostos dos meus usuários João e Maria. Imagina que temos duas pessoas, João e Maria, e 3 filmes, X, Y e Z. Imagina que sabemos uma nota numérica (um float) para cada pessoa que diz quanto ela gosta ou não de algum gênero de filme, por exemplo comédia. J_C é quanto o João gosta de comédia e M_C é quanto a Maria gosta de média. Notas altas, dirão que a pessoa gosta muito de comédia, nota 0 diz que a pessoa é indiferente à comédia, e notas negativas dirão que a pessoa não gosta de comédia. Exemplo: João adora comédia e Maria gosta só um pouco de comédia $J_C = 50$ e $M_C = 2$. Queremos agora descobrir uma nota numérica que diz quanto de comédia tem em cada filme. C_X é quanto de comédia tem no filme X.

Um modelo linear bem simples quanto João e Maria vão gostar do filmes (levando em consideração só comédia) é:

$$\begin{array}{ll} J_X = J_C \cdot C_X & M_X = M_C \cdot C_X \\ J_Y = J_C \cdot C_Y & M_Y = M_C \cdot C_Y \\ J_Z = J_C \cdot C_Z & M_Z = M_C \cdot C_Z \end{array}$$

1. Escreva o modelo linear de maneira matricial.
2. Determine quanto cada filme é de comédia dado que sabemos que $J_C = 50$, $M_C = 50$, $J_X = 50$, $J_Y = 50$, $J_Z = 50$, $M_X = 50$,$M_Y = 50$,e $M_Z = 3$ por mínimos quadrados.

Capítulo 2.3

Modelos Dinâmicos

2.3.1 Truque

Queremos encontrar uma fórmula fechada para x_k tal que $x_k = A^k x_0$. No entanto, quando A é uma matriz densa, essa tarefa pode ser complexa por estarmos tratando de recorrências lineares de ordem superior a 1. Para simplificar esse problema, buscamos modificar a equação $x_k = A^k x_0$ de forma que as recorrências provenientes do formato de A apresentem ordem 1.

Veja que, se A for uma matriz diagonal, o sistema pode ser resolvido de forma simples e eficiente. Por exemplo, no caso (3×3):

$$\begin{bmatrix} x_k \\ y_k \\ z_k \end{bmatrix} = \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}^k \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix}.$$

Pelo teorema T22, temos que esse sistema é o mesmo que:

$$\begin{bmatrix} x_k \\ y_k \\ z_k \end{bmatrix} = \begin{bmatrix} a_{11}^k & 0 & 0 \\ 0 & a_{22}^k & 0 \\ 0 & 0 & a_{33}^k \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix},$$

que, por sua vez, é simples de se resolver visto que aplicar a matriz no vetor é o mesmo que multiplicar cada posição i do vetor pelo elemento de índice ii da matriz.

Por exemplo, em um caso onde:

$$\begin{bmatrix} x_k \\ y_k \\ z_k \end{bmatrix} = \begin{bmatrix} 2^k & 0 & 0 \\ 0 & 5^k & 0 \\ 0 & 0 & 8^k \end{bmatrix} \begin{bmatrix} 3 \\ 5 \\ 1 \end{bmatrix},$$

temos que:

$$\begin{cases} x_k = 3 \cdot 2^k, \\ y_k = 5 \cdot 5^k, \\ z_k = 1 \cdot 8^k. \end{cases}$$

Assim, é visível que, quando A é uma matriz diagonal, a recorrência se torna muito mais simples de ser resolvida.

Quando um vetor é multiplicado repetidamente por uma matriz diagonal, cada uma de suas coordenadas é escalada de acordo com os elementos da diagonal da matriz. No caso da matriz A definida acima, essa transformação afeta cada componente do vetor de maneira independente.

Se uma das entradas diagonais da matriz for maior em magnitude do que as outras, a coordenada correspondente do vetor crescerá mais rapidamente ao longo das sucessivas multiplicações, dominando a direção do vetor resultante. No caso da matriz A diagonal definida acima, temos que um vetor aplicado muitas vezes nessa matriz apresentará sua terceira coordenada muito mais do que as outras, visto que aplicar A nesse vetor faz com que a terceira coordenada desse vetor estique em 8 vezes. Nesse caso em específico, podemos dizer que o vetor resultant tenderá a um vetor de direção $\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^\mathsf{T}$ pois a terceira coordenada cresce muito mais rápido do que as demais.

2.3.2 Troca de Variáveis

Por meio da diagonalização, temos a garantia de que, se a base de autovetores de uma matriz A for completa, então A pode ser decomposta no formato $A = VDV^{-1}$, onde V é uma matriz inversível e D é uma matriz diagonal.

Essa decomposição simplifica significativamente o problema: enquanto o cálculo de x_k tal que $x_k = A^k x_0$ é trabalhoso e ineficiente, o cálculo de x_k tal que $V^{-1}x_k = D^k V^{-1}x_0$ é muito mais simples e eficiente de se resolver visto que D é uma matriz diagonal. Além da própria computação de x_k por meio da multiplicação da matriz D por ela mesma k vezes ser mais barata, esse sistema também evidencia a facilidade de encontrar uma fórmula fechada para x_k.

A vantagem de utilizarmos a diagonalização da matriz A para resolver sistemas na forma $x_k = A^k x_0$ se dá pelo fato de que elevar uma matriz A, tal que $A = VDV^{-1}$, por uma potência k, faz com que o cálculo de A^k seja muito mais simples pelo fato de $A^k = VD^kV^{-1}$, como mostrado pelo teorema T8.

Teorema T8 (Matriz Diagonalizável Elevada a um Natural)**.** *Seja A uma matriz quadrada. Se $A = VDV^{-1}$, onde V é uma matriz invertível e D é uma matriz diagonal, então, $\forall k \in \mathbb{N}$, $A^k = VD^kV^{-1}$.*

Demonstração.

Caso Base: $k = 1$: $A^1 = VD^1V^{-1}$. Logo, o caso base é válido.

Hipótese de Indução: $\forall k \in \mathbb{N}$, $A^k = VD^kV^{-1}$.

Passo Indutivo: $k+1$;

$$
\begin{aligned}
& A^{k+1} \\
= & \quad \{\ A = VDV^{-1}\ \} \\
& (VDV^{-1})^{k+1} \\
= & \quad \{\ \text{Expandindo a exponenciação}\ \} \\
& VDV^{-1}(VDV^{-1})^{k} \\
= & \quad \{\ \text{Hipótese de Indução}\ \} \\
& VDV^{-1}VD^{k}V^{-1} \\
= & \quad \{\ V^{-1}V = I\ \} \\
& VDD^{k}V^{-1} \\
= & \quad \{\ DD^{k} = D^{k+1}\ \} \\
& VD^{k+1}V^{-1}
\end{aligned}
$$

Portanto, provamos que se A é diagonalizável, então, $\forall k \in \mathbb{N}$, $A^k = VD^kV^{-1}$. □

Teorema T9 (Troca de Variáveis de Modelos Dinâmicos)**.** *Seja A uma matriz diagonalizável tal que $A = VDV^{-1}$, sendo V uma matriz inversível e D uma matriz diagonal. Sejam x_k e x_0 vetores de modo que $\bar{x}_k = V^{-1}x_k$ e $\bar{x}_0 = V^{-1}x_0$. Se $x_k = A^kx_0$, então $\bar{x}_k = D^k\bar{x}_0$.*

Demonstração.

$$
\begin{aligned}
& x_k = A^kx_0 \\
\iff & \quad \{\ A = VDV^{-1}\ \} \\
& x_k = (VDV^{-1})^kx_0 \\
\iff & \quad \{\ \text{Teorema T8}\ \} \\
& x_k = VD^kV^{-1}x_0 \\
\iff & \quad \{\ \text{Multiplicando por } V^{-1} \text{ de ambos os lados}\ \} \\
& V^{-1}x_k = V^{-1}VD^kV^{-1}x_0 \\
\iff & \quad \{\ V^{-1}V = I\ \} \\
& V^{-1}x_k = D^kV^{-1}x_0 \\
\iff & \quad \{\ V^{-1}x_i = \bar{x}_i\ \} \\
& \bar{x}_k = D^k\bar{x}_0
\end{aligned}
$$

□

Essa equação é muito mais simples de se resolver porque $V^{-1}x_k$ e $V^{-1}x_0$ envolvem apenas transformações lineares (multiplicações por uma matriz fixa V^{-1}), enquanto D^k é uma matriz diagonal, cuja potência é computada, segundo o teorema T22, simplesmente elevando os elementos da diagonal à potência k. Essa simplificação elimina a necessidade de cálculos iterativos para encontrar A^k.

Além disso, é interessante notar que fatorar a matriz A em VDV^{-1} e, finalmente, multiplicar V^{-1} por x, equivale a tomar x na base dos autovetores de A. Nesse sentido, para resolver o sistema dinâmico, estamos inicialmente expressando o vetor x_0 na base dos autovetores da matriz A, resolvendo a recorrência com a matriz D — em um sistema desacoplado, pois D é diagonal — e, por fim, retornando à base original ao multiplicar o resultado por V.

Por fim, descobrimos uma forma eficiente de resolver o sistema de recorrências ao transformá-lo em um problema de recorrências de ordem 1 e, em seguida, voltar à base original e acoplar os resultados das recorrências já resolvidas.

2.3.3 Fatoração (Diagonalização)

2.3.3.1 Objetivo e Ideia da Diagonalização

Os autovetores v_i de uma matriz $A_{n\times n}$, $\forall i \in \{1, \ldots, n\}$, são os vetores que satisfazem a equação

$$Av_i = \lambda_i v_i.$$

Sendo λ_i autovalores. Estamos ordenando os autovetores e autovalores de 1 a n de modo que $\lambda_1 > \lambda_2 > \cdots > \lambda_n$ e que v_i seja o autovetor relativo ao autovalor λ_i. Pela definição de autovetor, temos o seguinte sistema de equações:

$$\begin{cases} Av_1 &= \lambda_1 v_1, \\ &\vdots \\ Av_i &= \lambda_i v_i \\ &\vdots \\ Av_n &= \lambda_n v_n. \end{cases}$$

Esse sistema pode ser reescrito como

$$A \underbrace{\begin{bmatrix} | & & | & & | \\ v_1 & \ldots & v_i & \ldots & v_n \\ | & & | & & | \end{bmatrix}}_{V} = \underbrace{\begin{bmatrix} | & & | & & | \\ v_1 & \ldots & v_i & \ldots & v_n \\ | & & | & & | \end{bmatrix}}_{V} \underbrace{\begin{bmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \lambda_i & & \\ & & & \ddots & \\ & & & & \lambda_n \end{bmatrix}}_{D}.$$

Dessa forma, temos então que $AV = VD$. Veja que, se V for uma matriz inversível, então

$$
\begin{aligned}
& AV = VD \\
\iff \quad & \{ \text{ Multiplicando por } V^{-1} \text{ a direita } \} \\
& AVV^{-1} = VDV^{-1} \\
\iff \quad & \{ \ VV^{-1} = I \ \} \\
& A = VDV^{-1}.
\end{aligned}
$$

Ou seja,

$$
A = \underbrace{\begin{bmatrix} | & & | & & | \\ v_1 & \dots & v_i & \dots & v_n \\ | & & | & & | \end{bmatrix}}_{V} \underbrace{\begin{bmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \lambda_i & & \\ & & & \ddots & \\ & & & & \lambda_n \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} | & & | & & | \\ v_1 & \dots & v_i & \dots & v_n \\ | & & | & & | \end{bmatrix}^{-1}}_{V^{-1}}.
$$

De maneira mais formal, dizemos que A é uma matriz diagonalizável se existem matrizes V inversível e D diagonal tal que $A = VDV^{-1}$.

2.3.3.2 Algoritmo para Diagonalização

Vimos na seção 2.3.3.1 que se uma matriz A de dimensão $(n \times n)$ apresentar uma base de n autovetores linearmente independentes $v_1, \dots, v_i, \dots, v_n$, então A pode ser fatorado como

$$
A = \underbrace{\begin{bmatrix} | & & | & & | \\ v_1 & \dots & v_i & \dots & v_n \\ | & & | & & | \end{bmatrix}}_{V} \underbrace{\begin{bmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \lambda_i & & \\ & & & \ddots & \\ & & & & \lambda_n \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} | & & | & & | \\ v_1 & \dots & v_i & \dots & v_n \\ | & & | & & | \end{bmatrix}^{-1}}_{V^{-1}}.
$$

Dessa maneira, temos então que uma forma de decompor uma matriz A em VDV^{-1} é a partir do cálculo de seus autovetores e autovalores.

2.3.4 Extensões

2.3.4.1 Método do Ponto Fixo

Mostramos em 2.1.3 que resolver o sistema $Ax = b$ requer o uso de uma fatoração matricial provinda da Eliminação Gaussiana, que possui alta complexidade computacional ($O(n^3)$). É

evidente que, no caso de A ser uma matriz esparsa, diversas computações feitas pela Eliminação Gaussiana se tornam inúteis por conta das multiplicações por 0. Portanto, torna-se essencial buscar métodos mais eficientes para resolver tais sistemas.

Um conceito fundamental em métodos numéricos é o de ponto fixo. Dada uma função real f, dizemos que x é ponto fixo de f se

$$f(x) = x.$$

O método de ponto fixo estabelece que, se f satisfizer um critério de convergência adequado, então a sequência definida por

$$x_{k+1} = f(x_k),$$

com x_k representando a aproximação para x na iteração k, converge para um ponto fixo de f. No contexto de sistemas lineares, podemos reescrever o sistema $Ax = b$ na forma $x = Mx + c$ de tal modo que a iteração do método do ponto fixo assume a forma

$$x_{k+1} = Mx_k + c$$

com x_0 sendo um chute inicial.[1]

Teorema T10 (Critério de Convergência do Método do Ponto Fixo)**.** *Considere o processo iterativo definido por $x_{k+1} = Mx_k + c$, onde M é uma matriz quadrada de ordem n e c é um vetor. Se M for uma matriz diagonalizável, ou seja, se existirem matrizes V inversível e D diagonal tais que $M = VDV^{-1}$, o método do ponto fixo converge para qualquer escolha de x_0 $\iff$ os autovalores λ_i de M satisfazerem:*

- $\forall i \in \{1, \ldots, n\}$, $\lambda_i \in (-1, 1]$;
- $\forall i \in \{1, \ldots, n\}$, *se* $(V^{-1}c)_i \neq 0$, *então* $\lambda_i < 1$.

