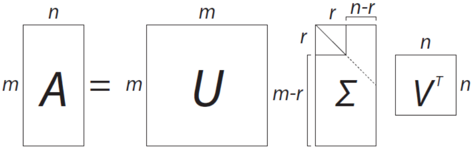
Decomposição em Valores Singulares (SVD)

Uma aplicação para *prunning* de redes neurais

André Pacheco
Universidade Federal do Espírito Santo



Novembro de 2016

SUMÁRIO

INTRODUÇÃO

- A decomposição em valores singulares (SVD) é uma importante ferramenta na ciência da computação, análise de dados e estatística
- É utilizada em diversas aplicações, dentre elas:
  - Reconhecimento de padrões
  - Filtro de sinais
  - Problemas de otimização
  - Compressão de dados
  - Recuperação da informação
- Além das aplicações práticas, os resultados da SVD produz diversos corolários na álgebra linear

INTRODUÇÃO

- A ideia principal da SDV é decompor uma matriz alvo no produto de três outras matrizes
  - O intuito é facilitar a resolução de um determinado problema
- O *pruning* de redes neurais é uma técnica que visa encontrar um número adequado de neurônios em uma camada oculta
  - Importante para o treinamento da rede
  - Evitar *overfitting* e convergência lenta

OBJETIVO

- O objetivo deste seminário é introduzir os conceitos da SVD e apresentar uma aplicação para *pruning* de redes neurais

CONCEITOS BÁSICOS

- Seja $\vec{v}$ um vetor de tamanho n, sua **norma** é definida como:

$$|\vec{v}| = \sqrt{\sum_{i=1}^{n} v_i^2} \tag{1}$$

- Seja $\vec{v^1}$ e $\vec{v^2}$ dois vetores de tamanho n, **o produto interno** (produto escalar ou *dot*) entre eles é defino como:

$$dot(\vec{v^1}, \vec{v^2}) = \sum_{i=1}^{n} = v_i^1 v_i^2 \tag{2}$$

CONCEITOS BÁSICOS

- Dois vetores são **ortogonais** se o produto interno entre eles é igual a 0
  - Em um espaço bidimensional é equivalente a dizer que os vetores são perpendiculares entre si, ou seja, o ângulo formado entre eles é de 90º
- Um **vetor normal** (ou unitário) é aquele com norma = 1. Qualquer vetor com norma > 0 pode ser normalizado dividindo o mesmo pela norma
  - Seja $\vec{v} = [2, 4, 1, 2]$, então $|\vec{v}| = \sqrt{2^2 + 4^2 + 1^2 + 2^2} = \sqrt{25} = 5$
  - Fazendo $\vec{u} = [\frac{2}{5}, \frac{4}{5}, \frac{1}{5}, \frac{2}{5}]$, $\vec{u}$ se torna normal, pois: $|\vec{u}| = \sqrt{\frac{2}{5}^2 + \frac{4}{5}^2 + \frac{1}{5}^2 + \frac{2}{5}^2} = \sqrt{1} = 1$

CONCEITOS BÁSICOS

- Sejam dois vetores **normais** $\vec{u}$ e $\vec{v}$. Se esses vetores **ortogonais** entre si, eles são chamados de **vetores ortonormais**. Por exemplo:
  - Seja $\vec{u} = [\frac{2}{5}, \frac{1}{5}, \frac{-2}{5}, \frac{4}{5}]$ e $\vec{v} = [\frac{3}{\sqrt{65}}, \frac{-6}{\sqrt{65}}, \frac{4}{\sqrt{65}}, \frac{2}{\sqrt{65}}]$
  - Ambos são normalizados, ou seja, $|\vec{u}| = 1$ e $|\vec{v}| = 1$
  - O produto interno $dot(\vec{u}, \vec{v}) = 0$
  - Os vetores $\vec{u}$ e $\vec{v}$ são **ortonormais**

CONCEITOS BÁSICOS

- Uma matriz A é **ortogonal** se os vetores que formam sua coluna são **ortogonais**. Por exemplo, seja
$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3/5 & -4/5 \\ 0 & 4/5 & 3/5 \end{bmatrix}$$
  - $A^T A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3/5 & -4/5 \\ 0 & 4/5 & 3/5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3/5 & 4/5 \\ 0 & -4/5 & 3/5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
- A matriz ortogonal A satisfaz as seguintes propriedades:
  - $A^T A = I$
  - $A^{-1} = A^T$
  - O produto de duas matrizes ortogonais gera outra matriz ortogonal

