Trabajo Tema 4 MED (MUM)

Beatriz Coronado Sanz

Curso 2018-2019

Índice

Conjunto de datos

- Cada fila de los conjuntos de datos contiene un número: la variable 785 identifica al dígito y las variables 1 hasta 784 se corresponden con los 28×28 píxeles que forman la imagen del número.
- Transformamos las variables 1 a 784 al intervalo [0,1] dividiendo cada una de ellas por 255

Selección de la muestra de entrenamiento y test

- Elegimos aleatoriamente 5 dígitos entre 0 y 9 y restringimos nuestros conjuntos de entrenamiento y test a esos números.
- Extraemos aleatoriamente 3000 casos del nuevo conjunto de entrenamiento y 1000 del nuevo conjunto de test.
- Los dígitos para nuestro problema serán 2, 4, 6, 8 y 9.

Representación del conjunto de entrenamiento

- En nuestra muestra tenemos el siguiente número de casos para cada dígito:

Dígito	2	4	6	8	9
Número de casos	597	582	614	602	605

- Observamos que tenemos una muestra balanceada donde el número de casos por dígito es parecido.

Representación del conjunto de entrenamiento

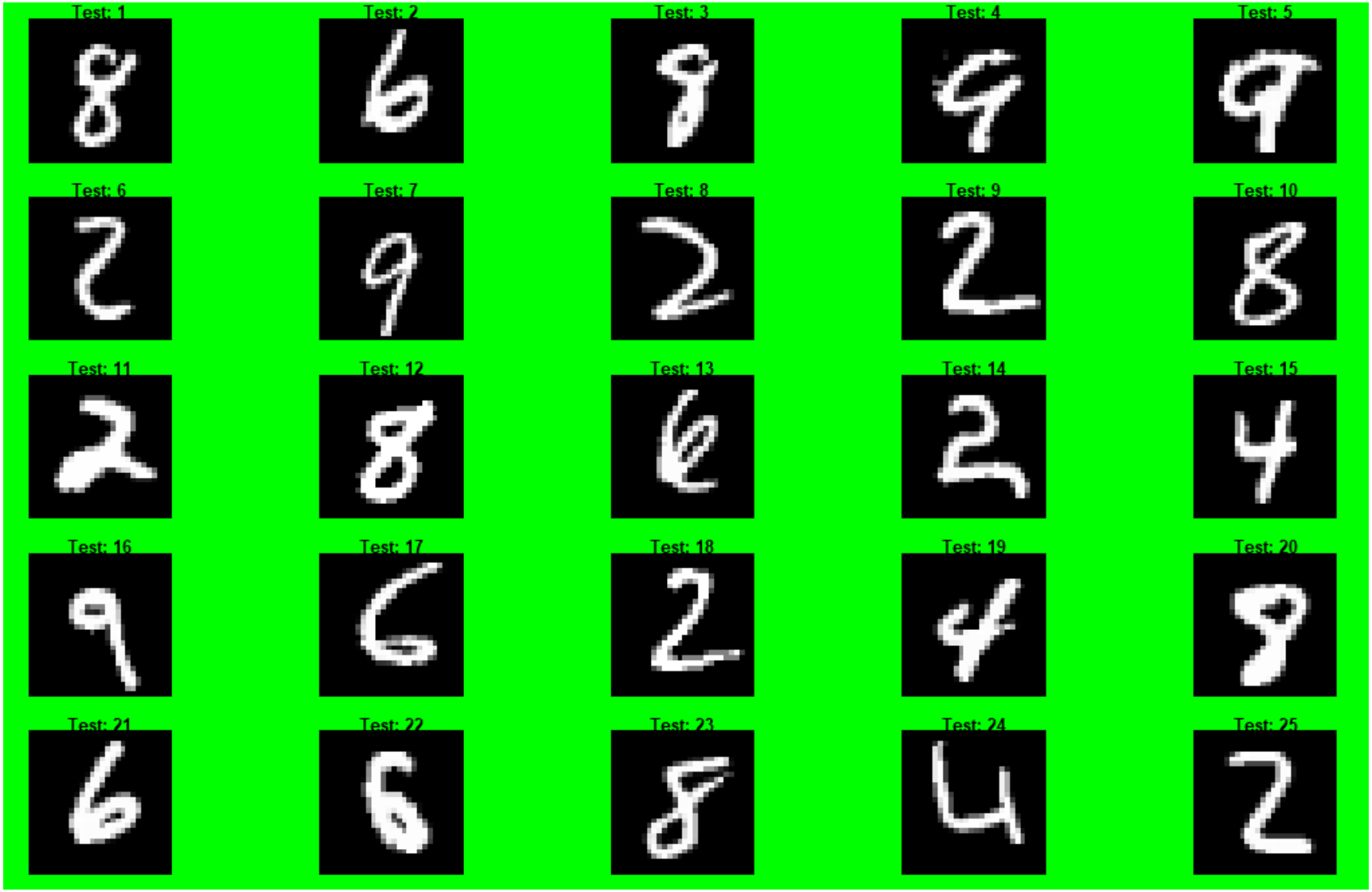
Representación del conjunto de test

- En nuestra muestra tenemos el siguiente número de casos para cada dígito:

Dígito	2	4	6	8	9
Número de casos	212	205	209	181	193

- Observamos que tenemos una muestra balanceada donde el número de casos por dígito es parecido.

Representación del conjunto de test



Eliminación de variables constantes

- El último preprocesado que hacemos a los datos consiste en eliminar aquellos píxeles que no varían para todos los elementos de nuestra muestra de entrenamiento.
- Eliminamos esas variables (píxeles) tanto en el conjunto de entrenamiento como en el de test.
- Esto nos permite pasar de 785 variables a 624.