Demonstração.

$$\begin{aligned}
& x_{k+1} = Mx_k + c \text{ converge para todo } x_0 \\
\iff \quad & \{ \text{ Indução } \} \\
& x_k = M^k x_0 + \sum_{i=0}^{n-1} M^i c \text{ converge para todo } x_0 \\
\iff \quad & \{ \ M = VDV^{-1} \ \}
\end{aligned}$$

[1] Naturalmente, diversas modelagens resultam diretamente na forma $x_{k+1} = Mx_k + c$, fazendo então com que uma mudança de forma do sistema seja desnecessária. Uma comparação — principalmente na etapa de modelagem — entre os problemas 1.1.2 e 1.3.6 pode ser feita e a similaridade entre sistemas na forma $Ax = b$ e $x_{k+1} = Mx_k + c$ fica ainda mais evidente. Duas formas de transformar um sistema da forma $Ax = b$ em um sistema na forma $x = Mx + c$ podem ser encontradas na seção 2.3.4.2.

$$
\begin{aligned}
& x_k = (VDV^{-1})^k x_0 + \sum_{i=0}^{n-1} (VDV^{-1})^i c \text{ converge para todo } x_0 \\
\iff & \quad \{ \text{ Teorema T22 } \} \\
& x_k = VD^kV^{-1}x_0 + \sum_{i=0}^{n-1} VD^iV^{-1}c \text{ converge para todo } x_0 \\
\iff & \quad \{ \ \bar{x}_0 = V^{-1}x_0 \ \} \\
& x_k = VD^k\bar{x}_0 + \sum_{i=0}^{n-1} VD^iV^{-1}c \text{ converge para todo } \bar{x}_0 \\
\iff & \quad \{ \text{ Distributiva } \} \\
& x_k = V(D^k\bar{x}_0 + \sum_{i=0}^{n-1} D^iV^{-1}c) \text{ converge para todo } \bar{x}_0 \\
\iff & \quad \{ \text{ Multiplicando pela inversa de } V \text{ (Não muda a convergência) } \} \\
& V^{-1}x_k = V^{-1}V(D^k\bar{x}_0 + \sum_{i=0}^{n-1} D^iV^{-1}c) \text{ converge para todo } \bar{x}_0 \\
\iff & \quad \{ \ V^{-1}V = I \text{ e } \bar{x}_k = V^{-1}x_k \ \} \\
& \bar{x}_k = D^k\bar{x}_0 + \sum_{i=0}^{n-1} D^iV^{-1}c \text{ converge para todo } \bar{x}_0 \\
\iff & \quad \{ \text{ Tomando } \bar{x}_0 = 0 \text{ e Aplicando propriedades do limite } \} \\
& \bar{x}_k = D^k\bar{x}_0 \text{ converge para todo } \bar{x}_0 \text{ e } \sum_{i=0}^{n-1} D^iV^{-1}c \text{ converge}
\end{aligned}
$$

Analisando ambas as partes separadamente. A primeira parte:

$$
\begin{aligned}
& \bar{x}_k = D^k\bar{x}_0 \text{ converge para todo } \bar{x}_0 \\
\iff & \quad \{ \text{ Mudando a representação do sistema para } \textit{point-full} \ \} \\
& \bar{x}_{k,i} = \lambda_i^k \bar{x}_{0,i} \text{ converge para todo } \bar{x}_{0,i},\ i \in \{1, \ldots, n\} \\
\iff & \quad \{ \text{ Convergência da progressão geométrica } \} \\
& \lambda_i \in (-1, 1],\ \forall i \in \{1, \ldots, n\}
\end{aligned}
$$

A segunda parte:

$$
\begin{aligned}
& \bar{x}_k = \sum_{i=0}^{n-1} D^iV^{-1}c \text{ converge} \\
\iff & \quad \{ \text{ Mudando a representação do sistema para } \textit{point-full} \ \} \\
& \bar{x}_{k,j} = \sum_{i=0}^{n-1} \lambda_j^i (V^{-1}c)_j \text{ converge} \\
\iff & \quad \{ \text{ Convergência da soma da progressão geométrica } \}
\end{aligned}
$$

$$
\begin{aligned}
& (V^{-1}c)_j = 0 \vee |\lambda_j| < 1,\ \forall j \in \{1, \ldots, n\} \\
\iff & \quad \{\ \text{Lógica}\ \} \\
& V^{-1}c \neq 0 \implies |\lambda_j| < 1,\ \forall j \in \{1, \ldots, n\}
\end{aligned}
$$

Assim, provamos as condições suficientes e necessárias para a convergência do método do ponto fixo. □

O teorema T10 se prova importante visto que ele cobre todos os casos possíveis de convergência do Método do Ponto Fixo, sob a hipótese que M é uma matriz diagonalizável, fazendo assim com que os casos onde $\forall i \in \{1, \ldots, n\}$, se $|\lambda_i| < 1$, então sabemos que a sequência converge e, como aprofundado na seção 2.3.4.3, também temos a garantia da convergência do caso onde um autovalor da matriz M é igual a 1.

2.3.4.2 Método de Gauss-Seidel e Gauss-Jacobi

Na seção 2.3.4.1 mostramos que, se um sistema no formato $Ax = b$ for convertido para um sistema no formato $x = Mx + c$ e se os autovalores de M respeitarem as condições suficientes e necessárias de convergência, então a sequência $x_{k+1} = Mx_k + c$ convergirá para o ponto fixo. Ainda assim, não mostramos exatamente como essa conversão pode ser feita. Veremos como essa conversão pode ser feita de duas maneiras diferentes.

Teorema T11. *Dado um sistema $Ax = b$, se A apresenta diagonal não nula, então o sistema pode ser reescrito como $x = Mx + c$ onde $M = -U^{-1}R$, $c = U^{-1}b$ e $A = U + R$, onde U é uma matriz triangular superior e $R = A - U$.*

Demonstração.

$$
\begin{aligned}
& Ax = b \\
\iff & \quad \{\ A = U + R\ \} \\
& (U + R)x = b \\
\iff & \quad \{\ \text{Distributiva}\ \} \\
& Ux + Rx = b \\
\iff & \quad \{\ \text{Subtraindo } Rx \text{ de ambos os lados}\ \} \\
& Ux = -Rx + b \\
\iff & \quad \{\ \text{Teorema T2}\ \} \\
& x = U^{-1}(-Rx + b) \\
\iff & \quad \{\ \text{Distributiva}\ \} \\
& x = -U^{-1}Rx + U^{-1}b \\
\iff & \quad \{\ M = -U^{-1}R \text{ e } c = U^{-1}b\ \} \\
& x = Mx + c
\end{aligned}
$$

□

Teorema T12. *Dado um sistema $Ax = b$, se A apresenta diagonal não nula, então o sistema pode ser reescrito como $x = Mx + c$ onde $M = -D^{-1}R$, $c = D^{-1}b$ e $A = D + R$, onde D é uma matriz diagonal e $R = A - D$.*

Demonstração.

$$
\begin{aligned}
& Ax = b \\
\iff & \quad \{\ A = D + R\ \} \\
& (D + R)x = b \\
\iff & \quad \{\ \text{Distributiva}\ \} \\
& Dx + Rx = b \\
\iff & \quad \{\ \text{Subtraindo } Rx \text{ de ambos os lados}\ \} \\
& Dx = -Rx + b \\
\iff & \quad \{\ \text{Teorema T2}\ \} \\
& x = D^{-1}(-Rx + b) \\
\iff & \quad \{\ \text{Distributiva}\ \} \\
& x = -D^{-1}Rx + D^{-1}b \\
\iff & \quad \{\ M = -D^{-1}R \text{ e } c = D^{-1}b\ \} \\
& x = Mx + c
\end{aligned}
$$

□

Apesar dessas conversões serem extremamente semelhantes e ambas precisarem que A apresente diagonal não nula, a diferença fundamental entre elas está no formato do vetor c e, principalmente, da matriz M. Enquanto a primeira conversão utiliza a matriz triangular superior U, a segunda faz uso da matriz diagonal D. Uma análise mais precisa sobre essas duas abordagens revela que a escolha entre U e D pode impactar a rapidez da convergência:

- Ao utilizarmos a decomposição $A = D + R$, onde D é a matriz diagonal, temos a forma iterativa $x_{k+1} = -D^{-1}Rx_k + D^{-1}b$. Como D é diagonal, cada componente de x_{k+1} é calculado exclusivamente a partir dos valores da iteração anterior (x_k), sem considerar atualizações feitas dentro da própria iteração. Isso significa que a atualização de cada componente ocorre apenas ao final de uma iteração completa. Esse método é conhecido como o método de Gauss-Jacobi.
- Ao utilizarmos a decomposição $A = L + R$, onde L é triangular inferior (incluindo a diagonal) e R é triangular superior estrita, temos a iteração $x_{k+1} = -L^{-1}Rx_k + L^{-1}b$. Como L é triangular inferior, sua inversa permite que os novos valores de x sejam computados sequencialmente e imediatamente usados na mesma iteração. Isso faz com que o método convirja mais rapidamente que o método de Gauss-Jacobi, pois ele sempre utiliza a informação mais atualizada disponível. Esse método é conhecido como o método de Gauss-Seidel.

Apesar do método de Gauss-Seidel convergir mais rapidamente, o método de Gauss-Jacobi se mostra interessante pois cada iteração do método é completamente independente, permitindo

que, pelo paralelismo, diferentes componentes de cada iteração possam ser calculados simultaneamente.

2.3.4.3 Método da Potência

Vimos na seção 2.3.2 que aplicar uma matriz $A = VDV^{-1}$ k vezes em um vetor inicial x é o mesmo que aplicar D^k no vetor inicial x escrito na base dos autovetores de A. Aplicar uma matriz D várias vezes em x (desde que não seja nulo) fará com que a coordenada do vetor resultante de índice igual à posição da matriz diagonal de maior valor em módulo prevaleça visto que essa coordenada crescerá mais rapidamente que as demais.[2] Além disso, como D representa os autovalores da matriz A e a matriz V^{-1} que o multiplica representa uma escrita na base dos autovetores de A, temos então que, ao aplicar D várias vezes em qualquer vetor inicial, esse vetor inicial tende ao autovetor de A associado ao maior autovalor, assim, servindo como uma espécie de proporção entre cada posição do vetor resultante após a aplicação de A várias vezes.

Teorema T13 (Convergência do Método da Potência)**.** *Dado uma matriz A com n autovalores λ_i ordenados de forma não-decrescente e seus n respectivos autovetores v_i, $\forall i \in \{1, \ldots, n\}$, e dado um vetor x, $\lim_{k\to\infty} A^k x = \lambda_1^k c_1 v_1$, sendo c_1 uma constante real.*

Demonstração.

$$
\begin{aligned}
& \lim_{k\to\infty} A^k x \\
= \quad & \{\ x = \sum_{i=1}^{n} c_i v_i \text{ (Combinação linear dos autovetores de } A)\ \} \\
& \lim_{k\to\infty} A^k \sum_{i=1}^{n} c_i v_i \\
= \quad & \{\ \text{Distributiva}\ \} \\
& \lim_{k\to\infty} \sum_{i=1}^{n} c_i A^k v_i \\
= \quad & \{\ A v_i = \lambda_i v_i\ \} \\
& \lim_{k\to\infty} \sum_{i=1}^{n} c_i \lambda_i^k v_i \\
= \quad & \{\ \text{Evidenciando o primeiro termo do somatório}\ \} \\
& \lim_{k\to\infty} c_1 \lambda_1^k v_1 + \sum_{i=2}^{n} c_i \lambda_i^k v_i \\
= \quad & \{\ \text{Colocando } \lambda_1^k \text{ em evidência}\ \}
\end{aligned}
$$

[2]Mais precisamente, os autovalores da matriz de transição A (ou seja, os valores diagonais de D) dizem muito a respeito do comportamento da população com o passar do tempo (ou então o que acontece com um vetor após a matriz ser aplicada a ele.). Perceba que, pelas fórmulas fechadas, sabemos que se $\lambda = 1$, então o seu respectivo autovetor v não crescerá conforme a matriz de transição for aplicada. Em outras palavras, a população está estável (Assumindo que o sistema seja $x_{k+1} = Ax_k$, sem uma constante sendo somada. No caso de uma constante sendo somada, verifique o teorema T10). De maneira análoga, se $\lambda > 1$, então a população está em crescimento contínuo e se $\lambda < 1$ então a população está em declínio.

$$
\begin{aligned}
& \lim_{k\to\infty} \lambda_1^k \left(c_1 v_1 + \sum_{i=2}^{n} \frac{c_i \lambda_i^k}{\lambda_1^k} v_i \right) \\
= \quad & \{\ \lambda_1 > \lambda_j \implies \lim_{k\to\infty} \left(\frac{\lambda_j}{\lambda_1}\right)^k = 0,\ \forall j \in \{2,\ldots,n\}\ \} \\
& \lim_{k\to\infty} \lambda_1^k c_1 v_1
\end{aligned}
$$

□

Dessa forma, temos que uma maneira iterativa de descobrir o autovetor associado ao maior autovalor em módulo de uma matriz A é pelo método conhecido como Método da Potência, onde, dado um vetor x_0 como chute inicial, desde que $x_0 \neq 0$, a matriz A é aplicada suficientes k vezes em x_0 até que $x_k = A^k x_0$ convirja para próximo do autovetor.

Ainda assim, veja que, pelo teorema T13, o termo $\lambda_1^k c_1$ acompanha o vetor x_k após a matriz A ser aplicada k vezes em x_0. Dessa maneira, se $|\lambda_1| > 1$, então o vetor x_0 crescerá após A ser aplicada e, se $|\lambda_1| < 1$, então x_0 diminuirá. Assim, uma forma de evitar que x_k convirja para o vetor nulo e nem para um vetor gigantesco que, apesar de ser aproximadamente um múltiplo de v, talvez não possa ser representado pelo computador, é, após cada aplicação de A em x_k, dividir x_k pela sua norma, garantindo assim que ele sempre seja um vetor unitário. Dessa maneira, uma escrita para o algoritmo que faz o especificado pelo método da potência seria:

```
function power_method(A, x, k)
  for i in 1:k
    x = Ax / norm(Ax)
  end
  return x
end
```

O teorema T13 nos diz que ao aplicar uma matriz várias vezes em um vetor não nulo, teremos que esse vetor estará cada vez mais próximo do autovetor associado ao maior autovalor em módulo. Vimos em seções como 1.3.4 e 1.3.5 a presença de matrizes probabilísticas nas quais não provamos nada a respeito de seus autovalores e, portanto não podemos ter certeza do que acontece exatamente com a convergência do método da potência ao aplicar a matriz várias vezes em um vetor inicial não-nulo.

Teorema T14. *Se P é uma matriz probabilística (Ou seja, a soma dos elementos de suas colunas é 1), então P apresenta 1 como um de seus autovalores.*

Demonstração.

Seja $v = \begin{bmatrix} 1 & \ldots & 1 & \ldots & 1 \end{bmatrix}^\mathsf{T}$.

$$
\begin{aligned}
& (P^\mathsf{T} v)_i \\
= \quad & \{\ \text{Definição}\ \} \\
& \sum_{j=1}^{n} P_{ij}^\mathsf{T} v_j \\
= \quad & \{\ v_j = 1,\ \forall j \in \{1,\ldots,n\}\ \}
\end{aligned}
$$

$$
\begin{aligned}
& \sum_{j=1}^{n} P_{ij}^{\mathsf{T}} \\
= & \quad \{ \text{ Definição de matriz probabilística (transposta) } \} \\
& 1
\end{aligned}
$$

Ou seja, $P^{\mathsf{T}}v = 1v$, em que v é autovetor de P^{T} e 1 é seu autovalor associado. E que, pelo teorema T30, temos então que v também é autovetor de P e 1 é seu autovalor associado. □

Teorema T15. *Se P é uma matriz probabilística, então seu maior autovalor em módulo é* 1.

Intuitivamente, sabemos que se a matriz P é probabilística, então a soma das probabilidades deve ser 1 e, portanto, para qualquer estado x_k em que o vetor inicial se encontra após k aplicações da matriz P, a distribuição de probabilidades sobre o possível futuro estado deve ser completa. Caso o maior autovalor em módulo da matriz P fosse maior do que 1, temos que o vetor x_k cresceria de maneira exponencial, o que viola a propriedade de conservação da soma das probabilidades.

Dessa forma, temos que, como o maior autovalor de uma matriz probabilística P é 1, então temos que, pelo teorema T10, o método converge e que, pelo teorema T13, converge para o autovetor associado ao maior autovalor de P.

Dessa forma, temos que o autovetor associado ao maior autovalor em módulo da matriz P representa a distribuição de probabilidade estável para a qual o sistema converge após muitas iterações. Ou seja, dado um vetor x_0 de distribuições iniciais, o vetor $P^k x_0$ representa uma aproximação cada vez melhor, conforme k aumenta, do estado estacionário (estável) do sistema.

2.3.4.4 Método do Gradiente Descendente

Nas seções 2.3.4.1, 2.3.4.2 e 2.3.4.3 vimos que aplicar matrizes iterativamente em vetores é uma possível estratégia para aproximar o ponto fixo de uma função, aproximar sistemas lineares ou aproximar vetores específicos. Outro método iterativo, mas dessa vez com o intuito que aproximar o valor mínimo de uma função, é o método do Gradiente Descendente.