CONCEITOS BÁSICOS

- Um **autovetor** é um vetor não nulo que satisfaz a equação

$$A\vec{v} = \lambda\vec{v} \tag{3}$$

  - A é uma matriz quadrada, λ são os **autovalores** e $\vec{v}$ os autovetores
  - Os autovalores podem ser calculados solucionando o sistema linear obtido a partir da equação $\det(A - \lambda I) = 0$
  - Os autovetores associados aos autovalores são obtidos por meio da equação $(\lambda I - R)\vec{v} = 0$
- Os **valores singulares** de uma matriz são definidos como:

$$\sigma_j = \sqrt{\lambda_j(A^T A)} \tag{4}$$

CONCEITOS BÁSICOS

- **Processo de ortogonalização de Gram-Schmidt**
  - Tem como objetivo transformar um conjunto de vetores em vetores ortogonais
  - O algoritmo possui as seguintes regras:
    1. Organizar os vetores em uma matriz. Cada vetor será uma coluna;
    2. Normalizar o primeiro vetor;
    3. Iterativamente reescrever os demais vetores em termos deles mesmos menos os vetores já normalizados:

$$\vec{w}_k = \vec{v}_k - \sum_{i=1}^{k-1} dot(\vec{u}_i, \vec{v}_k) \times \vec{u}_i \tag{5}$$

sendo $\vec{w}_k$ o vetor resultante, $\vec{v}_k$ o vetor atual, $\vec{u}_i$ o vetor ortonormal anterior e k o número de vetores

CONCEITOS BÁSICOS

- Considere 3 vetores organizados na forma da matriz A, sendo cada coluna, um vetor:

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 2 & 0 \\ 2 & 3 & 1 \\ 1 & 1 & 0 \end{bmatrix} \tag{6}$$

- **Passo 1:** Normalizar o vetor $\vec{v_1} = [1, 0, 2, 1]$. Como $|\vec{v_1}| = \sqrt{6}$, o vetor normalizado de $\vec{v_1}$ será $\vec{u_1} = [\frac{1}{\sqrt{6}}, 0, \frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}]$
- **Passo 2:** Calcular os vetores restantes de acordo com a equação 4. Neste exemplo:

$$\vec{w_2} = \vec{v_2} - dot(\vec{u_1}, \vec{v_2}) \times \vec{u_1}$$
$$\vec{w_3} = \vec{v_3} - dot(\vec{u_2}, \vec{v_3}) \times \vec{u_2}$$

CONCEITOS BÁSICOS

- **Passo 2:** Calcular os vetores restantes de acordo com a equação 4.
  $\vec{w_2} = \vec{v_2} - dot(\vec{u_1}, \vec{v_2}) \times \vec{u_1}$
  $= [2, 2, 3, 1] - dot([\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, 0, \frac{2}{\sqrt{6}}], [2, 2, 3, 1]) \times [\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, 0, \frac{2}{\sqrt{6}}]$
  $\vec{w_2} = [\frac{1}{2}, 2, 0, \frac{-1}{2}]$
  Normalizando $\vec{w_2}$: $\vec{u_2} = [\frac{\sqrt{2}}{6}, \frac{2\sqrt{2}}{3}, 0, \frac{-\sqrt{2}}{6}]$
- Fazendo esse mesmo processo para $\vec{w_3} = \vec{v_3} - dot(\vec{u_2}, \vec{v_3}) \times \vec{u_2}$, é obtido: $\vec{u_3} = [\frac{2}{3}, \frac{-1}{3}, 0, \frac{-2}{3}]$
- A matriz ortogonal será: $A = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{\sqrt{2}}{6} & \frac{2}{3} \\ 0 & \frac{2\sqrt{2}}{3} & \frac{-1}{3} \\ \frac{2}{\sqrt{6}} & 0 & 0 \\ \frac{1}{\sqrt{6}} & \frac{-\sqrt{2}}{6} & \frac{-2}{3} \end{bmatrix}$

DECOMPOSIÇÃO EM VALORES SINGULARES (SVD)

- Dado uma matriz $A \in \mathbb{R}$, a SVD de A é a fatorização de A representada pelo produto de três matrizes:

$$A_{m \times n} = U_{m \times m} D_{m \times n} V^T_{n \times n} \tag{7}$$

  - U é uma matriz ortogonal, sendo que suas colunas são autovetores ortonormais de AA^T
  - V é uma matriz ortogonal, sendo que suas colunas são autovetores ortonormais de $A^T A$
  - D é uma matriz diagonal contendo a raíz quadrada dos autovalores de U ou V em ordem decrescente
  - A ortogonalização das matrizes é obtida por meio do **Processo de Gram-Schmidt**