Red neuronal sobre PCA

- Debido a que tenemos muchas variables para predecir el dígito de cada imagen (623), no podemos generar una red neuronal con todas ellas.
- Para solucionar esto aplicamos PCA a nuestro conjunto de entrenamiento y nos quedamos con los píxeles más representativos de cada imagen.
- De esta forma pasamos de tener 623 variables a tener 60, lo que representa un $84,75\,\%$ de la varianza total de nuestros datos.

Resumen PCA

- En la siguiente imagen podemos observar un resumen de las primeras componentes principales que hemos calculado:

	Eigenvalue	Percentage	Cumulative percent.
[1,]	4.28	8.40	8.40
[2,]	4.10	8.05	16.45
[3,]	3.07	6.03	22.48
[4,]	2.92	5.74	28.22
[5,]	2.40	4.72	32.94
[6,]	1.82	3.57	36.51
[7,]	1.55	3.04	39.55
[8,]	1.48	2.91	42.46
[9,]	1.29	2.54	44.99
[10,]	1.17	2.29	47.29
[11,]	1.14	2.24	49.53
[12,]	0.96	1.89	51.43
[13,]	0.85	1.67	53.10

Conversión de los dígitos

- Lo último que hacemos antes de construir nuestra red neuronal es transformar cada uno de los dígitos que queremos averiguar a un formato vectorial de 5 elementos.
- De esta forma, una imagen será de la clase i si el i-ésimo elemento del vector vale 1 y el resto 0.
- Nuestro dígitos quedan de la forma:

```
     1 2 3 4 5
[1,] 0 0 0 1 0
[2,] 0 0 1 0 0
[3,] 0 0 0 0 1
[4,] 0 0 1 0 0
[5,] 0 0 0 0 1
[6,] 0 0 0 0 1
```

Resultados en el conjunto de entrenamiento

- Para una red neuronal de 15 nodos ocultos y parámetro de regularización igual a 0 tenemos los siguientes resultados para el conjunto de entrenamiento:

	2	4	6	8	9
2	566	11	2	13	5
4	11	539	4	4	24
6	3	2	604	5	0
8	12	1	3	581	5
9	7	19	1	8	570

- El porcentaje de acierto de la red neuronal es del $95,33\,\%$.

Resultados en el conjunto de test

- Para una red neuronal de 15 nodos ocultos y parámetro de regularización igual a 0 tenemos los siguientes resultados para el conjunto de test:

	2	4	6	8	9
2	192	3	8	6	3
4	1	194	1	2	7
6	5	2	198	2	2
8	8	1	3	166	3
9	3	8	0	8	174

- El porcentaje de acierto de la red neuronal es del $92,4\,\%$.

Interpretación de los resultados

- Observamos que los resultados son muy buenos en ambos conjuntos pues el acierto está por encima del 90 %.
- La red neuronal es un poco mejor en el conjunto de entrenamiento que en el de test. Lo que es normal debido a que el conjunto de entrenamiento es con el que se ha entrenado la red.
- Observamos así que nuestra red neuronal tiene una gran capacidad de generalización al acertar casi siempre los nuevos dígitos que le llegan.
- Vemos que hay algunas confusiones entre el 4 y el 9 (en total 43 fallos en el conjunto de entrenamiento) y entre el 2 y el 8 (en total 25 fallos).

Red mediante aprendizaje profundo

- Vamos a generar una red de aprendizaje profundo con 2 capas ocultas de 200 neuronas cada una.
- Tendremos un porcentaje de dropout del 20 % y entrenaremos la red como mucho 50 veces.
- Para generar esta red usaremos la herramienta h2o.
- En este caso podemos evaluar nuestra muestra de entrenamiento con todas las variables (las que no son constantes), sin necesidad de aplicar PCA.

Resultados para el conjunto de entrenamiento

- Para el conjunto de entrenamiento obtenemos los siguientes resultados:

	2	4	6	8	9	Error	Rate
2	592	1	1	3	0	0.0084 =	5 / 597
4	0	575	1	0	6	0.0120 =	7 / 582
6	0	0	614	0	0	0.0000 =	0 / 614
8	0	0	1	600	1	0.0033 =	2 / 602
9	0	0	0	1	604	0.0017 =	1 / 605
Totals	592	576	617	604	611	0.0050 =	15 / 3.000

- El porcentaje de acierto de la red neuronal es del $99,5\,\%$.

Resultados para el conjunto de test

- Para el conjunto de test obtenemos los siguientes resultados:

	2	4	6	8	9	Error	Rate
2	197	3	4	7	1	0.0708 =	15 / 212
4	1	198	2	1	3	0.0341 =	7 / 205
6	3	3	200	3	0	0.0431 =	9 / 209
8	0	1	1	175	4	0.0331 =	6 / 181
9	1	2	0	5	185	0.0415 =	8 / 193
Totals	202	207	207	191	193	0.0450 =	45 / 1.000

- El porcentaje de acierto de la red neuronal es del $95,5\,\%$.

Interpretación de los resultados

- En el conjunto de entrenamiento practicamente se aciertan todos los casos y en el de test el acierto está por encima del $95,5\,\%$
- Observamos que en el conjunto de entrenamiento siempre se predice correctamente el número 6.

Conclusiones finales

- Comprobamos que la eficacia de una red de aprendizaje profundo es superior a la de una red neuronal.
- Podemos plantearnos si generar una red de aprendizaje profundo compensa para este problema ya que los resultados obtenidos de la red neuronal calculada son muy buenos.