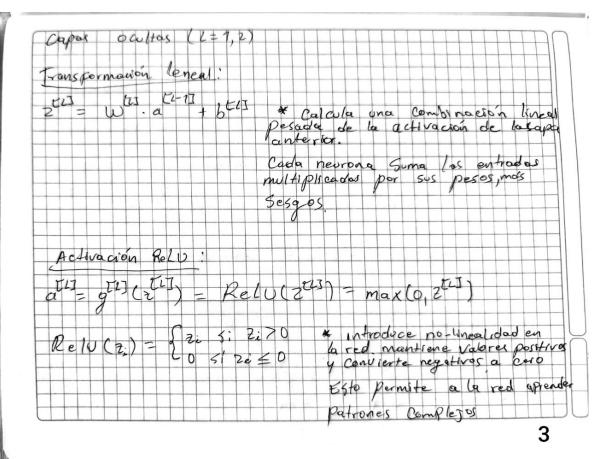
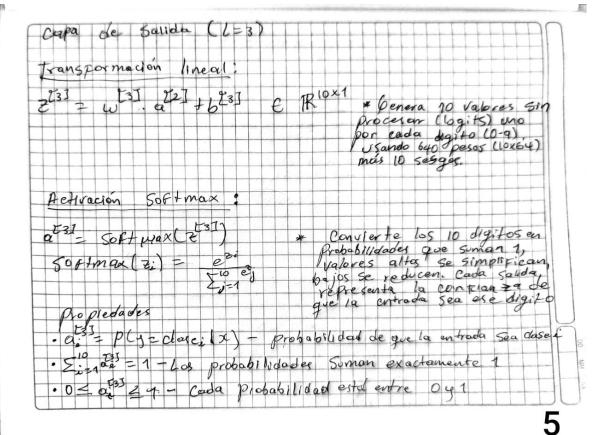
arquite coura de red neuronal elegida: Red neuronal multicapa (multilager perceptron - MLP) para Clasificación de digitos MNIST: Entrada: 784 neurones (28x28) pixeles explanados Capa Oculta 1: 120 neuronas Cactivación ReLU) Capa ocultaz: 64 neuronas (activación ReLU) capa de Salida: 10 novronas (activación Softmax) Notación matemática: Símbolo Descripción Valores en la red número total de capas 4=3 na1= 784, ntil= 128, nt=64, nt=10 n CLI numero de neuron os en capa l Dimensiones (nti ntil) matriz de pesos de capa 2

Simb	olo Desc	xipción	Valores en la red
2547	Vector o	de Sesgos de capa	U Dimensiones (51,1)
a <sup>[1]</sup>	Activac	iones de capa L	Salida de la capa
211	Pre-ad	Ivaciones de cap	oa U Antos de punción de activación
m	Tamaño	del batch	Numero de ejemplos
Ø	Tousa de	aprendizaje	ui per parametro
Foru	ord pas	5 (propagació	n hocia adelante)
Capa e	de entroda		
F0] = :			entrado, Como los activaciones de la ca
4.	Donde	x E Ries e	el vector de entrada (píxeles normaliz
	entre	b y 1)	





Prediction

g = a, Chare prediction

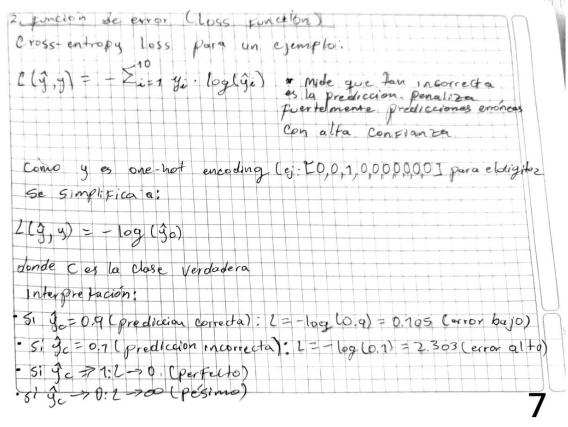
g = a, Chare prediction

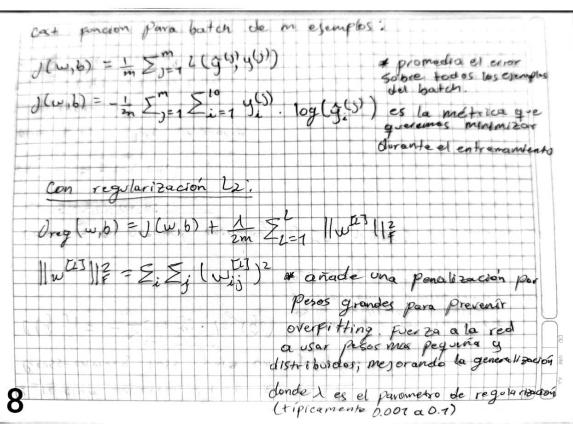
expense of probabilities

expense of the prediction

expense of the probabilities

expense of





3. back propagation ( propagation havia atras)

Noto don de gradientes:

Simbolo Significado
ac actual
actual
autil
dotil actual
and actual
capa de Salida (1=3):