DECOMPOSIÇÃO EM VALORES SINGULARES (SVD)

- Demonstrando o algoritmo por meio de um exemplo, considera a matriz:

$$A = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}_{2\times 3} \tag{8}$$

- Primeiramente, vamos calcular $U_{2\times 2}$

$$AA^T = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} \times \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}_{2\times 2}$$

  - Na sequência, precisamos calcular os autovalores e autovetores de AA^T
  - Realizando cálculos mais rápido do que a velocidade da luz, temos os autovalores 12 e 10, com autovetores associados sendo $[1, 1]$ e $[-1, 1]$
  - Esses autovetores se tornam colunas da matriz $\breve{U}$, ordenados pelos autovalores correspondentes (decrescente)

Decomposição em valores singulares (SVD)

- Portanto: $\breve{U} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$
- Para obter a matriz ortogonal U, é realizado o processo de Gram-Schmidt (calculando novamente mais rápido que a velocidade da luz):

$$U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}_{2\times 2} \tag{9}$$

- De maneira similar, calculamos a matriz ${V^T}_{3\times 3}$:

$$A^T A = \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix}_{3\times 3}$$

DECOMPOSIÇÃO EM VALORES SINGULARES (SVD)

- Os autovalores de $A^T A$ são: 0, 10 e 12. Os autovetores associados são: [1, 2, -5], [2, -1, 0] e [1, 2, 1], respectivamente
- Portanto: $\breve{V} = \begin{bmatrix} 1 & 2 & 1 \\ 2 & -1 & 2 \\ 1 & 0 & -5 \end{bmatrix}$
- Aplicando o processo de Gram-Schmidt e em seguida transpondo a matriz ortogonalizada, obtemos:

$$V^T = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix}_{3\times3} \tag{10}$$

DECOMPOSIÇÃO EM VALORES SINGULARES (SVD)

- Por fim, vamos obter a matriz $D_{2\times3}$
  - Utilizamos a raíz dos autovalores de U ou V
  - Os autovalores são colocados na diagonal principal em ordem decrescente
  - Os autovalores diferentes de zeros de U e V serão sempre iguais
  - A última coluna de zeros mantém as propriedades da multiplicação de matrizes

$$D = \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix}_{2\times3} \tag{11}$$

- A diagonal de D contém os valores singulares de A

DECOMPOSIÇÃO EM VALORES SINGULARES (SVD)

- Portanto, A foi decomposta nas 3 matrizes U, D e V^T:

$$A = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}_{2\times 2} \times \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix}_{2\times 3} \times \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix}_{3\times 3} \tag{12}$$

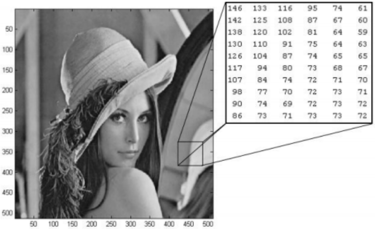
- Podemos visualizar A, da seguinte forma:

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T \tag{13}$$

Sendo r o posto da matriz

COMPRESSÃO DE DADOS COM SVD

- Assim como PCA, SVD também pode ser utilizado como uma maneira de comprimir dados
- Considere a imagem 512 $\times$ 512 *pixels* em níveis de cinza:

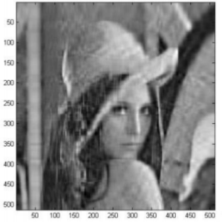


146	133	116	95	74	61
142	125	108	87	67	60
138	120	102	81	64	59
130	110	91	75	64	63
126	104	87	74	65	65
117	94	80	73	68	67
107	84	74	72	71	70
98	77	70	72	73	71
90	74	69	72	73	72
86	73	71	73	73	72

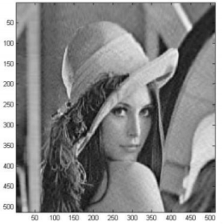
- Ao decompor a matriz X, que representa a imagem, os 512 valores singulares se distribuem entre $[5 \times 10^4, 2 \times 10^{-2}]$
- Portanto: $X = \sum_{j=1}^{512} \sigma_j u_j v_j^T$
- Cada componente vai contribuir de forma bem diferente devido aos valores singulares