Dado uma função $f(x) : \mathbb{R}^n \to \mathbb{R}$ que seja continuamente diferenciável em seu domínio, temos que o gradiente de f em x é definido como o vetor de derivadas parciais de f em relação ao vetor x. Ou seja,

$$
\nabla_x \, f = \left(\frac{\partial f}{dx_1}, \ldots, \frac{\partial f}{dx_i}, \ldots, \frac{\partial f}{dx_n} \right)^{\mathsf{T}}, \; \forall i \in \{1, \ldots, n\},
$$

com direção de maior crescimento da função f. Dessa forma, temos que $-\nabla_x \, f$ aponta na direção de maior decrescimento de f. Assim, a ideia por trás do método do Gradiente Descendente é, dado um chute inicial x_0, iterativamente calcular o gradiente de f no valor atual

do vetor de chute e ajustá-lo para um vetor mais próximo do vetor que minimiza a função.[3] Em outras palavras,

$$x_{k+1} = x_k - \alpha \nabla_x \ f(x_k),$$

onde α é uma constante que quantifica a intensidade do quanto na direção do gradiente o vetor x_{k+1} deve avançar. Intuitivamente, é notável que valores grandes de α farão com que pontos muito precisos de mínimo não possam ser alcançados, enquanto valores muito pequenos farão com que muitas iterações sejam necessárias para o método convergir.

No caso específico onde $n = 2$, é possível intuitivamente enxergar o que o método do Gradiente Descendente faz ao visualizar a figura F20 citada na seção 1.2.2, que é a mesma imagem da figura F39.

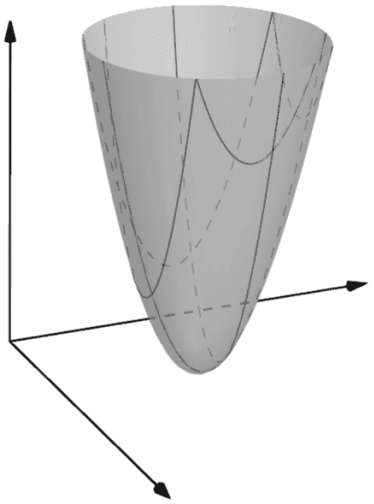


Figura F39: Paraboloide gerado pela função erro do caso da figura F19.

Supondo que a função da figura F39 seja a que queremos encontrar o mínimo[4], o que o método do gradiente fará é apontar sempre para o ponto de mínimo do paraboloide e "caminhar" em sua direção. Para exemplificar o comportamento do método do gradiente descendente, podemos fazer a seguinte analogia:

Imagine que o vetor x_k representa uma bola de gude colocada sobre o piso de um cômodo irregular, onde a altura em cada ponto é dada pela função $f(x)$. Nesse cenário:

- A altura do piso $f(x)$ representa o valor da função que queremos minimizar;

[3] Apesar do método do Gradiente Descendente poder ser utilizado para aproximar um ponto de mínimo local de qualquer função f, é extremamente comum que f seja uma função de erro, em que calcular o mínimo de f se torna crucial na medida que se busca minimizar o erro de uma aproximação.

[4] Não surpreendentemente, é comum que funções que representem o erro de um método apresentem um ponto de mínimo global, como é o caso da função de erro de aproximação da figura F19. Para funções como essas (que apresentam mínimo global), é bem conveniente que um método iterativo como o próprio gradiente descendente seja utilizado no lugar da fatoração QR, por exemplo, que apresenta maior complexidade computacional.

- A inclinação do piso em determinado ponto representa o valor do gradiente de f;
- Cada iteração do método equivale à passagem de um intervalo de tempo (em um tempo discretizado) em que a bola rola para a direção de maior inclinação em seu entorno. Ou seja, o vetor tende a um ponto de mínimo local da função f.

Após algumas iterações, é esperado que a bola se mova para um mínimo local — onde a inclinação da função f é zero. Além disso, veja que, se o piso apresentar diferentes pontos de mínimo, temos então que a bola de gude, desde que seja solta em um lugar aleatório do cômodo a cada novo lance, não irá parar sempre no mesmo lugar. Dessa forma, temos que uma possível melhoria para o método é realizá-lo algumas vezes com a esperança de que uma das iterações encontre um mínimo menor do que o mínimo encontrado anteriormente.

É possível que, apesar do método utilizar o gradiente da função, ele possa ser representado genericamente por um sistema matriz-vetor.[5] Supondo que, por exemplo, a função seja

$$f(\begin{bmatrix} x \\ y \end{bmatrix}) = 3x^2 + y + 3xy + 2,$$

temos então que:

$$\begin{aligned}
& -\alpha \nabla_x \ f(x_k) \\
= & \quad \{ \text{ Cálculando o gradiente de } f \ \} \\
& -\alpha \begin{bmatrix} 6x + 3y + 0 \\ 3x + 0y + 1 \end{bmatrix} \\
= & \quad \{ \text{ reescrita na forma matriz-vetor } \} \\
& -\alpha \begin{bmatrix} 6 & 3 \\ 3 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
= & \quad \{ \text{ Distributiva } \} \\
& \begin{bmatrix} -\alpha 6 & -\alpha 3 \\ -\alpha 3 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\end{aligned}$$

definindo $A = \begin{bmatrix} -\alpha 6 & -\alpha 3 \\ -\alpha 3 & 0 \end{bmatrix}$ e $c = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$:

$$\begin{aligned}
& x_{k+1} = x_k + \begin{bmatrix} -\alpha 6 & -\alpha 3 \\ -\alpha 3 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
\iff & \quad \{ \text{ Definição } \} \\
& x_{k+1} = Ax_k + x_k + c \\
\iff & \quad \{ \text{ Distributiva } \}
\end{aligned}$$

[5] Estamos assumindo que o gradiente da função apresenta apenas componentes lineares. Caso contrário, teríamos um sistema não-linear, que complicaria absurdamente o problema.

$$x_{k+1} = (A + I)x_k + c$$

que, por sua vez, apresenta formato $x_{k+1} = Mx_k + c$, onde $M = A + I$, que sabemos que converge sob determinadas condições a respeito do vetor c e a respeito dos autovalores da matriz M pelo teorema T10.

2.3.5 Exercícios de Modelos Dinâmicos

Exercício E59:

Sabemos que $a_{k+2} + 4a_k = 5a_{k+1}$ e que $a_4 = 1$ e $a_5 = 2$.

1. Determine uma fórmula fechada (fórmula rápida) para a_{100} usando a teoria de autovetores e autovalores.
2. Determine $\lim_{k\to\infty} \frac{a_k}{a_{k+2}}$.

Exercício E60:

Seja $\begin{bmatrix} 1 & -1 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{bmatrix}^k \begin{bmatrix} 1 \\ 7 \\ 5 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$.

Determine *exatamente* b_2 em função de k (uma fórmula fechada para b_2) com a teoria de autovalores e autovetores.

Exercício E61:

Num país politicamente instável 30% dos defensores da república passam a apoiar a monarquia a cada ano e 20% dos defensores da monarquia passa a apoiar a república a cada ano. Portanto, denotando por r_k e m_k o número de republicanos e monarquistas, respectivamente, a cada ano k.

1. Qual é o código para calcular r_k e m_k?
2. Sabendo que hoje metade da população apoia a república, em 10 anos qual será o percentual que apoia a república?
3. A longo prazo qual será o percentual de republicanos e monarquistas?

Exercício E62:

População de bactérias.

A população de uma certa espécie de bactéria pode ser compreendida da seguinte maneira. Existem bactéria novas, maduras e velhas. A cada mês:

- 80% das bactérias novas chegam à maturidade, e 20% morrem.
- 50% das bactérias maduras tornam-se velhas, e 50% morrem.
- 100% das bactérias velhas morrem.
- Uma a cada duas bactérias maduras geram uma nova bactéria.
- Uma a cada cinco bactérias velhas geram uma nova bactéria.

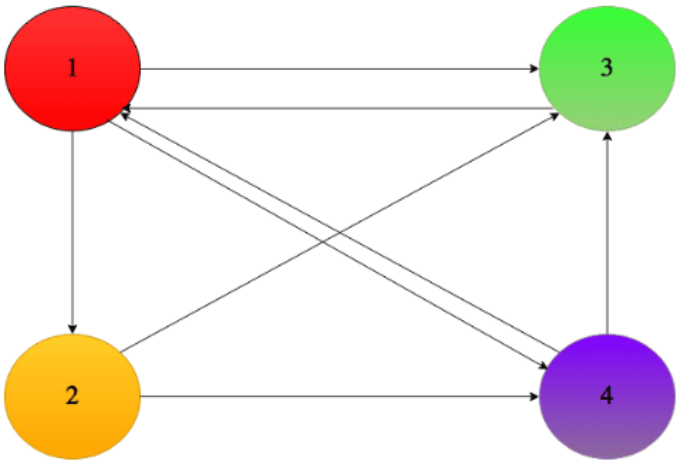
1. Modele o sistema populacional descrito acima – ou seja, determine o significado de cada coordenada do vetor que representa a população em um dado mês, e a matriz que representa a transição de um mês para o seguinte.
2. Se, no mês $t = 0$, existem apenas 250 bactérias novas (e 0 bactérias em outras faixas de idade), determine usando calculadora ou programa a população nos cinco primeiros meses.
3. Dada uma população inicial não-nula, suponha que, após um certo tempo, a proporção entre as populações de cada faixa de idade se estabilize. Determine essa proporção.
4. Qual método foi escolhido para encontrar essa proporção, e por quê ele funciona?

Exercício E63:

PageRank.

O modelo do surfista aleatório é muito usado quando se deseja realizar o ranqueamento de uma rede, não é a toa que o Google desenvolveu o *PageRank*!

Vamos pensar num caso simples onde, partindo de algum site, temos uma mesma probabilidade de visitar link de vizinhos. A estrutura da rede é a seguinte:



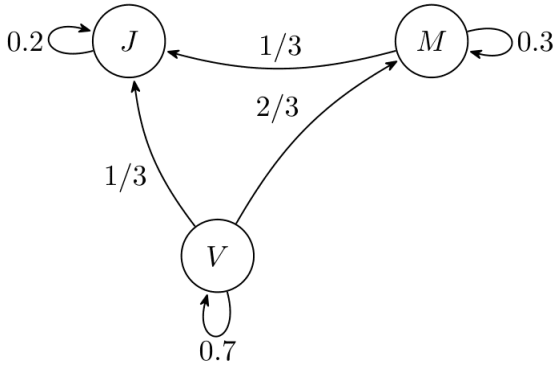
[H]

1. Descreva o sistema dinâmico linear representado acima, encontrando a matriz de transição que representa um "passo" ou "clique" do surfista aleatório.

2. Escreva um programa que simule esse sistema dinâmico. Realize a transição de um site para o outro 100 vezes para cada estado inicial possível. Os resultados são similares? Se sim, por quê?

Exercício E64:

Sabemos que a dinâmica de uma população de animais, dividida em 3 faixas etárias: jovens, maduros e velhos, por ano é determinada pelo modelo gráfico.



1. Existe algum valor de a que esses animais vão entrar em extinção para qualquer condição inicial?

2. Agora suponha que $a = 2$ e suponha também que sabemos que a proporção de jovens para maduros depois de muito tempo é aproximadamente de 2 para 1. Determine b.

Dica: use o fato que é fácil calcular determinante de matrizes triangulares.

Exercício E65:

Vampiros.

Suponha que vampiros existem. Será que as criaturas da escuridão seriam capazes de manter sua existência em segredo? Será que eles dominariam o mundo, ou acabariam sendo extintos? Suponha que a cada ano:

- A taxa de natalidade humana (nascimentos/população) é de 2%.
- A taxa de contaminação (humanos transformadas em vampiros/população) é de 1%.
- Muitas vezes, vampiros brigam entre si. Às vezes, a briga resulta em morte. A taxa de morte vampiresca por competição interna (vampiros mortos/população) é também de 1%.

1. Descreva matematicamente o sistema dinâmico linear descrito acima.

2. Alguma das duas espécies será extinta?

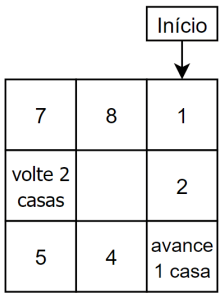
3. Suponha que, caso a população de vampiros passasse de 1% da população total (humanos + vampiros) do mundo, eles seriam descobertos. Esse evento aconteceria?

Exercício E66:

(Desafio) Considere o seguinte problema: uma empresária meio esquecida possui dois guarda-chuvas que ela usa no percurso de casa para ou trabalho e vice-versa. Se estiver chovendo e um guarda-chuva estiver disponível em seu local atual, ela o pega. Se não estiver chovendo, ela sempre esquece de levar o guarda-chuva. Suponha que a probabilidade de chuva é p a cada vez que ela se desloca de um lugar pra outro, independente de deslocamentos anteriores. Nosso objetivo é determinar a fração de deslocamentos com que a empresária irá se molhar com a chuva. Escreva um sistema linear dinâmico que descreva a situação.

Exercício E67:

Dado o seguinte tabuleiro de Banco Imobiliário:



- Determine qual é a casa sobre a qual é mais provável que o jogador termine uma rodada.
- Determine qual é a probabilidade do jogador terminar uma rodada em cada casa.

Exercício E68:

Dado

$$a_n = 2a_{n-1} + a_{n-2} - 2a_{n-3} \quad \text{e} \quad a_0 = 1,\ a_1 = 1 \text{ e } a_2 = 1,$$

determine a_{100} com um código rápido.

Exercício E69:

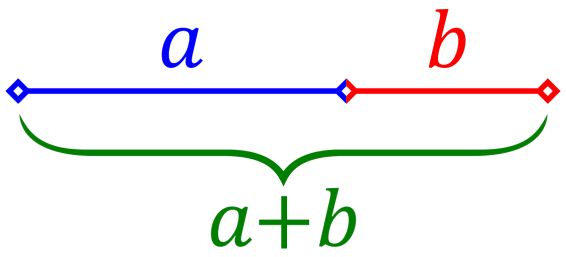
Sequência de Fibonacci.

A sequência de Fibonacci é definida pelas fórmulas:

$$F_0 = 0$$
$$F_1 = 1$$
$$F_{t+1} = F_t + F_{t-1}$$

Os 13 primeiros números da sequência são 0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144.

Esta famosa sequência tem uma profunda conexão com o número irracional ϕ, conhecido como Proporção Áurea. Esta proporção possui a seguinte propriedade geométrica:



$$\frac{a}{b} = \phi = \frac{a+b}{a}$$

1. Seja $v = \begin{bmatrix} F_t & F_{t+1} \end{bmatrix}^\mathsf{T}$ um vetor cuja primeira coordenada é um elemento da sequência e a segunda coordenada é o elemento seguinte. Determine qual é a matriz A que avança o vetor v ao longo da sequência, ou seja, $Av = A \begin{bmatrix} F_t & F_{t+1} \end{bmatrix}^\mathsf{T} = \begin{bmatrix} F_{t+1} & F_{t+2} \end{bmatrix}^\mathsf{T}$.
2. Determine os autovetores e autovalores da matriz A. Sabendo que o resultado da aplicação repetida de uma transformação linear tende ao autovetor de maior autovalor associado daquela transformação (*Método da Potência*), escreva em Português o que os autovetores e autovalores nos dizem sobre a sequência de Fibonacci e sua relação com a proporção áurea.
3. Dada a lista de números da sequência de Fibonacci acima, confira se as conclusões às quais você chegou no item anterior se verificam.

Exercício E70:

Seja $x = \begin{bmatrix} 1 & 0 \end{bmatrix}^\mathsf{T}$ e $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. Determine uma aproximação para z_1/z_2 tal que $z = A^{1000000}x$.

Exercício E71:

P é uma transformação que permuta de maneira cíclica uma lista de 5 números tal que

$$P \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{bmatrix}^\mathsf{T} = \begin{bmatrix} x_5 & x_1 & x_2 & x_3 & x_4 \end{bmatrix}^\mathsf{T}.$$

1. Determine um autovalor de P.
2. Determine o autovetor correspondente ao autovalor do item anterior.

Exercício E72:

Seja

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 1 & 1 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 4 \end{bmatrix}^{1000} \begin{bmatrix} 1 \\ 7 \\ 5 \\ 2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}.$$

Aproxime $\dfrac{b_1}{b_3}$.

