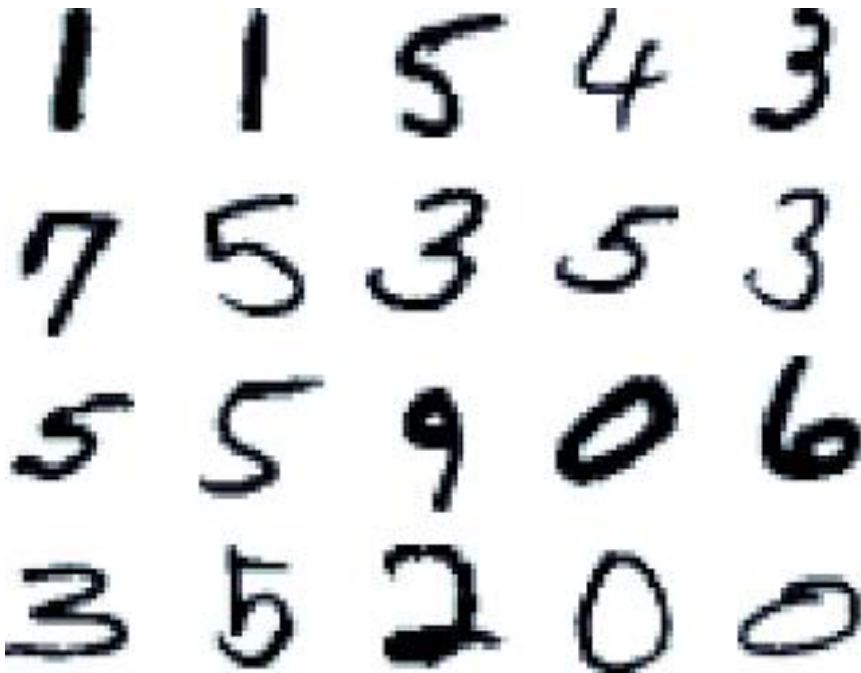


Digit Recognizer

Kaggle

Autor: Andrés Herrera Poyatos



GitHub



Repositorio en GitHub con el código:
<https://github.com/andreshp/Kaggle>

Digit Recognizer

Kaggle

- ¿Qué es Kaggle?
- Digit Recognizer
- Resolviendo el problema
 1. ¿Qué hacer en primer lugar?
 2. Visualización
 3. Preprocesamiento
 4. Deep Learning

Digit Recognizer

Kaggle

- ¿Qué es Kaggle?
- Digit Recognizer
- Resolviendo el problema
 1. ¿Qué hacer en primer lugar?
 2. Visualización
 3. Preprocesamiento
 4. Deep Learning

¿Qué es Kaggle?

- **Kaggle** es una plataforma web que mantiene competiciones de análisis de datos.
- Reconocidas empresas patrocinan competiciones con premios en metálico.
- ¡Participan los mejores científicos de datos del mundo!

kaggle.com

Digit Recognizer


Kaggle

- ¿Qué es Kaggle?
- **Digit Recognizer**
- Resolviendo el problema
 1. ¿Qué hacer en primer lugar?
 2. Visualización
 3. Preprocesamiento
 4. Deep Learning

Digit Recognizer

- Desarrollar un reconocedor de dígitos es uno de los **problemas clásicos** de la ciencia de datos.
- Sirve de **benchmark** para probar los nuevos algoritmos. ¡Ni un humano acierta el 100%!
- **Aplicación práctica:** detección de matrículas, conversión de escritura a mano en texto ...