COMPRESSÃO DE DADOS COM SVD

- Considerando os 20 primeiros (e maiores) valores singulares e na sequência os 50 primeiros, obtemos:



a) 20 valores singulares



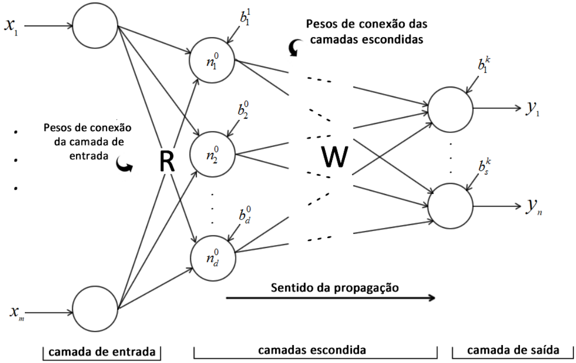
b) 50 valores singulares

Pruning DE REDES NEURAIS UTILIZANDO SVD

- Determinar o **número de neurônios de uma camada** de uma rede neural é estritamente ligado ao desempenho da mesma
  - Uma má escolha pode ocasionar *overfitting* ou não convergência do treinamento
- Uma maneira atacar este problema é **alargar** o número de neurônios no início do algoritmo e de acordo com o resultado **diminuir** este número (*pruning*)
- Em 1994, Psichogios e Ungar [2], propuseram o SVD-NET
  - Algoritmo que aplicava *pruning* de maneira automática para remodelar camadas escondidas

Pruning DE REDES NEURAIS UTILIZANDO SVD

- Considere uma rede neural de uma camada oculta



- A priori é escolhido um número grande de neurônios escondidos

Pruning DE REDES NEURAIS UTILIZANDO SVD

- Inicializado os pesos R e W de maneira aleatória, a SVD-NET atua da seguinte forma:
    1. Os pesos de R são ajustados por meio da minimização do erro quadrático. Os valores de W se mantém fixos;
    2. Conhecidos os valores de R, cada neurônio oculto n pode ser calculado a parti das entradas x, gerando uma matriz $N_{m \times d}$;
    3. Os pesos de W são ajustados minimizando $E = \mid R \times W - \hat{y} \mid$
    4. Aplicar SVD na matriz A em busca de neurônios ocultos redundantes
        - 4.1 Um número não nulo de valores singulares pequenos indicam redundância
        - 4.2 Aplicar *prunning* nos neurônios ocultos
- Os autores afirmam que o mesmo procedimento pode ser aplicado para redes de mais camadas

RESULTADOS

- No trabalho Abid e Najim [3], os autores aplicam essa metodologia para dois problemas simples: aproximar a função seno e reconhecer padrões de uma entrada
- Para ambos os casos, foram utilizadas uma rede neural tradicional e a SVD-NET. Ambas com apenas uma camada.

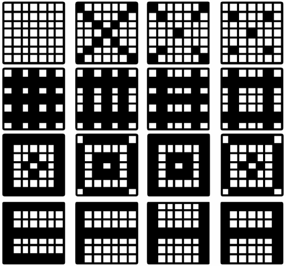
Problema I: aproximar $f(x) = sin(x)$ no intervalo $0 < x < 4\pi$

	Neurônios	Iteracoes para converg.	Tempo (s)
NN	15	2238	12.13
SVD-NET	11	1776	10.88

Table: Resultado para aproximação da função seno

RESULTADOS

Problema II: reconhecer os padrões apresentados na figura abaixo



	Neurônios	Iteracoes para converg.	Tempo (s)
NN	38	632	34.25
SVD-NET	29	479	29.36

Table: Resultado para o reconhecimento dos padrões

CONSIDERAÇÕES FINAIS

- O SVD é uma metodologia clássica da álgebra linear com aplicações em diversos problemas reais
- Utilizar o SVD para *pruning* de redes neurais foi proposto há mais de 2 décadas
  - É um problema relevante se tratando de redes neurais
  - Todavia, a metodologia não teve tanta aceitação
- Um dos motivos é nova técnicas mais recentes, como *dropout* que trata do mesmo problema de maneira mais fácil e rápida
- Por outro lado, SDV continua sendo bastante utilizada para compressão de dados, juntamente com PCA

REFERÊNCIAS

1. Baker, Kirk. "Singular value decomposition tutorial." The Ohio State University 24 (2005).
2. Psichogios, Dimitris C., and Lyle H. Ungar. "SVD-NET: An algorithm that automatically selects network structure." IEEE Transactions on Neural Networks 5.3 (1994): 513-515.
3. Abid, S., F. Fnaiech, and M. Najim. "A new Neural Network pruning method based on the singular value decomposition and the weight initialisation." Signal Processing Conference, 2002 11th European. IEEE, 2002.