UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
DISCIPLINA: REDES NEURAIS - DEEP LEARNING
Iª LISTA DE EXERCÍCIO) - 2024.1
Data de entrega da lista com apresentação: 04/04/2023

Objetivos da lista: consolidar os conceitos de: redes neurais, processos de aprendizagem de máquinas e estimular uma revisão dos conceitos de álgebra linear, cálculo de múltiplas variáveis, métodos de otimização e os fundamentos de aprendizagem de máquinas.

1-) As métricas de distancia entre vetores são aplicadas nos estudos de aprendizagem de máquina como medidas de similaridade/dissimilaridade entre vetores que representam padrões. Apresente um estudo sobre as seguintes distancias: Distancia Euclidiana, Distancia de Minkowski, Distancia City Block, Distancia de Mahalanobis, Coeficiente de Correlação de Pearson e Similaridade Cosseno. Apresente neste estudo aplicações onde cada tipo de métricas de distância é mais adequada.

Sugestão de Aplicações: Classificação de padrões, Clustering, Reconhecimento de padrões, etc.

2) Considere a função $E(\mathbf{w})$ onde $\mathbf{w} = [w_1, w_2, \ldots, w_n]^t$ é um vetor com múltiplas variáveis. Usando a expansão em série de Taylor a função pode ser expressa como $E(\mathbf{w}(n) + \mathbf{\Delta w}(n)) = E(\mathbf{w}(n)) + \mathbf{g}^t(\mathbf{w}(n))\mathbf{\Delta w}(n) + \frac{1}{2}\mathbf{\Delta w}^t(n)\mathbf{H}(\mathbf{w}(n))\mathbf{\Delta w}(n) + O(\|\mathbf{\Delta w}\|^3)$, onde $\mathbf{g}(\mathbf{w}(n))$ é o vetor gradiente local definido por $\mathbf{g}(\mathbf{w}) = \dfrac{\partial E(\mathbf{w})}{\partial \mathbf{w}}$ e $\mathbf{H}(\mathbf{w})$ é matriz Hessiana, definida por $\mathbf{H}(\mathbf{w}) = \dfrac{\partial^2 E(\mathbf{w})}{\partial \mathbf{w}^2}$.

Demonstre com base na expansão em série de Taylor:

a-) que o método do gradiente da descida mais íngreme é dado por

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \eta \mathbf{g}(\mathbf{w}(n)).$$

b-) que o método de Newton é dado por

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mathbf{H}^{-1}(\mathbf{w}(n))\mathbf{g}(\mathbf{w}(n)).$$

c-) que o passo ótimo é dado por $\mathbf{g}(\mathbf{w})\mathbf{g}^t(\mathbf{w})/\mathbf{g}^t(\mathbf{w})\mathbf{H}(\mathbf{w})\mathbf{g}(\mathbf{w})$, assumindo a matriz $\mathbf{H}$ definada positiva

d-) Qual será o valor do passo sob as condições acima se vetor gradiente estiver alinhado com a autovetor correspondente ao autovalor máximo.

3-) A matriz Hessiana pode ser expressa aproximadamente como $\mathbf{H}(n) = \sum_{k=1}^{n} \mathbf{g}(k)\mathbf{g}^t(k)$, onde $\mathbf{g}(\mathbf{w}(n))$ é o vetor gradiente local definido por $\mathbf{g}(\mathbf{w}) = \dfrac{\partial E(\mathbf{w})}{\partial \mathbf{w}}$, sendo E (w) a função custo.

a-) Demonstre que $\mathbf{H}(n) = \mathbf{H}(n-1) + \mathbf{g}(n)\mathbf{g}^t(n)$

b-) Usando o lema da inversão matricial tal que se $\mathbf{A} = \mathbf{B}^{-1} + \mathbf{CDC}^t$ então $\mathbf{A}^{-1} = \mathbf{B} - \mathbf{BC}(\mathbf{D} + \mathbf{C}^t\mathbf{BC})^{-1}\mathbf{C}^t\mathbf{B}$, sendo $\mathbf{A}$ e $\mathbf{B}$ matrizes definidas positivas, $\mathbf{C}$ e $\mathbf{D}$ são duas outras matrizes, demonstre que

$$\mathbf{H}^{-1}(n) = \mathbf{H}^{-1}(n-1) - \frac{\mathbf{H}^{-1}(n-1)\mathbf{g}(n)\mathbf{g}^t(n)\mathbf{H}^{-1}(n-1)^t}{1 + \mathbf{g}^t(n)\mathbf{H}^{-1}(n-1)\mathbf{g}(n)}$$

Isto é uma forma iterativa para gerar a matriz Hessiana com base na informação do gradiente, fazendo $\mathbf{H}(0)=\mathbf{I}$.

4-) Dada a função

$$E(w_1, w_2, w_3) = w_1^2 - w_2^2 + 2w_3^2 - 0.1w_1w_2 + 0.5w_1w_3 + 0.08w_2w_3 - 0.2w_1 + 0.3w_2 - 0.7w_3 + 0.4$$

a) represente a mesma na forma matricial,

$E(\mathbf{w}) = \mathbf{w}^\mathbf{t}\mathbf{A}\mathbf{w} + \mathbf{b}^\mathbf{t}\mathbf{w} + a$, onde $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3]^\mathbf{t}$, onde $a \in \mathbf{R}, \mathbf{b} \in \mathbf{R}^3$ e $\mathbf{A} \in \mathbf{R}^{3\times 3}$.

b) determine o seu gradiente, $\mathbf{g}(\mathbf{w}) = \dfrac{\partial E(\mathbf{w})}{\partial \mathbf{w}}$

c) determine a sua matriz Hessiana, $\mathbf{H}(\mathbf{w}) = \dfrac{\partial^2 E(\mathbf{w})}{\partial \mathbf{w}^2}$

d) verifique se esta função é estritamente convexa

e) Considerando que em um determinado ponto **w o** gradiente da função E(**w)** é igual zero,isto é, **g(w)=0. Isto não é suficiente para indicar que este ponto corresponde a um ponto de mínimo, máximo ou um ponto de sela. Com base** nos autovalores da matriz Hessiana $\mathbf{H}(\mathbf{w}) = \dfrac{\partial^2 E(\mathbf{w})}{\partial \mathbf{w}^2}$ apresente a condição para este ponto ser: (i) um ponto de mínimo, (ii) um ponto de máximo (iii) um ponto de sela da função.