9	6	6	5	4	0	7	4	0	1
3	1	3	4	7	2	7	1	2	1
1	7	4	2	3	5	1	2	4	4



9	6	6	6	4	0	7	4	0	1
3	1	3	4	7	2	7	1	2	1
1	7	4	2	3	5	1	2	4	4

Digit Recognizer

- **Kaggle** mantiene una competición pública:

	Digit Recognizer Classify handwritten digits using the famous MNIST data	9 months 433 teams Knowledge
--	--	------------------------------------

<http://www.kaggle.com/c/digit-recognizer>

- Datos a analizar: MNIST DATA

<http://yann.lecun.com/exdb/mnist/>

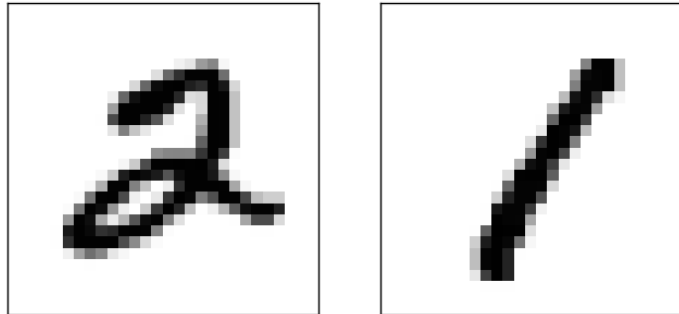
Digit Recognizer

- Data Set:

- Training Set: 42.000 Imágenes
- Test Set: 28.000 Imágenes

- Imagen:

- 10 clases: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
- 28x28 píxeles
- Ejemplo:



- Puntuación para la clasificación general: índice de acierto sobre un 25% del Test Set.

Digit Recognizer

Kaggle

- ¿Qué es Kaggle?
- Digit Recognizer
- Resolviendo el problema
 1. ¿Qué hacer en primer lugar?
 2. Visualización
 3. Preprocesamiento
 4. Deep Learning

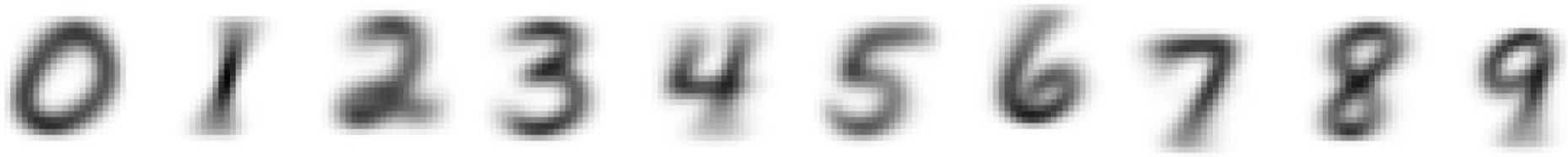
1. ¿Qué hacer en primer lugar?

■ Respuesta:

1. Probar los algoritmos más conocidos para usarlos como benchmark
 - KNN con $k = 10$ \longrightarrow 0.96557 en Kaggle
 - Random Forest con 1000 árboles \longrightarrow 0.96829 en Kaggle
2. Optimizar los parámetros de un algoritmo sencillo
 - Cross Validation sobre KNN para encontrar el mejor valor de k .
Solución: $K=1$ \longrightarrow 0.97114 en Kaggle

2. Visualización

- Media de todas las imágenes del training set por clases:

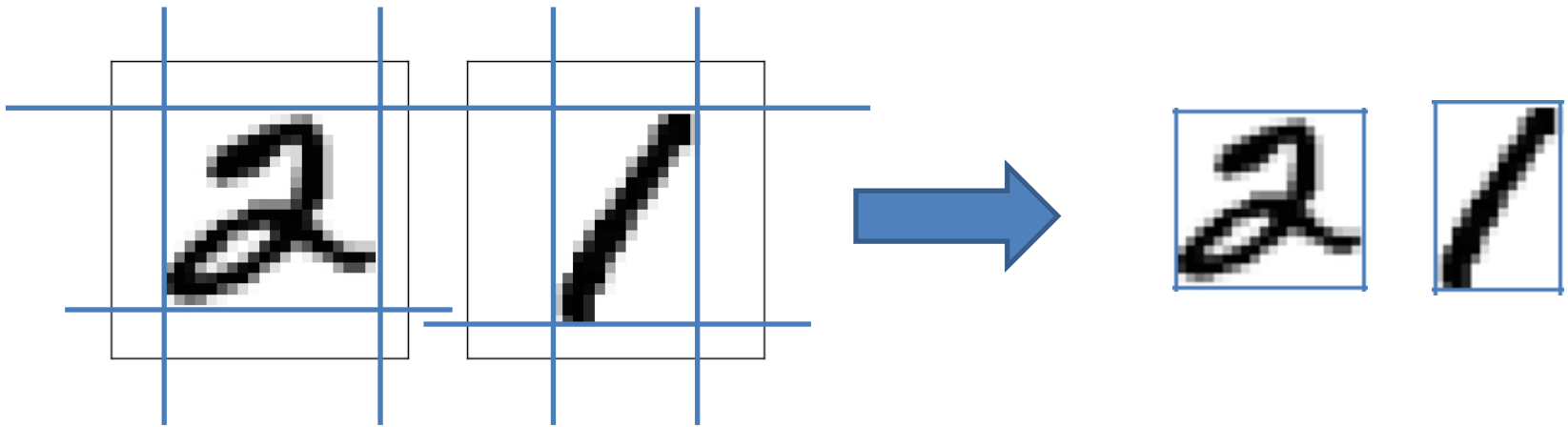


- **Observación:** Incluso las medias no están centradas (ver 6 y 7). Esto provoca problemas para clasificarlas correctamente.

Solución: Preprocesamiento

3. Preprocesamiento

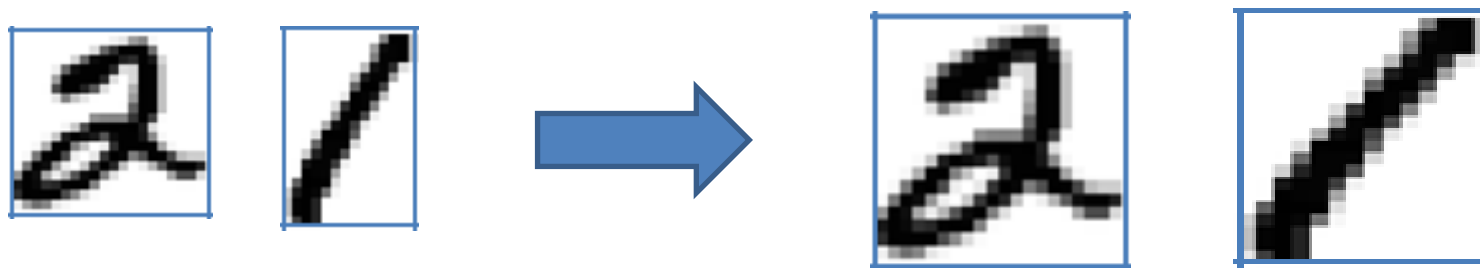
- **Idea:** Eliminar las filas y columnas de píxeles en blanco.



- **Problema:** Las nuevas imágenes tienen diferentes dimensiones.

3. Preprocesamiento

- **Solución: Redimensionar** las imágenes a 20x20 píxeles (tras el proceso anterior la imagen más grande tiene esa dimensión)



- Media de las imágenes del training set preprocesadas:



- ¡Todas están centradas!
- KNN con $k=1$ sobre los datos preprocesados → 0.97557 en Kaggle

4. Deep Learning

- Deep Learning es un algoritmo de aprendizaje basado en redes neuronales que proporciona muy buenos resultados en el área de **Pattern Recognition** (Reconocimiento de patrones).
- Buena librería de Deep Learning: **h2o** <http://0xdata.com/>
 - Soporte para R, Hadoop y Spark
 - Récord del mundo en el problema MNIST sin preprocesamiento

<http://0xdata.com/blog/2015/02/deep-learning-performance/>

- Funcionamiento: Crea una máquina virtual con Java en la que optimiza el paralelismo de los algoritmos.

H₂O



Spark + **H₂O**

SPARKLING
WATER

4. Deep Learning

- Tiempo de Ejecución: 2.5 horas de cómputo con un Procesador Intel i5 a 2.5 GHz.
- Resultados conseguidos:
 - Deep Learning → 0.98229 en Kaggle
 - Preprocesamiento + Deep Learning → 0.98729 en Kaggle
- ¡Recordad que nuestro primer resultado era 0.96557!