Dica: use o fato de ser fácil calcular o determinante de uma matriz triangular.

Exercício E73:

Prove algebricamente que se A é uma matriz tal que os valores de cada coluna somam 1, então A tem um autovalor igual à 1. Use o fato que para qualquer matriz B, $\det\left(B^{\mathsf{T}}\right) = \det(B)$.

Dica: escreva isso matricialmente

Exercício E74:

Considere a função $f(x) = \frac{(x-7)^2}{6}$. Queremos minimizar essa função. Determine todos os tamanhos que o passo precisa ser para o método do gradiente descendente convergir para qualquer x inicial.

Exercício E75:

1. Faça dois passos do Método de Gauss-Jacobi para resolver o sistema

$$\begin{cases} 2x_1 + x_2 = 2 \\ \\ x_1 + x_2 = 2 \end{cases}$$

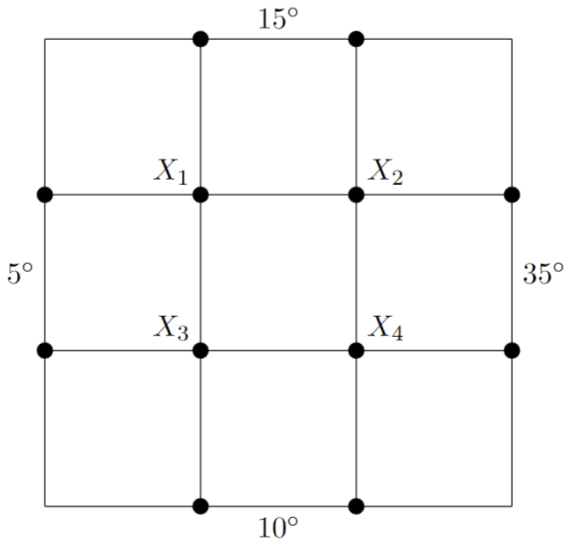
   com o chute inicial sendo o vetor nulo.
2. Usando autovalores, mostre se o método converge ou diverge.

Exercício E76:

Seja $f(x, y) = x^2 + y^2 - 2xy + 8x + 0.3y + 7$. Queremos minimizar essa função. Determine se o método do gradiente descendente converge para qualquer chute inicial com o tamanho do passo igual à 0.15.

Exercício E77:

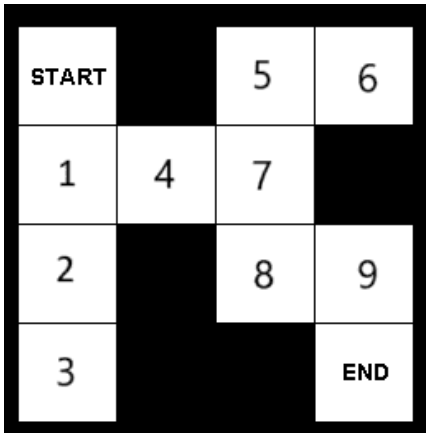
Problema da temperatura de um lago Queremos descobrir a temperatura em diferentes lugares no interior de um lago (vértices x_1, x_2, x_3 e x_4), mas só conseguimos medir a temperatura de 5, 10, 15 e 35 graus Celsius nas margens (laterais do quadrado na figura abaixo). Quando o calor está em equilíbrio, a temperatura em cada vértice no interior do lago é aproximadamente a média das temperaturas dos 4 vértices vizinhos.



1. Modele o problema como um sistema linear $Ax = b$.
2. Determine a temperatura nos 4 vértices do interior do quadrado com Gauss-Jacobi com 2 passos na mão e 50 passos no Julia.
3. Use a função *eigen* do Julia para ver se o método converge ou diverge?

Exercício E78:

Em um experimento, estamos medindo se um ratinho é realmente esperto e está aprendendo aonde está o queijo (END) ou se ele está andando aleatoriamente pelo labirinto. Essa semana, quando colocamos o ratinho no quadrado 8, ele acha o queijo (END) 70% das vezes. 30% das vezes ele volta para o começo (START). Vamos considerar que ele "perde o jogo", se ele volta para o começo.



Nesse exercício vamos calcular p_i que é a probabilidade da gente colocar o ratinho no quadro i e ele chegar no final (END) se ele estivesse andando aleatoriamente (com a mesma probabilidade de ir em qualquer direção). O rato é esperto ou não? Resolva esse problema em Julia com o método do ponto fixo.

Exercício E79:

Faça dois passos do Método de Jacobi para resolver o sistema

$$\begin{cases} 6z = 2 \\ \\ 2x + 5y + 3z = 5 \\ \\ 2y + 3z = 2 \end{cases}$$

com o chute inicial sendo o vetor nulo.

Exercício E80:

1. Faça dois passos do Método de Jacobi para resolver o sistema

$$\begin{cases} 2x_1 + x_2 = 2 \\ \\ -x_1 + 2x_2 = 2 \end{cases}$$

   com o chute inicial sendo o vetor nulo.

2. Usando autovalores, mostre se o método converge ou diverge.

3. Tem alguma maneira de ver que o método converge ou diverge nesse caso sem autovalores?

Exercício E81:

Considere o sistema linear

$$\begin{cases} 2x - y + z = -1 \\ \\ 2x + 2y + 2z = 4 \\ \\ -x - y + 2z = -5 \end{cases}$$

e mostre que o método de Jacobi não converge em Julia usando a função *eigen*.

Exercício E82:

Faça o gráfico do sistema linear

$$\begin{cases} 2x + y = 2 \\ \\ -x + 2y = 2 \end{cases} .$$

Desenhe as duas retas no plano cartesiano (eixo x e eixo y), desenhe a interseção das retas e depois desenhe os pontos (x, y) das iterações do método de Gauss-Jacobi no gráfico onde os eixos correspondem à x_1 e x_2 com chute inicial sendo o vetor nulo.

Exercício E83:

Considere o sistema linear

$$\begin{cases} 2x - y = -1 \\ \\ 2x + 2y = 4 \end{cases}$$

e determine se o método de Gauss-Seidel converge para qualquer chute inicial.

Exercício E84:

1. Faça dois passos do Método do ponto fixo para resolver o problema

$$\begin{cases} x_1 = 2 + x_1 + x_2 \\ \\ x_2 = 2 + x_1 \end{cases}$$

   com o chute inicial sendo o vetor nulo.

2. Usando autovalores, mostre se o método converge ou diverge.

Exercício E85:

Determine todos os valores de p tal que esse método diverge/converge.

```
function metodo(p)
  c = 0.3
```

```
  d = 0.7
  for i = 1:10000
    c = 2*c + 50 - p*d
    d = 0.8*d + 20
  end
  return c, d
end
```

Exercício E86:

Faça dois passos do Método de Jacobi para resolver o sistema

$$\begin{cases} 6z = 2 \\ \\ 2x + 5y + 3z = 5 \\ \\ 2y + 3z = 2 \end{cases}$$

com o chute inicial sendo o vetor nulo.

Exercício E87:

Descreva como você pode aproximar a solução de um sistema linear

$$\begin{bmatrix} 2 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \end{bmatrix}$$

com o gradiente descendente.

Dica: descreva uma função tal que o gradiente da função é zero no ponto (x_1, x_2) que resolve o sistema.

Exercício E88:

Use dois passos do método do gradiente descendente para encontrar um mínimo local de

$$f(x, y) = (y - 1)^4 + x^2y^2 + 1$$

começando no ponto $(1, 1)$ com o passo 0.1.

Exercício E89:

Seja

$$f(x, y) = 3x^2 + 4y^2 + 4xy + 6x + 10y + 7.$$

1. Use dois passos do método do gradiente descendente para encontrar um mínimo local de f começando no ponto $(1, 1)$ com o passo $p = 0.5$.
2. Usando o critério de convergência do ponto fixo, determine se o método do gradiente descendente converge para qualquer chute inicial com $p = 0.5$.

3. (Desafio) Determine todos os tamanhos que p pode ser para o método convergir para qualquer (x, y) inicial.

Exercício E90:

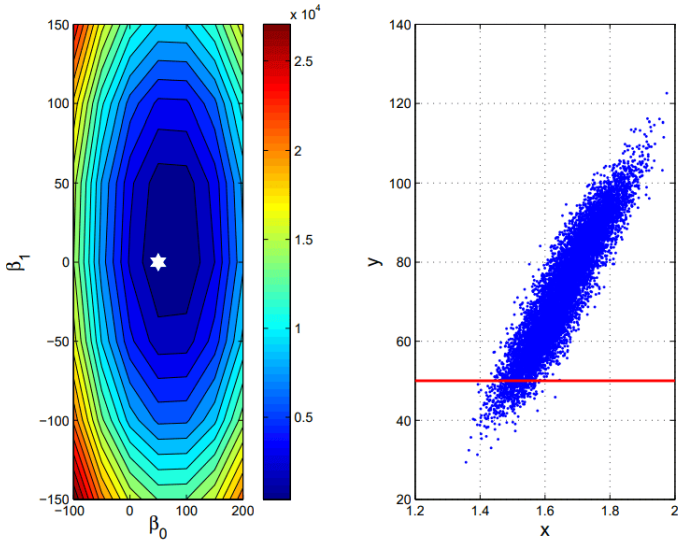
Determine qual é o tamanho que o passo precisa ser para o método convergir para qualquer x inicial considerando as seguintes funções :

1. $f(x) = \frac{(x-3)^2}{2}$.
2. $f(x) = \frac{(x-7)^2}{6}$.

Exercício E91:

Na figura abaixo, o ponto estrelado na esquerda corresponde a reta vermelha na direita. Estamos fazendo regressão linear na esquerda com o modelo que tem β_0 e β_1 como parâmetros.

1. Qual é o modelo linear sendo usado, $f(x) = \beta_0 + \beta_1 x$ ou $f(x) = \beta_1 + \beta_0 x$?
2. Estime o ponto na gráfico da esquerda aonde tem um ponto de mínimo usando o gráfico da direita.
3. Estime a direção do gradiente no ponto "estrelado".



Exercício E92:

Seja

$$f(x, y) = 2x^2 + y^2 + 2xy + 7x + 10y + 7$$

. Queremos achar um mínimo local de $f(x, y)$. O método do gradiente descendente converge para qualquer chute inicial com $p = 0.5$?

Exercício E93:

Seja a uma constante. Determine qual é a razão entre o módulo do vetor erro do método do ponto fixo na iteração 199 e na iteração 203 para o problema

$$\begin{cases} x_1 = x_1 + x_2 + 3 \\ \\ x_2 = -0.99x_1 - x_2 + a \end{cases}$$

com o chute inicial sendo o vetor nulo.

Exercício E94:

Já sabemos encontrar o autovetor associado ao maior autovalor. Agora, responda:

1. Como podemos achar o maior autovalor com o método da potência?
2. Como achar o segundo maior autovetor?
3. Como achar o menor autovetor com o método da potência?

Exercício E95:

Usei o método da potência no meu computador para a matriz

$$A = \begin{bmatrix} -4 & 18 & 18 \\ 14 & -37 & -40 \\ -14 & 42 & 45 \end{bmatrix}$$

com o vetor inicial $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$. O resultado foi alguma combinação linear de $\begin{bmatrix} 2.00001 \\ -5.9999 \\ 6.9999 \end{bmatrix}$. Determine um autovalor de $(A - 2I)^{-1}$.

Exercício E96:

Use dois passos do método do gradiente descendente para encontrar um mínimo local de

$$f(x, y) = (y - 1)^4 + x^2y^2 + 1$$

começando no ponto $(1, 1)$ com o passo$= 0.1$.

Exercício E97:

Determine qual é o tamanho que o passo precisa ser para o método convergir para qualquer x inicial considerando as seguintes funções:

1. $f(x) = \frac{(x-3)^2}{2}$.
2. $f(x) = \frac{(x-7)^2}{6}$.

Capítulo 2.4

Modelos de Reconhecimento de Padrões

2.4.1 Truque

Queremos encontrar $\text{argmax}_x \frac{\|Ax\|}{\|x\|}$ que, pelo teorema T37. No entanto, quando A é uma matriz densa, essa tarefa pode ser complexa por estarmos tratando de um problema de maximização bem acoplado. Para simplificar esse problema, buscaremos modificar a equação $\frac{\|Ax\|}{\|x\|}$ de forma que o problema de maximização seja o mais simples possível.

Veja que, se A for uma matriz diagonal, o sistema pode ser resolvido de forma simples e eficiente. Por exemplo, no caso (3×3):

$$\text{argmax}_x \frac{\left\| \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \right\|}{\left\| \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \right\|}.$$

Resolver esse sistema é simples pois o vetor x que maximiza a norma no numerador Ax da fração é um vetor seletor em que o índice não nulo i corresponde ao índice ii da matriz A que contém o maior elemento da diagonal principal.

Por exemplo, no caso da equação

$$\text{argmax}_x \frac{\left\| \begin{bmatrix} 3 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \right\|}{\left\| \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \right\|},$$

um vetor x que maximiza é $\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^\mathsf{T}$ visto que o numerador é maximizado com o valor 7 e o denominador fixado com o valor 1. Além disso, perceba que se o vetor fosse $\begin{bmatrix} 0 & 2 & 0 \end{bmatrix}^\mathsf{T}$, por exemplo, continuaríamos tendo 7 como o valor máximo da equação. Dessa forma, tanto o vetor $\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^\mathsf{T}$ quanto qualquer um de seus múltiplos fazem parte do conjunto dos vetores que maximiza essa equação. Em outras palavras,

$$\forall \lambda \in \mathbb{R},\ \lambda \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \in \text{argmax}_x \frac{\left\| \begin{bmatrix} 3 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \right\|}{\left\| \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \right\|}.$$

2.4.2 Troca de Variáveis

Por meio da decomposição em valores singulares, temos a garantia de que toda matriz A pode ser decomposta no formato $A = U\Sigma V^\mathsf{T}$, onde U e V^T são matrizes ortogonais e Σ é uma matriz diagonal.

Essa decomposição faz com que o cálculo de x tal que $x \in \text{argmax}_x \frac{\|Ax\|}{\|x\|}$ seja muito mais simples. Como V^T é uma matriz ortogonal, temos a garantia de que sua inversa V existe e que, consequentemente, podemos reescrever a equação $\text{argmax}_x \frac{\|Ax\|}{\|x\|}$ de maneira que o problema seja reduzido a outro muito mais simples.

Teorema T16 (Troca de Variáveis de Modelos de Reconhecimento de Padrões)**.** *Seja A uma matriz tal que $A = U\Sigma V^\mathsf{T}$, onde U e V^T são matrizes ortogonais e Σ é uma matriz diagonal e seja x um vetor. $argmax_x \frac{\|Ax\|}{\|x\|} = V\, argmax_x \frac{\|\Sigma x\|}{\|x\|}$.*

Demonstração.

$$\begin{aligned} & \text{argmax}_x \frac{\|Ax\|}{\|x\|} \\ = & \quad \{\ A = U\Sigma V^\mathsf{T}\ \} \\ & \text{argmax}_x \frac{\left\|U\Sigma V^\mathsf{T}x\right\|}{\|x\|} \\ = & \quad \{\ \text{Teorema T4}\ \} \end{aligned}$$

$$
\begin{aligned}
& \text{argmax}_x \ \frac{\left\|\Sigma V^{\mathsf{T}} x\right\|}{\|x\|} \\
= \quad & \{ \text{ Teorema T4 } \} \\
& \text{argmax}_x \ \frac{\left\|\Sigma V^{\mathsf{T}} x\right\|}{\left\|V^{\mathsf{T}} x\right\|} \\
= \quad & \{ \text{ Teorema T38 } \} \\
& V\text{argmax}_x \ \frac{\|\Sigma x\|}{\|x\|}
\end{aligned}
$$

□

Dessa forma, temos que se $A = U\Sigma V^{\mathsf{T}}$, sendo U e V^{T} matrizes ortogonais e Σ uma matriz diagonal, então a solução x tal que $x \in \text{argmax}_x \ \frac{\|Ax\|}{\|x\|}$ é a mesma solução x tal que $x \in V\text{argmax}_x \ \frac{\|\Sigma x\|}{\|x\|}$.