5-) Apresente um estudo dos seguintes algoritmos de otimização: **Gradiente Estocástico (SGM)**, **AdaGrad, RMSProp e Adam.** Estes métodos ou otimizadores são utilizados no processo de aprendizagem de redes neurais/deep learning.

6-) O modelo de neurônio artificial de Mc-Culloch-Pitts faz uso da função de ativação para resposta do neurônio artificial. As função sigmoíde (ou função logística) e a função tangente hiperbólica (tangsigmoíde) são normalmente utilizadas nas redes neurais perceptrons de múltiplas camadas tradicionais (uma ou duas camadas ocultas - shallow network). A função ReLu (retificador linear) é normalmente utilizadas nas camadas ocultas das redes Deep Learning. Segue abaixo as expressões matemáticas de cada uma:

a-) $\varphi(v)=\dfrac{1}{1+\exp(-av)}$ (sigmoíde); b-) $\varphi(v)=\dfrac{1-\exp(-av)}{1+\exp(-av)}=\tanh(\dfrac{av}{2})$ (tangente hiperbólica ou tangsigmoíde) ; c-) $\varphi(v)=\max(0,v)$ (Re-Lu).

i) Faça uma análise comparativa de cada uma destas funções apresentando de forma gráfica a variação da função e da sua derivada com relação a v (potencial de ativação)

ii) Mostre que $\varphi'(v)=\dfrac{d\varphi(v)}{dv}=a\varphi(v)[1-\varphi(v)]$ para função sigmoíde.

iii) Mostre que $\varphi'(v)=\dfrac{d\varphi(v)}{dv}=\dfrac{a}{2}[1-\varphi^2(v)]$ para função tangsigmoíde.

7-) As funções de saída das redes neurais dependem do modelo que se busca gerar com a rede. Faça uma análise sobre a escolhas destas funções considerando os seguintes problemas: (i) Classificação de padrões com duas classes. (iii) Classificação de padrões com múltiplas classes. (iii) Problema de regressão (aproximação de funções)

8-) O método da máxima verossimilhança (Max-Likelihood) é aplicado na determinação de parâmetros desconhecidos de um modelo de distribuição de probabilidades. O método é bastante utilizado para o desenvolvimento de algoritmos de aprendizagem de máquinas em particular em redes neurais. O problema considerado nesta questão envolve o modelo probabilístico de uma distribuição gaussiana multivariada . Determine os parâmetros estimados da distribuição gaussiana.

9-) Implementações Computacionais de Redes Neurais.

Para cada um dos problemas abaixo apresente a solução fazendo uso de redes. Apresente na solução a curva do erro de treinamento e o erro de validação:

9.1-) Defina a estrutura de uma rede perceptron de múltiplas camadas para aproximar a

$$f(\mathbf{x})=16x_1^2+x_1x_2+8x_2^2-x_1-x_2+\ln(1+x_1^2+x_2^2)$$

9.2-) Considere o problema das espirais. Sendo a espiral 1 uma classe e a espiral 2 outra classe. Gere as curvas das espirais usando as seguintes equações:

para espiral 1 $x = \frac{\theta}{4}\cos\theta \qquad y = \frac{\theta}{4}\,\mathrm{sen}\,\theta \qquad \theta \geq 0$

para espiral 2 $x = (\frac{\theta}{4} + 0.8)\cos\theta \qquad y = (\frac{\theta}{4} + 0.8)\,\mathrm{sen}\,\theta \qquad \theta \geq 0$

Solucione este problema utilizando uma rede perceptron de múltiplas camadas. Gere a partir das equações os dados para treinamento e teste. Determine a matriz de confusão.

9.3-) Considere o problema de classificação de padrões bidimensionais constituído neste caso de 5 padrões. A distribuição dos padrões tem como base um quadrado centrado na origem interceptando os eixos nos pontos +1 e -1 de cada eixo. Os pontos +1 e -1 de cada eixo são centros de quatro semicírculos que se interceptam no interior do quadrado originando uma classe e as outras quatro classes nas regiões de não interseção. Após gerar aleatoriamente dados que venham formar estas distribuições de dados, selecione um conjunto de treinamento e um conjunto de validação. Treine uma rede perceptron para classificar os padrões associados a cada uma das classes. Verifique o desempenho do classificador usando o conjunto de validação e calculando a matriz de confusão.

10 - Trabalho:

Apresente um estudo com uma aplicação fazendo uso de uma rede percetron de múltiplas camadas. A aplicação pode ser na área de estudo que for de seu interesse.

A entrega e apresentação dos trabalhos correspondem a um processo de avaliação. Portanto a presença é obrigatória.

O trabalho e a lista podem ser feitos de forma individual ou em grupo de dois componentes. Na apresentação os componentes poderão ser submetidos a questionamentos sobre a solução da lista e o desenvolvimento dos trabalhos.