Para Sortmax + Cross-entropy

dutas = d 2 cs (23) 1

dutas = d 2 cs (23) 1

dutas = d 2 cs (23) 1

U.U. ( UU. 1)

gradiente respecto a activaciones previos

datis = (wtitis) T dz titis

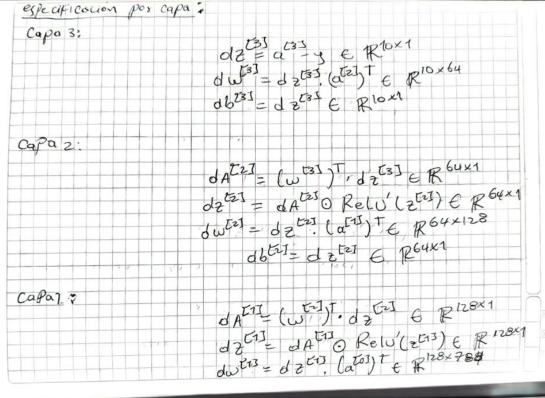
gradiente respecto a pre-activaciones (con derivado de Relo)

dztis = datis (ztis) = Relu'(ztis) = (1 sizis)

gradientes de parametros

dotis = dztis (atis) T

dotis = dztis (atis) T



para batch de m ejemplos

distil = 1 detis. (Atilija

distil = 1 \ (soma por pi/ae de de zel)

distil = 1 detis. (Atilija

distil = 1 detis. (Atilija

distil = 1 detis. (Atilija)

distil = 1 detis.

4. descenso por gradiente (gradient descent)

Ecuaciones de actualización:

Para codo capa l = 1,2,3

L[1]:=W - a.dw[1]

b[2]:=b[1] a.db[1]

dende a es la taxa de aprendizaje (tipicamente entre 0,001 y 0.7)

Variantes:

7. Batch gradient descent

Usa todo el dataset en lada Heración

2. Sto Chastic gradient olesicalant (560)

Actualiza con un ejemplo a la vez

3. Mini-batch gradien descent

Usa batches pequeños (32,64,128 ejemplos) maís Councir

momentum:

tis = B. vis + (1-B). dw de gradientes plesados.

Crea inercia que acelera el esprendita de gradientes y amos tigua oscilaciones.

Lonsistentes y amos tigua oscilaciones.

Velo := B. vis + (1-B). dotis 

Consistentes y amos tigua oscilaciones.

Velo := B. vis + (1-B). dotis 

Velocidad acumulada en lugar.

Printer momento (momentum): molu = R1 molu + (1-B1) du

promedio provil exponencial del gradiente

moly:= 184 moly + (1-B1) do

segundo momento (RMS prop): B2: Volu + (1-B2) dut

promedo movil exponencial del gradiente all cuadrodo movil
(magnitud). adapta la tasa de caprendizaje por parametro

Volo:= B2 Volo + (1-B2) blb2

Corrección de bias: moju:= moju; moly:= moly & corrige el sesso inicial
7-B1; houa caro en las princess

Volu:= Volo:= Volo:= Volo:= 1-B2

Outralización: W:= W-a - moly ole la varianza normaliza el paso:
Volute parametros con gradientes el paso:
reciben pares mos poquenes y vice versa
es como tener una Lasa de aprendizaje
adaptativa por parametro

b:= b-a multiple

Tipicamente: B1 = 09 (momentum), B2 = 0.999 (vomenza), E=10 Estabilidad numérica

combina lo mejor de momentum y RMS prop. es el optimizador

muls popular por que funciona bien "out ox The box" sin mucho trang

5. Explicación intuitiva:

El proceso de aprendizajo:

conalogía: la función de error j(W,b) es como un paísaje montañoso.

9 veremos encentrar el valle mas bajo (mínimo

El gradiente: Vo = (aw, db) indica:

• Dirección: hacia donde sobe más rápido

• Magnitud: que fan emprada es la pendiente

gradient descent: Nos movemos Cuesta abajo (dirección opuesta al gradiente)

Ciclo de entrenamiento:

1 Prodecir: con posos achales (Forward Pass)

2 Medir: el error (Loss Function)

3 Calcular: Como Cambiar pesas (Backpropagation)

4 Ajustar: pesas en ethrección correcta (gradient descent)

5. Repetir: hosta convergencia

Interpretación de gradientes:

• Si dw<sup>[2]</sup> > 0: aumentar Wij incrementa error > dismovir pesa

• Si dw; [1] co aumentar Wij dismovye error > aumentar pesa

• Si dw; [1] cs grande: Ese pero tiene gran impacto

• Si ldw; [1] cs pequeño: ese peso tiene poco impacto

Valor	epecto	Consecuencia
Grande (0.7)	pasos grandes	Rápido paro Puede Oscilar
Pequeño (0.001)	Pasos pequeños	lento pero estable
optimo	Balance	Convergencia eficiente
Moniforeo del e	intrenamento:	
	un entrenamiento:	
	ye consistente	
· accuracy		
· Convergencia		
Problemas Com	wnes:	
· loss oscila-	or muy grande	2
		wente o mínimo local
		1 -> overfitting

Ejemplo numerico simplificado: para un peso u simple con a = 0.1% u g érror gradiente nuevo u 0.54 0.54 0.85 0.023 -0.30 0.57 0.57 0.91 0.008 -0.18 0.59 el error disminuje y w coverge gradualmente al valor optimo Resumen de domensjones TLT 2643 btes w 123 capa 128x1 12821 128 X 784 128×1 64x1 64×1 64 x 128 64×1 10 x 1 10×1 2027 10 ×64