Essa troca de variáveis nos é vantajosa pois resolver o problema de maximização com uma matriz diagonal, conforme discutido na seção 2.4.1, é muito mais simples do que resolver a maximização com uma matriz densa.

A interpretação do passo a passo contido no teorema T16 pode ser vista como uma reescrita da matriz A na base dos vetores da matriz V^{T}, em que os vetores de A são expressos como múltiplos dos vetores canônicos. Além disso, a transformação V sobre os elementos do conjunto de vetores que maximiza essa razão ocorre devido à mudança de base, já que A não está mais representada na base original, mas sim na base associada a V^{T}.

2.4.3 Fatoração (SVD)

2.4.3.1 Objetivo e Ideia do SVD

A decomposição em valores singulares (SVD) consiste em fatorar uma matriz A em um produto de matrizes U, Σ e V^{T}, onde U e V^{T} são matrizes ortogonais e Σ é uma matriz diagonal.

Há diferentes formas de mostrarmos a existência da fatoração SVD. Uma delas é a partir do próprio problema que queremos resolver após nossa modelagem, que é $\text{argmax}_x \ \frac{\|Ax\|^2}{\|x\|^2}$. Podemos utilizar uma regra de cálculo como a regra do quociente ou o método dos multiplicadores de Lagrange para encontrarmos sob quais condições ambas as funções $f(x) = \|Ax\|^2$ de maximização quanto a função $g(x) = \|x\|^2$ de restrição são colineares. Dessa forma:

$$
\begin{aligned}
& \text{argmax}_x \ \frac{\|Ax\|^2}{\|x\|^2} \\
= \quad & \{ \text{ Teorema T37 } \}
\end{aligned}
$$

$$
\begin{aligned}
& \operatorname*{argmax}_{\|x\|=1} \|Ax\|^2 \\
= & \quad \{\ \|x\| = 1 \iff \|x\|^2 = 1\ \} \\
& \operatorname*{argmax}_{\|x\|^2=1} \|Ax\|^2 \\
\iff & \quad \{\ \text{Método de Lagrange}\ \} \\
& \nabla_x\ \|Ax\|^2 = \lambda \nabla_x\ \|x\|^2 - 1 \\
\iff & \quad \{\ \nabla_x\ -1 = 0\ \} \\
& \nabla_x\ \|Ax\|^2 = \lambda \nabla_x\ \|x\|^2 \\
\iff & \quad \{\ \text{Definição}\ \} \\
& \nabla_x\ (Ax)^\mathsf{T}(Ax) = \lambda \nabla_x\ \|x\|^2 \\
\iff & \quad \{\ \text{Teorema T20}\ \} \\
& \nabla_x\ x^\mathsf{T} A^\mathsf{T} Ax = \lambda \nabla_x\ \|x\|^2 \\
\iff & \quad \{\ \text{Teoremas T39 e T41}\ \} \\
& 2A^\mathsf{T} Ax = 2\lambda x \\
\iff & \quad \{\ \text{Dividindo por 2 dos dois lados}\ \} \\
& A^\mathsf{T} Ax = \lambda x \\
\iff & \quad \{\ \sigma^2 = \lambda\ \} \\
& A^\mathsf{T} Ax = \sigma^2 x
\end{aligned}
$$

Ou seja, encontramos que a direção de maior crescimento de $\dfrac{\|Ax\|^2}{\|x\|^2}$ ocorre quando:

- x é autovetor de $A^\mathsf{T}A$;
- σ^2 é autovalor de $A^\mathsf{T}A$.[1]

Além disso veja que, se multiplicarmos ambos os lados dessa equação por A:

$$
\begin{aligned}
& A^\mathsf{T} Ax = \sigma^2 x \\
= & \quad \{\ \text{Multiplicando por } A\ \} \\
& AA^\mathsf{T} Ax = \sigma^2 Ax \\
= & \quad \{\ \sigma u = Ax\ \} \\
& AA^\mathsf{T} \sigma u = \sigma^2 \sigma u \\
= & \quad \{\ \text{Dividindo por } \sigma\ \} \\
& AA^\mathsf{T} u = \sigma^2 u
\end{aligned}
$$

[1] O teorema T37 só pode ser utilizado nessa demonstração pois estamos interessados na direção do vetor (e não em sua magnitude). Com essa restrição, podemos dizer que os conjuntos $\operatorname{argmax}_x \dfrac{\|Ax\|^2}{\|x\|^2}$ e $\operatorname*{argmax}_{\|x\|^2=1} \|Ax\|^2$ são iguais.

Assim, temos que:

- u é autovetor de AA^T;
- σ^2 é autovalor de AA^T.

Além disso, por x e u poderem ser qualquer autovetor de $A^\mathsf{T}A$ e AA^T, respectivamente, então temos que $\forall i \in \{1,\ldots,n\}, j \in \{1,\ldots,m\}$, sendo n a quantidade de autovetores de $A^\mathsf{T}A$ e m a quantidade de autovetores de AA^T, então:

$$
\begin{aligned}
& Ax_i = \sigma_j u_j \\
= \quad & \{ \text{ Reescrevendo esse sistema na forma matriz-vetor } \} \\
& A \begin{bmatrix} | & & | & & | \\ x_1 & \ldots & x_i & \ldots & x_n \\ | & & | & & | \end{bmatrix} = \begin{bmatrix} | & & | & & | \\ u_1 & \ldots & u_j & \ldots & u_m \\ | & & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_j & & \\ & & & \ddots & \\ & & & & \sigma_m \end{bmatrix} \\
= \quad & \{ \text{ Reescrita na forma pointfree } \} \\
& AV = U\Sigma \\
= \quad & \{ \text{ Multiplicando por } V^\mathsf{T} \} \\
& AVV^\mathsf{T} = U\Sigma V^\mathsf{T} \\
= \quad & \{ \ VV^\mathsf{T} = I \text{ pelos teoremas T17 e T18 } \} \\
& A = U\Sigma V^\mathsf{T}
\end{aligned}
$$

Dessa forma, fatoramos A como sendo o produto de U, Σ e V^T, sendo, pelo teorema T18, U e V^T matrizes ortogonais e Σ uma matriz diagonal. O nome dessa decomposição é decomposição em valores singulares pois os vetores da matriz U são chamados de vetores singulares a esquerda, os valores de σ_j em Σ são chamados de valores singulares e os vetores da matriz V^T são chamados de vetores singulares a direita.

Essa fatoração se prova extremamente conveniente para a resolução do modelo de reconhecimento de padrões na medida que, por U e V^T serem matrizes formadas pelos dos autovetores de $A^\mathsf{T}A$ e AA^T, temos que, pelo teorema T17, sabemos que $A^\mathsf{T}A$ e AA^T são matrizes simétricas e, portanto, pelo teorema T18, sabemos que U e V^T são matrizes ortogonais.

Teorema T17. *Seja A uma matriz qualquer. $A^\mathsf{T}A$ e AA^T são matrizes simétricas.*

Demonstração.

Provando para $A^\mathsf{T}A$:

$$
\begin{aligned}
& (A^\mathsf{T}A)^\mathsf{T} \\
= \quad & \{ \text{ Teorema T20 } \} \\
& A^\mathsf{T}(A^\mathsf{T})^\mathsf{T} \\
= \quad & \{ \ (A^\mathsf{T})^\mathsf{T} = A \ \}
\end{aligned}
$$

$$A^\mathsf{T}A$$

Provando para AA^T:

$$\begin{aligned}
& (AA^\mathsf{T})^\mathsf{T} \\
= \quad & \{ \text{ Teorema T20 } \} \\
& (A^\mathsf{T})^\mathsf{T}A^\mathsf{T} \\
= \quad & \{ \ (A^\mathsf{T})^\mathsf{T} = A \ \} \\
& AA^\mathsf{T}
\end{aligned}$$

□

Teorema T18. *Se C é uma matriz simétrica, então seus autovetores associados a autovalores distintos são ortogonais entre si.*

Demonstração.

Sejam $\lambda_1, \ldots, \lambda_n$ autovalores de C com $\lambda_i \neq \lambda_j$ para $i \neq j$ tal que $i, j \in \{1, \ldots, n\}$, e que $v_1, \ldots, v_n$ sejam os autovetores correspondentes a $\lambda_1, \ldots, \lambda_n$, respectivamente. Assim, temos:

$$\begin{aligned}
& Cv_i = \lambda_i v_i \\
\iff \quad & \{ \text{ Teorema T23 } \} \\
& \langle Cv_i \,|\, v_j \rangle = \langle v_i \,|\, Cv_j \rangle \\
\iff \quad & \{ \text{ Definição } \} \\
& (Cv_i)^\mathsf{T} v_j = v_i^\mathsf{T}(Cv_j) \\
\iff \quad & \{ \ Cv_i = \lambda_i v_i \text{ e } Cv_j = \lambda_j v_j \ \} \\
& (\lambda_i v_i)^\mathsf{T} v_j = v_i^\mathsf{T}(\lambda_j v_j) \\
\iff \quad & \{ \text{ Linearidade do produto interno } \} \\
& \lambda_i (v_i^\mathsf{T} v_j) = \lambda_j (v_i^\mathsf{T} v_j) \\
\iff \quad & \{ \ \lambda_i \neq \lambda_j \ \} \\
& v_i^\mathsf{T} v_j = 0
\end{aligned}$$

Assim, temos que v_i e v_j são ortogonais sempre que $i \neq j$. Dessa forma, mostramos que os autovetores associados a autovalores distintos de uma matriz simétrica são ortogonais. □

Além disso, veja que:

$$\begin{aligned}
& A^\mathsf{T}A \\
= \quad & \{ \ A = U\Sigma V^\mathsf{T} \ \} \\
& (U\Sigma V^\mathsf{T})^\mathsf{T}(U\Sigma V^\mathsf{T}) \\
= \quad & \{ \text{ Teorema T20 } \}
\end{aligned}$$

$$
\begin{aligned}
& (V^\mathsf{T})^\mathsf{T}\Sigma^\mathsf{T}U^\mathsf{T}U\Sigma V^\mathsf{T} \\
= & \quad \{\ (V^\mathsf{T})^\mathsf{T} = V\ \} \\
& V\Sigma^\mathsf{T}U^\mathsf{T}U\Sigma V^\mathsf{T} \\
= & \quad \{\ U^\mathsf{T}U = I \text{ pelos teoremas T17 e T18}\ \} \\
& V\Sigma^\mathsf{T}\Sigma V^\mathsf{T} \\
= & \quad \{\ D = \Sigma^\mathsf{T}\Sigma\ \} \\
& VDV^\mathsf{T},
\end{aligned}
$$

que é a diagonalização de $A^\mathsf{T}A$, onde V é a matriz com os autovetores de $A^\mathsf{T}A$ e, pelo teorema T21, D é a matriz diagonal com o quadrado dos valores singulares de A. De maneira análoga, temos:

$$
\begin{aligned}
& AA^\mathsf{T} \\
= & \quad \{\ A = U\Sigma V^\mathsf{T}\ \} \\
& (U\Sigma V^\mathsf{T})(U\Sigma V^\mathsf{T})^\mathsf{T} \\
= & \quad \{\ \text{Teorema T20}\ \} \\
& U\Sigma V^\mathsf{T}(V^\mathsf{T})^\mathsf{T}\Sigma^\mathsf{T}U^\mathsf{T} \\
= & \quad \{\ (V^\mathsf{T})^\mathsf{T} = V\ \} \\
& U\Sigma V^\mathsf{T}V\Sigma^\mathsf{T}U^\mathsf{T} \\
= & \quad \{\ V^\mathsf{T}V = I \text{ pelos teoremas T17 e T18}\ \} \\
& U\Sigma\Sigma^\mathsf{T}U^\mathsf{T} \\
= & \quad \{\ D = \Sigma^\mathsf{T}\Sigma\ \} \\
& UDU^\mathsf{T}
\end{aligned}
$$

que é a diagonalização de AA^T, onde U é a matriz com os autovetores de AA^T e, pelo teorema T21, D é a matriz diagonal com o quadrado dos valores singulares de A.

De maneira geral, temos:

$$
A = \underbrace{\begin{bmatrix} | & & | & & | \\ u_1 & \dots & u_j & \dots & u_m \\ | & & | & & | \end{bmatrix}}_{U}
\underbrace{\begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_j & \\ & & & \ddots \\ & & & & \sigma_m \end{bmatrix}}_{\Sigma}
\underbrace{\begin{bmatrix} \text{———} & x_1 & \text{———} \\ & \vdots & \\ \text{———} & x_i & \text{———} \\ & \vdots & \\ \text{———} & x_n & \text{———} \end{bmatrix}}_{V^\mathsf{T}}.
$$

Dessa forma, temos que os teoremas T4 e T38 podem ser aplicados de maneira a simplificar o problema da maximização da razão $\dfrac{\|Ax\|}{\|x\|}$.

2.4.4 Extensões

2.4.4.1 Clusterização e K-means

Dado um conjunto de dados Ω tal que $\Omega = \bigcup_{i=1}^{n} A_i$ e $\bigcap_{i=1}^{n} A_i = \emptyset$, gostaríamos de separar Ω em n subconjuntos A_i, chamados de clusters, de tal forma que os elementos desse conjunto estejam agrupados por algum critério de semelhança.

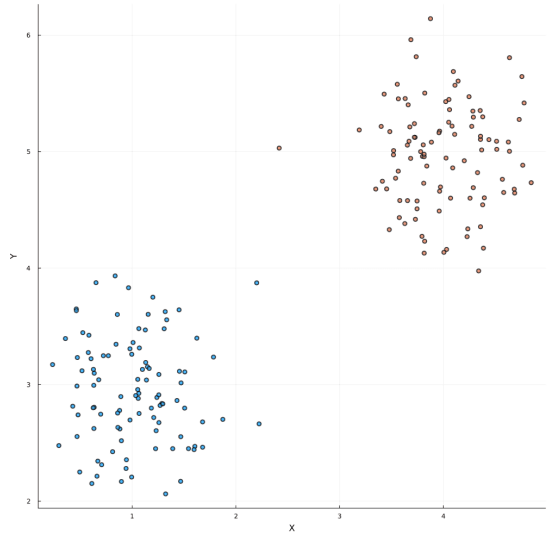


Figura F40: Exemplo de dados em $\mathbb{R}^2$ separados por cor pelo processo de clusterização.

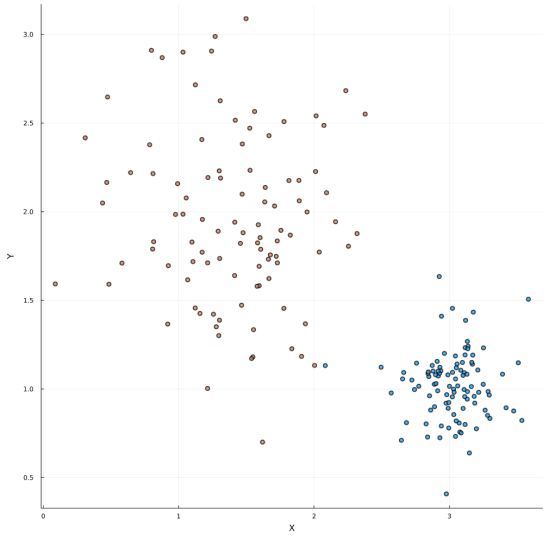


Figura F41: Outro exemplo de dados em $\mathbb{R}^2$ separados por cor pelo processo de clusterização.

Veja que para realizarmos um agrupamento com base em um critério de semelhança, precisamos definir precisamente o que essa significa essa semelhança. Definimos então que, dado n clusters e k pontos, com $k > n$, o ponto p_l pertence ao cluster A_i se a distância de p_l para o representante de A_i é menor ou igual a distância de p_l para o representante de A_j, $\forall i, j \in \{1, \ldots, n\}$ e $\forall l \in \{1, \ldots, k\}$.

Em outras palavras,

$$p_l \in A_i \iff \text{dist}(p_l, \text{rep}(A_i)) \leq \text{dist}(p_l, \text{rep}(A_j)) \ \forall i, j \in \{1, \ldots, n\} \text{ e } \forall l \in \{1, \ldots, k\}.$$

Dessa forma, podemos definir o erro da representação de dados por clusters em função de seus representantes como

$$\sum_{i=1}^{n} \sum_{a \in A_i} \|a - \text{rep}(A_i)\|^2,$$

sendo $\text{rep}(A_i)$ o ponto que representa o cluster A_i. Conforme provado em T27, sabemos

que, dado k pontos, o ponto que melhor os aproxima é a sua média. Sendo assim, definimos o representante de um cluster A_i como sendo a média entre todos os seus pontos.

Para resolver esse problema, utilizamos uma heurística conhecida como K-means. A ideia do algoritmo é: sortear n pontos aleatórios como os representantes dos n clusters e, em seguida, atribuir os pontos aos clusters com base na proximidade aos representantes. Após essa atribuição, recalculamos os representantes como a média dos pontos atribuídos a cada cluster. Se houver mudanças nos clusters, repetimos o processo até que a atribuição de pontos e os representantes se estabilizem.

Perceba que o algoritmo apresentado acima é uma heurística. Em outras palavras, é uma solução rápida, prática e eficiente mas sem a garantia de que o resultado encontrado é o melhor possível. No caso dessa heurística, temos a garantia que, a cada iteração do algoritmo, o erro não aumenta, mas não necessariamente a clusterização encontrada é a melhor possível visto que a função pode convergir para um mínimo local.

Teorema T19. *O erro do algoritmo K-means não aumenta a cada iteração do algoritmo.*

Demonstração.

Na etapa de atribuição, atribuímos cada ponto ao cluster que minimize a sua distância ao representante daquele cluster. Logo, O erro é minimizado ou mantido.

Na etapa de atualização dos representantes, o representante é escolhido de tal forma que o somatórios das distâncias ao quadrado de cada ponto ao representante seja o menor possível visto que, pelo teorema T27, temos que essa é a forma de calcular o representante de modo a diminuir o erro. Assim, fazendo com que o erro seja minimizado ou mantido.

Sendo assim, em cada etapa do nosso algoritmo, o ponto é atribuído ao cluster que minimiza ou mantém o seu erro e o representante de cada cluster é escolhido de tal forma que o erro seja minimizado ou mantido. Dessa forma, temos a garantia de que o erro do nosso método é sempre minimizado ou mantido a cada iteração do algoritmo.

Dessa forma, provamos que o erro não aumenta a cada iteração do algoritmo K-means. □

Além disso, veja que, como consequência da clusterização, temos uma aproximação matricial. Para representar k pontos, podemos considerá-los vetores coluna de uma matriz A, seus representantes com vetores coluna de uma matriz B e uma indicação de em qual cluster determinado ponto está por meio de uma matriz indicadora C. Sendo assim, temos

$$\underbrace{\begin{bmatrix} | & & | & & | \\ p_1 & \dots & p_l & \dots & p_k \\ | & & | & & | \end{bmatrix}}_{A} \approx \underbrace{\begin{bmatrix} | & & | & & | \\ r_1 & \dots & r_i & \dots & r_n \\ | & & | & & | \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} c_{1,1} & \dots & c_{1,l} & \dots & c_{1,k} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ c_{i,1} & \dots & c_{i,l} & \dots & c_{i,k} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ c_{n,1} & \dots & c_{n,l} & \dots & c_{n,k} \end{bmatrix}}_{C},$$

em que 1 único elemento de cada coluna da matriz C pode ser 1 e o restante tem que necessariamente ser 0.

Dessa forma, descobrimos uma forma de separar um grande conjunto de dados em conjuntos menores a partir de critérios de similaridade. Além disso, dado um novo dado nesse conjunto, saberíamos julgar a qual subconjunto ele pertence com base na sua distância até os representantes de cada cluster.

2.4.5 Exercícios de Modelos de Redução de Dimensionalidade

Exercício E98:

Use propriedades algébricas e a definição de autovalores e autovetores para provar que os autovalores de $A^\mathsf{T}A$ são todos positivos.

Exercício E99:

Determine a melhor fatoração de posto 1 de $A = \begin{bmatrix} 1 & 0 \\ 1 & 3 \\ 1 & 1 \end{bmatrix}$ e o erro associado usando o PCA.

Exercício E100:

Dado os pontos $a_1 = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}$, $a_2 = \begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix}$, $a_3 = \begin{bmatrix} -3 \\ 0 \\ -1 \end{bmatrix}$ e $a_4 = \begin{bmatrix} -1 \\ 0 \\ -3 \end{bmatrix}$, determine a redução dos pontos em dimensão 1 e 2 com o PCA.

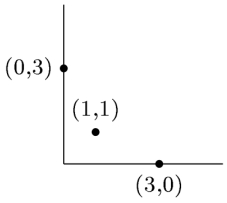
Exercício E101:

Seja B uma matriz simétrica e com autovalores distintos, prove que os seus autovetores são perpendiculares entre si (ortogonais entre si).

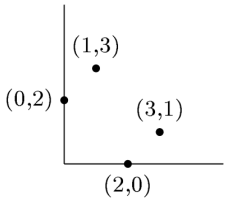
Exercício E102:

Determine uma matriz M de posto 1 (tal que $M = bc^\mathsf{T}$) que melhor representa a matriz A na norma de Frobenius:

1. $A = \begin{bmatrix} 0 & 1 & 3 \\ 3 & 1 & 0 \end{bmatrix}$



2. $A = \begin{bmatrix} 0 & 1 & 3 & 2 \\ 2 & 3 & 1 & 0 \end{bmatrix}$



Dica: utilize as simetrias do desenho.

Exercício E103:

Determine uma matriz de posto 1 que aproxima $A = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 1 \\ 3 & 0 & 0 \end{bmatrix}$ na norma de Frobenius melhor que $\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 3 & 6 & 3 \end{bmatrix}$. Justifique como você achou essa matriz.

Exercício E104:

Dadas duas tabelas (matrizes):

Matriz Usuário-gênero	**Drama**	**Ação**
User A	0.9	0.1
User B	0.8	0.2
User C	0.3	0.7
User D	0.6	0.4
User E	0.0	1.0

Matriz Filme-gênero	**Drama**	**Ação**
Titanic	0.9	0.1
Rocky	0.1	0.9
The Hobbit	0.5	0.5
Fight Club	0.0	1.0
Jurassic Park	0.2	0.8

1. Calcule a tabela (matriz) de usuários por filmes.
2. Usando a matriz do item (a), determine a aproximação de posto 1 em Julia usando a função autovalores e autovetores.
3. Desenhe os filmes em dimensão 1.
4. Desenhe os usuários em dimensão 1.
5. Qual filme você recomendaria para quem gostou de Titanic usando os itens anteriores?

Exercício E105:

Temos que $A = BC$, onde A é uma matrix $m \times n$, B é uma matrix $m \times p$ e C é uma matrix $p \times n$ tal que p é o menor valor possível (p é o posto). Em cada item, ache "no olho"(sem usar um algoritmo) matrizes B e C que não foram dadas.

1. Seja $A = \begin{bmatrix} 2 & 10 & 20 & -2 & 0 \\ 3 & 15 & 30 & -30 & 0 \\ 5 & 25 & 50 & -5 & 0 \\ 7 & 35 & 70 & -7 & 0 \end{bmatrix}$
2. Seja $A = \begin{bmatrix} 1 & 3 & 5 & 5000 & 31 \\ 4 & 5 & 9 & 9000 & 54 \\ 3 & 5 & 8 & 8000 & 53 \end{bmatrix}$

Exercício E106:

Filmes Seja U uma matriz com a preferência de 4 usuários por 5 filmes levando em consideração somente o nível de comédia:

$$U = \begin{bmatrix} 2 & 10 & 20 & 200 & -10 \\ -3 & -15 & -30 & -300 & 15 \\ 5 & 25 & 50 & 500 & -25 \\ 7 & 35 & 70 & 700 & -35 \end{bmatrix}$$

Determine uma possível solução para quanto cada usuário gosta (ou não gosta) de comédia e quanto cada filme é (ou não é) de comédia.

Exercício E107:

Considere a bandeira da Grécia como uma imagem azul (pixel azul vale 1) e branca (pixel branco vale 0) e modele com uma matriz A.

Figura F42: Bandeira da Grécia.

1. Qual é o posto da bandeira da Grécia (tal que $A = BC^{\mathsf{T}}$)? O que a matriz B e matriz C^{T} representam nesse caso? Explique com as suas próprias palavras.
2. Determine dois países tais quais suas bandeiras tem posto 1.
3. Determine dois países tais quais suas bandeiras tem posto 2.
4. Determine dois países tais quais suas bandeiras tem posto 3.

Exercício E108:

Qual é a melhor matriz de posto 1 que representa a matriz $A = \begin{bmatrix} 2 & 3 \\ 1 & 0 \end{bmatrix}$?

Exercício E109:

Escreva uma matriz $A_{(4\times6)}$ de posto 3 tal que a fatoração aproximada dela de posto 2 vai ter erro menor que 0.1.

Exercício E110:

Determine o primeiro componente principal v (em $\mathbb{R}^2$) de $A = \begin{bmatrix} 1 & 0 & 3 \\ 1 & 3 & 0 \end{bmatrix}$

Dica: Usa a simetria do gráfico para chutar o primeiro componente principal e depois faça com autovalores e autovetores.

Exercício E111:

Seja A uma matriz simétrica e com autovalores distintos, prove que os seus autovetores são perpendiculares.

Exercício E112:

Determine uma matriz M de posto 1 (tal que $M = bc^{\mathsf{T}}$) que melhor representa a matriz A e o erro:

1. $A = \begin{bmatrix} 0 & 1 & 3 \\ 3 & 1 & 0 \end{bmatrix}$
2. $A = \begin{bmatrix} 0 & 1 & 3 & 2 \\ 2 & 3 & 1 & 0 \end{bmatrix}$

Dica: usa a simetria dos pontos

Exercício E113:

Considere a seguinte imagem:

1. Insira esta imagem em uma matriz (5×5). Suponha que as sombras na imagem preta e branca que você vê são apenas valores 0, 0.5 e 1.
2. Qual é o posto da imagem?
3. Faça a análise em componentes principais com posto 1, 2, 3, 4 e 5 e verifique os seus erros de aproximação.

Exercício E114:

Considere a seguinte imagem.

Represente esta imagem em uma matriz $A_{(5\times5)}$ supondo que as sombras da figura são apenas valores 0, 0.5 e 1.

Determine a imagem da matriz B tal que $A = BC$.

Exercício E115:

A melhor aproximação de posto 1 da matriz A é xy^{T}. Determine a melhor matriz de posto 1 de A^{T}. Prove que sua resposta está correta.

Exercício E116:

Quanto se economiza de memória em porcentagem ao fatorar uma matriz 1000×1000 que tem posto 5 e armazenar a fatoração?

Exercício E117:

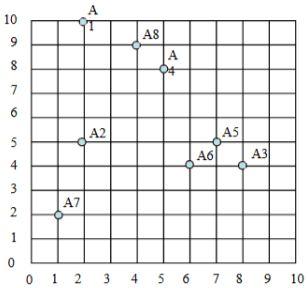
Seja $\begin{bmatrix} 2 & 2 & 10 \\ 5 & 6 & 25 \end{bmatrix} \approx \begin{bmatrix} 2 \\ 6 \end{bmatrix} \begin{bmatrix} 1 & 1 & 5 \end{bmatrix}$. Essa é a melhor fatoração de posto 1? Justifique sua resposta.

Exercício E118:

Encontre (ou desenhe) uma imagem A na internet, com mais ou menos 200×200 pixels (pode ser um pouco maior), que

1. exija mais que 3 componentes e menos que 6 no PCA para recuperar pelo menos 99% da norma de Frobenius total.
2. exija mais que 40 componentes para recuperar pelo menos 99% da norma de Frobenius total.

Exercício E119:



	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Use os pontos a_1, a_2, ..., a_8 e a tabela de distâncias acima para o exercício:

1. Faça só uma iteração (um passo) do algoritmo K-means com as sementes sendo a_1, a_2 e a_7.
2. Determine os clusters, os centróides, a fatoração matricial depois de uma iteração.
3. Represente graficamente o erro.

4. Quantas iterações você ia precisar fazer a mais até o algoritmo ter convergência?

Exercício E120:

Represente matricialmente o que seria pegar um dado com 100 pontos em dimensão 5, fazer uma redução de dimensionalidade com o PCA para duas dimensões e depois usar o K-means para clusterizar os pontos em dimensão 2.

Apêndice A

Teoremas de Algebra Linear

Teorema T20 (Transposta do Produto Entre Matrizes)**.** *Sejam $A_{m\times n}$ e $B_{n\times p}$ matrizes.* $(AB)^\mathsf{T} = B^\mathsf{T}A^\mathsf{T}$.

Demonstração.

$$
\begin{aligned}
& \left((AB)^\mathsf{T}\right)_{ij} \\
= & \quad \{ \text{ Definição de transposta } \} \\
& (AB)_{ji} \\
= & \quad \{ \text{ Definição de produto de matrizes } \} \\
& \sum_{k=1}^{n} A_{jk}B_{ki} \\
= & \quad \{ \text{ Substituição de transpostas } \} \\
& \sum_{k=1}^{n} (B^\mathsf{T})_{ik}(A^\mathsf{T})_{kj} \\
= & \quad \{ \text{ Definição de produto de } B^\mathsf{T}A^\mathsf{T} \} \\
& (B^\mathsf{T}A^\mathsf{T})_{ij}
\end{aligned}
$$

Portanto, provamos que $(AB)^\mathsf{T} = B^\mathsf{T}A^\mathsf{T}$. □

Teorema T21 (Produto entre Matrizes Diagonais)**.** *Se D e Λ são matrizes diagonais de dimensão $(n \times n)$, então $(D\Lambda)_{i,i} = D_{i,i}\Lambda_{i,i}$, $\forall i \in \{1, \ldots, n\}$.*

Demonstração.

$$
\begin{aligned}
& (D\Lambda)_{i,j} \\
= & \quad \{ \text{ Definição } \}
\end{aligned}
$$

$$
\begin{aligned}
& \sum_{l=1}^{n} D_{i,l}\Lambda_{l,j} \\
= & \quad \{\ D_{i,l} = 0 \text{ quando } l \neq i \text{ e } \Lambda_{l,j} = 0 \text{ quando } l \neq j\ \} \\
& \begin{cases} D_{i,i}\Lambda_{i,i}, & \text{se } i = j \\ \\ 0, & \text{se } i \neq j \end{cases}
\end{aligned}
$$

Portanto, provamos que $(D\Lambda)_{i,i} = D_{i,i}\Lambda_{i,i}$. □

Teorema T22 (Matriz Diagonal Elevada a um Natural)**.** *Se D é uma matriz diagonal de dimensão $(n \times n)$ e $k \in \mathbb{N}$, então $(D^k)_{i,j} = (D_{i,j})^k$, $\forall i \in \{1, \ldots, n\}$.*

Demonstração.

Caso Base: $k = 0$: temos que $D^0 = I$, que satisfaz $(D^0)_{i,j} = (D_{i,j})^0$.

Hipótese de Indução: k: $(D^k)_{i,j} = (D_{i,j})^k$.

Passo indutivo: $k + 1$:

$$
\begin{aligned}
& (D^{k+1})_{i,j} \\
= & \quad \{\ \text{Definição}\ \} \\
& (D^k D)_{i,j} \\
= & \quad \{\ \text{Teorema T21}\ \} \\
& D^k_{i,j} D_{i,j}
\end{aligned}
$$

□

Teorema T23 (Simétrica se e somente se Autoadjunta)**.** *Seja A uma matriz $(n \times n)$. $\forall x, y \in \mathbb{R}^n$, $\langle Ax \,|\, y\rangle = \langle x \,|\, Ay\rangle \iff A^\mathsf{T} = A$.*

Demonstração.

Precisamos mostrar ambas as implicações.

Quero mostrar que: $A^\mathsf{T} = A \implies \langle Ax \,|\, y\rangle = \langle x \,|\, Ay\rangle$:

$$
\begin{aligned}
& \langle Ax \,|\, y\rangle \\
= & \quad \{\ \text{Definição}\ \} \\
& (Ax)^\mathsf{T} y \\
= & \quad \{\ \text{Teorema T20}\ \} \\
& x^\mathsf{T} A^\mathsf{T} y \\
= & \quad \{\ A^\mathsf{T} = A\ \}
\end{aligned}
$$

$$
\begin{aligned}
& x^{\mathsf{T}} A y \\
= & \quad \{ \text{ Definição } \} \\
& \langle x \,|\, Ay \rangle
\end{aligned}
$$

Provamos que $A^{\mathsf{T}} = A \implies \langle Ax \,|\, y \rangle = \langle x \,|\, Ay \rangle$.

Quero mostrar que: $\langle Ax \,|\, y \rangle = \langle x \,|\, Ay \rangle \implies A^{\mathsf{T}} = A$

$$
\begin{aligned}
& A_{ij} \\
= & \quad \{ \text{ Reescrevendo o índice como produto interno } \} \\
& \langle e_i \,|\, Ae_j \rangle \\
= & \quad \{ \ \langle e_i \,|\, Ae_j \rangle = \langle Ae_i \,|\, e_j \rangle \ \} \\
& \langle Ae_i \,|\, e_j \rangle \\
= & \quad \{ \text{ Reescrevendo o produto interno como índice } \} \\
& A_{ji}
\end{aligned}
$$

Provamos que $\langle Ax \,|\, y \rangle = \langle x \,|\, Ay \rangle \implies A^{\mathsf{T}} = A$

Dessa forma, provamos que $\langle Ax \,|\, y \rangle = \langle x \,|\, Ay \rangle \iff A^{\mathsf{T}} = A$. □

Teorema T24. *A projeção de um vetor a em um vetor v é igual a* $\dfrac{\langle a \,|\, v \rangle}{\|v\|^2} v$.

Demonstração.

A projeção de a em v terá a direção de v logo:

$$\text{proj}_a(v) = \lambda v, \text{ sendo } \lambda \in \mathbb{R} \tag{1}$$

Sendo assim, podemos ver que:

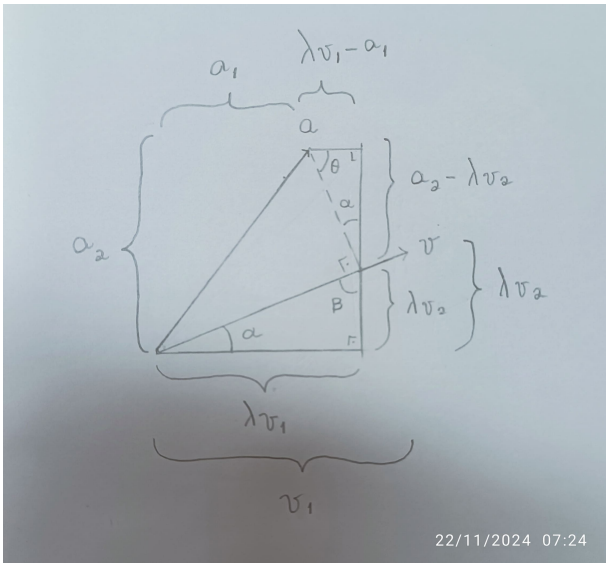


Figura F43: Representação dos comprimentos e ângulos gerados pela projeção de um vetor a em um vetor v.

Veja que $\alpha + 90° + \beta = 180°$ e $\alpha + 90° + \theta = 180$. Ou seja, temos que:

$$\begin{aligned} & \alpha + 90° + \beta = \alpha + 90° + \theta \\ \iff & \quad \{ \text{ Subtraindo } \alpha + 90° \text{ de ambos os lados } \} \\ & \theta = \beta \end{aligned}$$

Logo temos que os triângulos inferior e lateral superior direito são semelhantes. Sendo assim, por semelhança de triângulos temos que:

$$\begin{aligned} & \frac{\lambda v_2}{\lambda v_1} = \frac{\lambda v_1 - a_1}{a_2 - \lambda v_2} \\ = & \quad \{ \text{ Simplificando } \} \\ & \frac{v_2}{v_1} = \frac{\lambda v_1 - a_1}{a_2 - \lambda v_2} \\ = & \quad \{ \text{ Multiplicando por } v_1(a_2 - \lambda v_2) \text{ de ambos os lados } \} \\ & v_2(a_2 - \lambda v_2) = v_1(\lambda v_1 - a_1) \\ = & \quad \{ \text{ Distributiva } \} \\ & a_2 v_2 - \lambda v_2^2 = \lambda v_1^2 - a_1 v_1 \\ = & \quad \{ \text{ Somando } \lambda v_2^2 + a_1 v_1 \text{ de ambos os lados } \} \\ & \lambda v_1^2 + \lambda v_2^2 = a_2 v_2 + a_1 v_1 \\ = & \quad \{ \text{ Distributiva } \} \\ & \lambda(v_1^2 + v_2^2) = a_2 v_2 + a_1 v_1 \\ = & \quad \{ \text{ Definição } \} \\ & \lambda \|v\|^2 = \langle a \,|\, v \rangle \\ = & \quad \{ \text{ Dividindo ambos os lados por } \|v\|^2 \ \} \end{aligned}$$

$$\lambda = \frac{\langle a \,|\, v \rangle}{\|v\|^2}$$

Substituindo em (1):

$$\text{proj}_a(v) = \frac{\langle a \,|\, v \rangle}{\|v\|^2} v$$

Portanto, provamos a projeção de um vetor a em um vetor v é igual a $\dfrac{\langle a \,|\, v \rangle}{\|v\|^2} v$. □

Teorema T25. *O vetor que apresenta a menor distância entre a e v é o vetor gerado pela diferença de a pela projeção de a em v. Ou seja, dist*$(a, v) = a - proj_v(a)$.

Demonstração.

Perceba que a menor distância de um ponto a qualquer para um vetor v resulta da diferença do vetor que representa o ponto a para algum vetor que apresenta a mesma direção de v.

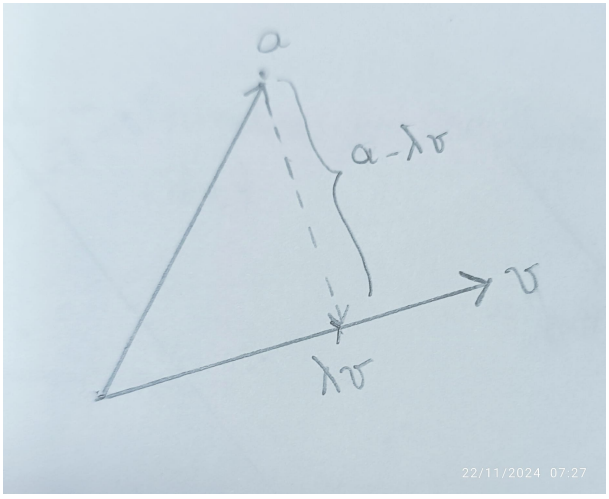


Figura F44: Representação da projeção de a em v.

Queremos calcular $\text{argmin}_\lambda \ \|a - \lambda v\|$. Ou seja, encontrar λ que minimize a distância de a a v. Sendo assim:

$$\begin{aligned}
& \text{argmin}_\lambda \ \|a - \lambda v\| \\
= \quad & \{ \text{ Teorema T31 } \} \\
& \text{argmin}_\lambda \ \|a - \lambda v\|^2 \\
= \quad & \{ \text{ Definição } \} \\
& \text{argmin}_\lambda \ (a - \lambda v)^\mathsf{T}(a - \lambda v) \\
= \quad & \{ \text{ Distributiva } \}
\end{aligned}$$

$$\text{argmin}_\lambda \; a^\mathsf{T}a - 2\lambda a^\mathsf{T}v + \lambda^2 v^\mathsf{T}v$$

Agora temos uma função de segundo grau em λ. Veja que a função apresenta concavidade para cima e, portanto, para encontrarmos o mínimo da função, podemos simplesmente encontrar o ponto em que sua derivada é igual a 0. Ou seja:

$$\begin{aligned}
& (a^\mathsf{T}a - 2\lambda a^\mathsf{T}v + \lambda^2 v^\mathsf{T}v)_\lambda = 0 \\
= \quad & \{ \text{ Derivada da soma é igual a soma das derivadas } \} \\
& (a^\mathsf{T}a)_\lambda - (2\lambda a^\mathsf{T}v)_\lambda + (\lambda^2 v^\mathsf{T}v)_\lambda = 0 \\
= \quad & \{ \ (a^\mathsf{T}a)_\lambda = 0, \, (2\lambda a^\mathsf{T}v)_\lambda = 2a^\mathsf{T}v \text{ e } (\lambda^2 v^\mathsf{T}v)_\lambda = 2\lambda v^\mathsf{T}v \ \} \\
& 0 - 2a^\mathsf{T}v + 2\lambda v^\mathsf{T}v = 0 \\
= \quad & \{ \text{ Somando } 2a^\mathsf{T}v \text{ de ambos os lados } \} \\
& 2\lambda v^\mathsf{T}v = 2a^\mathsf{T}v \\
= \quad & \{ \text{ Dividindo por 2 de ambos os lados } \} \\
& \lambda v^\mathsf{T}v = a^\mathsf{T}v \\
= \quad & \{ \text{ Dividindo por } v^\mathsf{T}v \text{ de ambos os lados } \} \\
& \lambda = \frac{a^\mathsf{T}v}{v^\mathsf{T}v} \\
= \quad & \{ \text{ Definição } \} \\
& \lambda = \frac{\langle a \,|\, v \rangle}{\|v\|^2}
\end{aligned}$$

Logo, O valor de λ que minimiza a distância entre a e λv é $\lambda = \frac{\langle a \,|\, v \rangle}{\|v\|^2}$. Veja também que isso significa que:

$$\lambda v = \frac{\langle a \,|\, v \rangle}{\|v\|^2} v$$

Que é a fórmula da projeção que encontramos em T24.

Logo, provamos que o vetor que representa a menor distância entre a e v é o vetor $a - \text{proj}_v(a)$. □

Teorema T26. *A menor distância entre um ponto a e a reta v é igual à* $\sqrt{\|a\|^2 - \left(\frac{\langle a \,|\, v \rangle}{\|v\|}\right)^2}$.

Demonstração.

Provamos em T25 que o vetor na direção de v que minimiza a distância até a é:

$$\text{proj}_v(a) = \frac{\langle a \,|\, v \rangle}{\|v\|^2} v$$

Sendo assim, visto que $\text{proj}_v(a)$ representa uma projeção ortogonal (ou seja, apresenta ângulo reto), podemos utilizar o teorema de Pitágoras tal que:

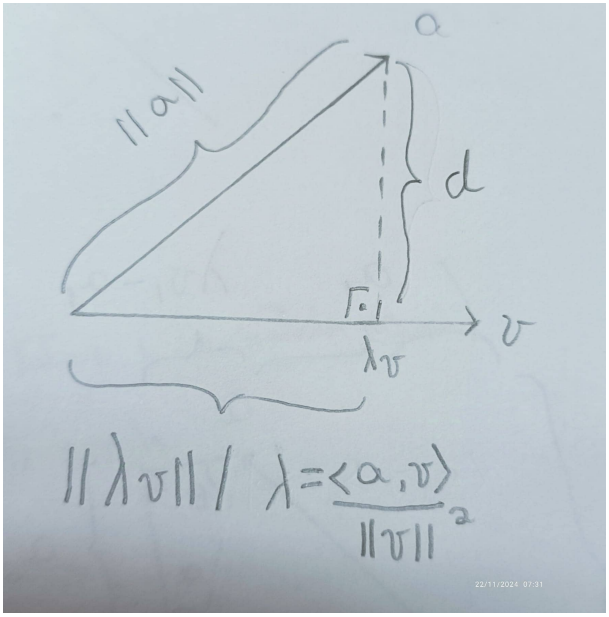


Figura F45: Representação da distância d entre um vetor a e sua projeção em um vetor v.

$$
\begin{aligned}
& \|a\|^2 = d^2 + \left\| \frac{\langle a \mid v \rangle}{\|v\|^2} v \right\|^2 \\
\iff & \quad \{\ \|\lambda v\|^2 = \lambda^2 \|v\|\ \} \\
& \|a\|^2 = d^2 + \left(\frac{\langle a \mid v \rangle}{\|v\|^2} \right)^2 \|v\|^2 \\
\iff & \quad \{\ \left(\frac{a}{b}\right)^2 = \frac{a^2}{b^2}\ \} \\
& \|a\|^2 = d^2 + \frac{\langle a \mid v \rangle^2}{\|v\|^4} \|v\|^2 \\
\iff & \quad \{\ \text{Simplificação}\ \} \\
& \|a\|^2 = d^2 + \frac{\langle a \mid v \rangle^2}{\|v\|^2} \\
\iff & \quad \{\ \frac{a^2}{b^2} = \left(\frac{a}{b}\right)^2\ \} \\
& \|a\|^2 = d^2 + \left(\frac{\langle a \mid v \rangle}{\|v\|} \right)^2 \\
\iff & \quad \{\ \text{Subtraindo } \left(\frac{\langle a \mid v \rangle}{\|v\|} \right)^2 \text{ de ambos os lados}\ \} \\
& d^2 = \|a\|^2 - \left(\frac{\langle a \mid v \rangle}{\|v\|} \right)^2 \\
\iff & \quad \{\ \text{Calculando a raiz quadrada de ambos os lados}\ \}
\end{aligned}
$$

$$d = \sqrt{\|a\|^2 - \left(\frac{\langle a \mid v \rangle}{\|v\|}\right)^2}$$

Assim, provamos que a menor distância entre um ponto a e uma reta v é $\sqrt{\|a\|^2 - \left(\frac{\langle a \mid v \rangle}{\|v\|}\right)^2}$.

□

Teorema T27 (Ponto que Melhor Representa um Conjunto de Pontos)**.** *Seja o erro entre um ponto p e sua aproximação c a distância entre eles. Dado um conjunto de n pontos p_i, $i \in \{1, \ldots, n\}$, em $\mathbb{R}^m$, o ponto c que minimiza a soma das distâncias (ou seja, o erro total) entre ele e os pontos dados é a média aritmética dos pontos.*

Demonstração.

$$
\begin{aligned}
& \operatorname{argmin}_c \sum_{i=1}^{n} \|p_i - c\|^2 \\
= \quad & \{ \text{ Expandindo o quadrado da norma } \} \\
& \operatorname{argmin}_c \sum_{i=1}^{n} (p_i - c)^\mathsf{T}(p_i - c) \\
= \quad & \{ \text{ Distribuindo o produto interno } \} \\
& \operatorname{argmin}_c \sum_{i=1}^{n} p_i^\mathsf{T} p_i - p_i^\mathsf{T} c - c^\mathsf{T} p_i + c^\mathsf{T} c \\
= \quad & \{ \text{ Usando que } p_i^\mathsf{T} c = c^\mathsf{T} p_i \ \} \\
& \operatorname{argmin}_c \sum_{i=1}^{n} \|p_i\|^2 - 2p_i^\mathsf{T} c + \|c\|^2 \\
= \quad & \{ \text{ Calculando o gradiente e igualando a zero } \} \\
& \sum_{i=1}^{n} \nabla_c \left(\|p_i\|^2 - 2p_i^\mathsf{T} c + \|c\|^2 \right) = 0 \\
= \quad & \{ \text{ Usando } \nabla_c \left(\|p_i\|^2\right) = 0,\ \nabla_c \left(-2p_i^\mathsf{T} c\right) = -2p_i, \text{ e } \nabla_c \left(\|c\|^2\right) = 2c \ \} \\
& \sum_{i=1}^{n} -2p_i + 2c = 0 \\
= \quad & \{ \text{ Separando o termo constante no somatório } \} \\
& \sum_{i=1}^{n} (-2p_i) + \sum_{i=1}^{n} (2c) = 0 \\
= \quad & \{ \text{ Fatorando } 2c \ \} \\
& \sum_{i=1}^{n} (-2p_i) + 2nc = 0 \\
= \quad & \{ \text{ Isolando } c \ \}
\end{aligned}
$$

$$
\begin{aligned}
& 2nc = \sum_{i=1}^{n} 2p_i \\
= \quad & \{ \text{ Dividindo por } 2n \text{ } \} \\
& c = \frac{1}{n} \sum_{i=1}^{n} p_i
\end{aligned}
$$

□

Teorema T28. *Se A é uma matriz que tem núcleo não trivial, ou seja, $N(A) \neq \{0\}$, então $det(A) = 0$.*

Teorema T29. *λ, $\lambda \in \mathbb{R}$, é autovalor de uma matriz $A \iff det(A - \lambda I) = 0$.*

Demonstração.

$$
\begin{aligned}
& \exists v \neq 0 \mid Av = \lambda v \\
\iff \quad & \{ \text{ } Iv = v \text{ } \} \\
& \exists v \neq 0 \mid Av = \lambda Iv \\
\iff \quad & \{ \text{ Subtraindo } \lambda Iv \text{ de ambos os lados } \} \\
& \exists v \neq 0 \mid Av - \lambda Iv = 0 \\
\iff \quad & \{ \text{ Distributiva } \} \\
& \exists v \neq 0 \mid (A - \lambda I)v = 0 \\
\iff \quad & \{ \text{ } v \neq 0 \implies A - \lambda I \text{ tem núcleo não trivial, logo, pelo Teorema T28 } \} \\
& \det(A - \lambda I) = 0
\end{aligned}
$$

□

Teorema T30 (Preservação dos Autovalores sob Transposição Matricial)**.** *Os autovalores de uma matriz A são iguais aos autovalores da matriz A^{T}.*

Demonstração.

$$
\begin{aligned}
& \lambda \text{ é autovalor de } A \\
\iff \quad & \{ \text{ Teorema T29 } \} \\
& \det(A - \lambda I) = 0 \\
\iff \quad & \{ \text{ } \det(A) = \det\left(A^{\mathsf{T}}\right) \text{ } \} \\
& \det\left((A - \lambda I)^{\mathsf{T}}\right) = 0 \\
\iff \quad & \{ \text{ Distributiva } \} \\
& \det\left(A^{\mathsf{T}} - \lambda I^{\mathsf{T}}\right) = 0 \\
\iff \quad & \{ \text{ } I^{\mathsf{T}} = I \text{ } \} \\
& \det\left(A^{\mathsf{T}} - \lambda I\right) = 0 \\
\iff \quad & \{ \text{ Teorema T29 } \}
\end{aligned}
$$

λ é autovalor de A^{T}

□

Apêndice B

Teoremas de Otimização

Teorema T31 (Preservação do argmax por Computar o Quadrado)**.** *Se* $\forall x,\ f(x) \geq 0$*, então* $argmin_x\ f(x) = argmin_x\ (f(x))^2$

Demonstração.

$$
\begin{aligned}
& \text{argmin}_x\ f(x) \\
= & \quad \{\ \text{Definição de mínimo}\ \} \\
& x' \text{ tal que } f(x') \leq f(x), \forall x \\
= & \quad \{\ \text{Elevando ao quadrado ambos os lados}\ \} \\
& x' \text{ tal que } (f(x'))^2 \leq (f(x))^2, \forall x \\
= & \quad \{\ \text{Definição de mínimo para } f^2(x)\ \} \\
& \text{argmin}_x\ (f(x))^2
\end{aligned}
$$

□

Teorema T32 (Preservação do argmax por Somar uma Constante)**.** *se* c *é uma constante, então* $argmin_x\ f(x) + c = argmin_x\ f(x)$ *e* $argmax_x\ f(x) + c = argmax_x\ f(x)$

Demonstração.

$$
\begin{aligned}
& \text{argmin}_x\ (f(x) + c) \\
= & \quad \{\ \text{Propriedade de translação em mínimos}\ \} \\
& x' \text{ tal que } f(x') + c \leq f(x) + c, \forall x \\
= & \quad \{\ \text{Substraindo } c \text{ de ambos os lados}\ \} \\
& x' \text{ tal que } f(x') \leq f(x), \forall x \\
= & \quad \{\ \text{Definição de mínimo de } f(x)\ \} \\
& \text{argmin}_x\ f(x)
\end{aligned}
$$

$$
\begin{aligned}
& \operatorname{argmax}_x \ (f(x) + c) \\
= & \quad \{ \text{ Propriedade de translação em máximos } \} \\
& x' \text{ tal que } f(x') + c \geq f(x) + c, \forall x \\
= & \quad \{ \text{ Substraindo } c \text{ de ambos os lados } \} \\
& x' \text{ tal que } f(x') \geq f(x), \forall x \\
= & \quad \{ \text{ Definição de máximo de } f(x) \ \} \\
& \operatorname{argmax}_x \ f(x)
\end{aligned}
$$

□

Teorema T33 (Preservação do argmin por Computar o Logarítmo)**.** *se* $\forall x,\ f(x) > 0$ *e* f *possui mínimo em um ponto* x'*, então* $argmin_x\ f(x) = argmin_x\ log\,(f(x))$

Demonstração.

$$
\begin{aligned}
& \operatorname{argmin}_x \ f(x) \\
= & \quad \{ \text{ Definição de mínimo } \} \\
& x' \text{ tal que } f(x') \leq f(x), \forall x \\
= & \quad \{ \text{ Aplicando logem ambos os lados, como logé crescente } \} \\
& x' \text{ tal que } \log\left(f(x')\right) \leq \log\left(f(x)\right), \forall x \\
= & \quad \{ \text{ Definição de mínimo de } \log\left(f(x)\right) \ \} \\
& \operatorname{argmin}_x \log\left(f(x)\right)
\end{aligned}
$$

□

Teorema T34 (Preservação do argmax por Computar o Logarítmo)**.** *se* $\forall x,\ f(x) > 0$ *e* f *possui mínimo em um ponto* x'*, então* $argmax_x\ f(x) = argmax_x\ log\,(f(x))$

Demonstração.

$$
\begin{aligned}
& \operatorname{argmax}_x \ f(x) \\
= & \quad \{ \text{ Definição de máximo } \} \\
& x' \text{ tal que } f(x') \geq f(x), \forall x \\
= & \quad \{ \text{ Aplicando logem ambos os lados, como logé crescente } \} \\
& x' \text{ tal que } \log\left(f(x')\right) \geq \log\left(f(x)\right), \forall x \\
= & \quad \{ \text{ Definição de máximo de } \log\left(f(x)\right) \ \} \\
& \operatorname{argmax}_x \log\left(f(x)\right)
\end{aligned}
$$

□

Teorema T35 (Inversão de argmin para argmax pela Troca de Sinal)**.** *$argmin_x\ f(x) = argmax_x\ -f(x)$*

Demonstração.

$$
\begin{aligned}
& \operatorname{argmin}_x\ f(x) \\
= & \quad \{\ \text{Definição de mínimo}\ \} \\
& x' \text{ tal que } f(x') \leq f(x), \forall x \\
= & \quad \{\ \text{Multiplicando por } -1 \text{ e invertendo a desigualdade}\ \} \\
& x' \text{ tal que } -f(x') \geq -f(x), \forall x \\
= & \quad \{\ \text{Definição de máximo de } -f(x)\ \} \\
& \operatorname{argmax}_x\ -f(x)
\end{aligned}
$$

□

Teorema T36 (Preservação do argmax por Multiplicar por uma Constante Positiva)**.** *se c é uma constante positiva, então $argmax_x\ cf(x) = argmax_x\ f(x)$*

Demonstração.

$$
\begin{aligned}
& \operatorname{argmax}_x\ cf(x) \\
= & \quad \{\ \text{Definição de máximo}\ \} \\
& x' \text{ tal que } cf(x') \geq cf(x), \forall x \\
= & \quad \{\ \text{Dividindo por } c \text{ de ambos os lados}\ \} \\
& x' \text{ tal que } f(x') \geq f(x), \forall x \\
= & \quad \{\ \text{Definição de máximo}\ \} \\
& \operatorname{argmax}_x\ f(x)
\end{aligned}
$$

□

Teorema T37. *Seja A uma matriz e x um vetor. $\underset{\|x\|=1}{argmax_x} \|Ax\| \supseteq \underset{x\neq 0}{argmax_x} \frac{\|Ax\|}{\|x\|}$*

Demonstração.

$$
\begin{aligned}
& \underset{\|x\|=1}{\operatorname{argmax}_x} \|Ax\| \\
= & \quad \{\ \text{Definição}\ \} \\
& \{x \mid \|x\| = 1 \wedge \forall v \mid \|v\| = 1, \frac{\|Ax\|}{\|x\|} \geq \frac{\|Av\|}{\|v\|}\} \\
= & \quad \{\ \|x\| = \|v\| = 1 \implies x \neq 0 \wedge v \neq 0 \wedge \|Ax\| = \frac{\|Ax\|}{\|x\|} \wedge \|Av\| = \frac{\|Av\|}{\|v\|}\ \}
\end{aligned}
$$

$$\{x \mid x \neq 0 \wedge \|x\| = 1 \wedge \forall v \mid v \neq 0 \wedge \|v\| = 1, \frac{\|Ax\|}{\|x\|} \geq \frac{\|Av\|}{\|v\|}\}$$

Escolha $w \mid \|w\| = 1$.

$$\begin{aligned}
& w \in \{x \mid x \neq 0 \wedge \|x\| = 1 \wedge \forall v \mid v \neq 0 \wedge \|v\| = 1, \frac{\|Ax\|}{\|x\|} \geq \frac{\|Av\|}{\|v\|}\} \\
\iff & \quad \{\ \|w\| = 1\ \} \\
& w \in \{x \mid x \neq 0 \wedge \forall v \mid v \neq 0 \wedge \|v\| = 1, \frac{\|Ax\|}{\|x\|} \geq \frac{\|Av\|}{\|v\|}\} \\
\iff & \quad \{\ \|v\| = 1 \iff \exists z \mid z = \lambda v\ \} \\
& w \in \{x \mid x \neq 0 \wedge \forall z \mid z \neq 0, \frac{\|Ax\|}{\|x\|} \geq \frac{\|A\lambda v\|}{\|\lambda v\|}\} \\
\iff & \quad \{\ \|\lambda v\| = |\lambda|\,\|v\|\ \} \\
& w \in \{x \mid x \neq 0 \wedge \forall z \mid z \neq 0, \frac{\|Ax\|}{\|x\|} \geq \frac{|\lambda|\,\|Av\|}{|\lambda|\,\|v\|}\} \\
\iff & \quad \{\ \text{Simplificando}\ \} \\
& w \in \{x \mid x \neq 0 \wedge \forall z \mid z \neq 0, \frac{\|Ax\|}{\|x\|} \geq \frac{\|Av\|}{\|v\|}\} \\
\iff & \quad \{\ \text{Definição}\ \} \\
& w \in \operatorname{argmax}_{\substack{x \\ \|x\| \neq 0}} \frac{\|Ax\|}{\|x\|}
\end{aligned}$$

Ou seja, todo elemento $w \in \operatorname{argmax}_{\substack{x \\ \|x\|=1}} \|Ax\|$ satisfaz $w \in \operatorname{argmax}_{x \neq 0} \frac{\|Ax\|}{\|x\|}$, o que significa que $\operatorname{argmax}_{\substack{x \\ \|x\|=1}} \|Ax\| \supseteq \operatorname{argmax}_{x \neq 0} \frac{\|Ax\|}{\|x\|}$. □

Teorema T38. *Se g for uma função inversível, então $argmax_x\ f(g(x)) = g^{-1}(argmax_x\ f(x))$*

Demonstração.

$$\begin{aligned}
& x \in \operatorname{argmax}_x\ f(g(x)) \\
\iff & \quad \{\ \text{Definição}\ \} \\
& \forall x' \mid f(g(x')) \leq f(g(x)) \\
\iff & \quad \{\ y = g(x') \text{ pois } g \text{ é sobrejetiva}\ \} \\
& \forall y \mid f(y) \leq f(g(x)) \\
\iff & \quad \{\ \text{Definição de argmax}\ \} \\
& g(x) \in \operatorname{argmax}_x\ f(x) \\
\iff & \quad \{\ z = g(x)\ \} \\
& \exists z \mid z = g(x) \wedge z \in \operatorname{argmax}_x\ f(x) \\
\iff & \quad \{\ z = g(x) \iff x = g^{-1}(z)\ \}
\end{aligned}$$

$$
\begin{array}{ll}
 & \exists z \mid x = g^{-1}(z) \wedge z \in \operatorname{argmax}_x f(x) \\
\iff & \quad \{ \text{ Reescrevendo como conjunto-imagem } \} \\
 & x \in g^{-1}(\operatorname{argmax}_x f(x))
\end{array}
$$

□

Apêndice C

Teoremas de Cálculo

Teorema T39 (Gradiente do Quadrado na Norma de um Produto Matriz-vetor)**.** *Seja x um vetor e A uma matriz.* $\nabla_x \ \|Ax\|^2 = \nabla_x \ x^\mathsf{T} A^\mathsf{T} Ax = 2A^\mathsf{T} Ax$

Teorema T40 (Gradiente do Produto Interno Entre Vetores)**.** *Sejam x e b vetores.* $\nabla_x \ x^\mathsf{T} b = b$.

Teorema T41 (Gradiente do Quadrado da Norma de um Vetor)**.** *Seja x um vetor.* $\nabla_x \ \|x\|^2 = 2x$.

Apêndice D

Símbolos e Notações

Segue uma lista de notações utilizadas nas diferentes seções do documento.

- $\|v\|$: Norma vetorial de um vetor v.
- $\|A\|_{\mathsf{F}}$: Norma de Frobenius de uma matriz A.
- $\det(A)$: Determinante de uma matriz A.
- $|x|$: Valor absoluto de x.
- $\langle v \,|\, w \rangle$: Produto interno entre vetores v e w.
- $\text{proj}_v(w)$: Projeção ortogonal do vetor w na direção do vetor v.
- $\min_f$: Valor de mínimo de uma função f.
- $\text{argmin}_x\ f(x)$: Conjunto com os valores de x no domínio de f tal que $f(x)$ é mínimo da função f.
- $\text{argmax}_x\ f(x)$: Conjunto com os valores de x no domínio de f tal que $f(x)$ é máximo da função f.
- $\nabla_v\ f$: Gradiente da função f em relação ao vetor v, representando o vetor de derivadas parciais de f em relação às componentes de v.
- $\text{rep}(A)$: Representante de um conjunto A, frequentemente usado para indicar um elemento único que caracteriza A em um sistema de representação.
- A^{T}: transposta de uma matriz A.
- $\bigcup_{i=1}^{n} A_i$: União dos conjuntos $A_1, A_2, ..., A_n$.
- $\bigcap_{i=1}^{n} A_i$: Interseção dos conjuntos $A_1, A_2, ..., A_n$.
- $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$: Representação de uma matriz.

- $\begin{bmatrix}1\\0\end{bmatrix}$ ou $\begin{bmatrix}1 & 0\end{bmatrix}^\intercal$: Representação de um vetor.
- $(m \times n)$: Dimensão de uma matriz com m linhas e n colunas ou de um vetor de tamanho m se $n = 1$.
- i, j: Letras geralmente utilizadas como índices de uma matriz, vetor ou lista.
- m, n: Letras geralmente utilizadas como sendo o valor máximo que um índice pode atingir ou tamanho (por exemplo, a dimensão de um espaço ou a quantidade de pontos).
- f_x: Derivada de uma função f na variável x.
- v, w, u, z, x (letras minúsculas): Vetores.
- A, B, C, M (letras maiúsculas): Matrizes.
- $\alpha, \beta, \gamma, \lambda$: Números reais ou ângulos.