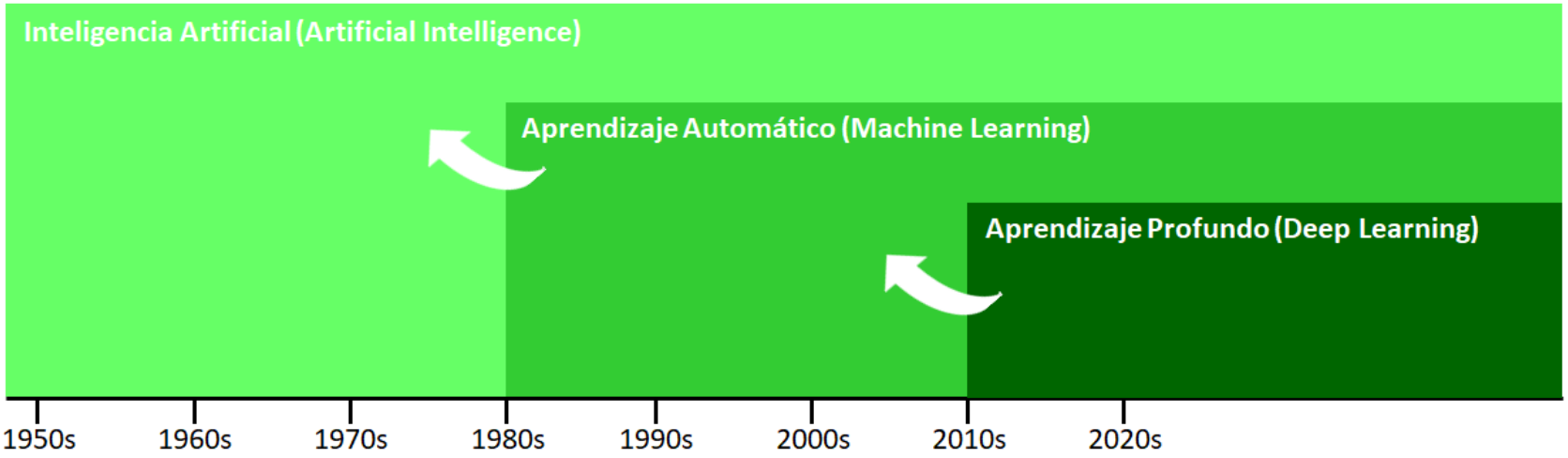
Introducción al curso

Aprendizaje automático 1

Juan R Gonzalez
juanr.gonzalez@isglobal.org

UAB - Department of Mathematics
BRGE - Bioinformatics Research Group in Epidemiology
ISGlobal - Barcelona Institute for Global Health
http://brge.isglobal.org

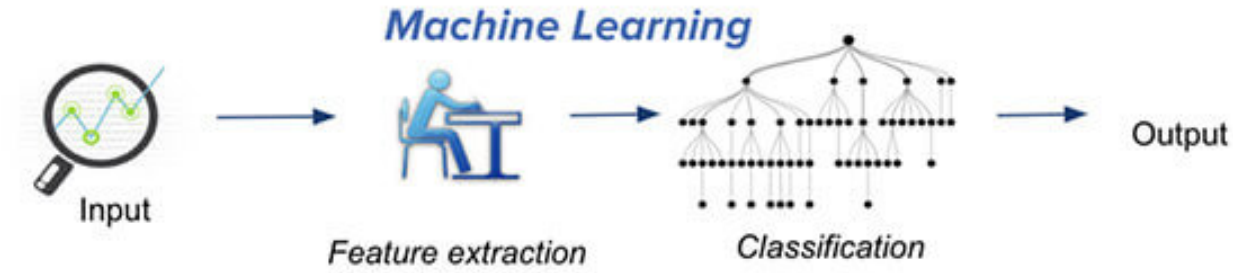
Introducción



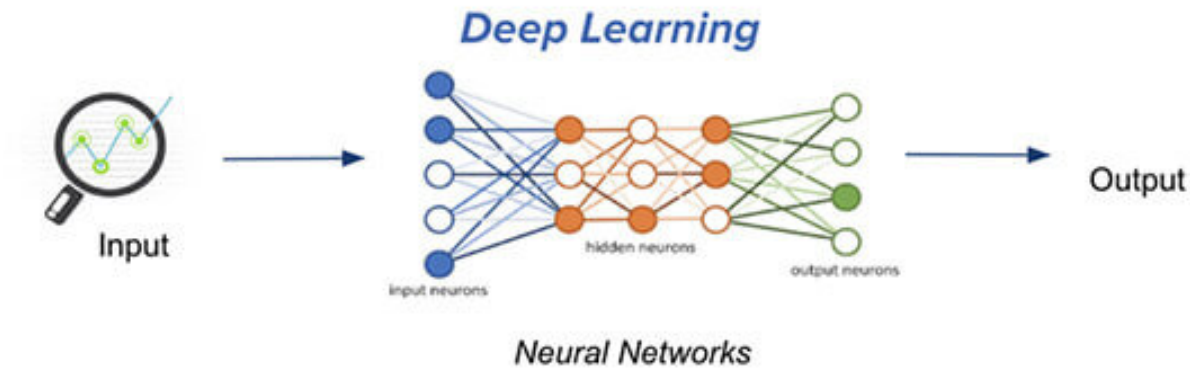
Sobre 2010 DL obtuvo una gran popularidad ya que ha permitido acercarse a sitemas de inteligencia artificial de forma más eficientge que ML. Los tres términos están ligados y cada uno forma una parte esencial de los otros. DL permite llevar a cabo ML, que en última instancia permite la AI.

No obstante, es más fácil aprender ML como herramienta para AI. En el siguente curso (AA_2) veréis cómo llevar a cabo técnicas de DL.

Introducción



Traditional machine learning uses hand-crafted features, which is tedious and costly to develop.

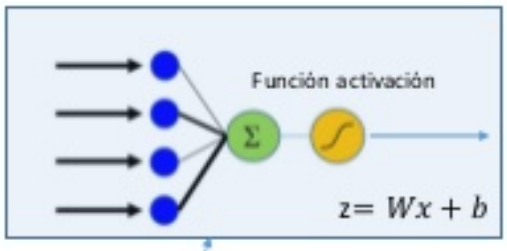


Deep learning learns hierarchical representation from the data itself, and scales with more data.

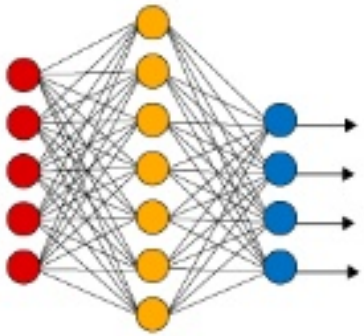
Ejemplos


Deep Learning

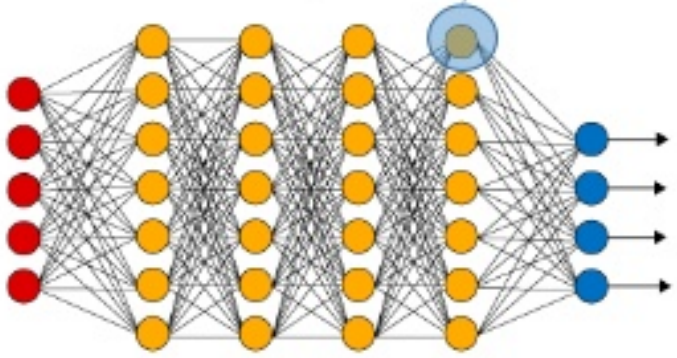
Redes Neuronales



Red neuronal simple

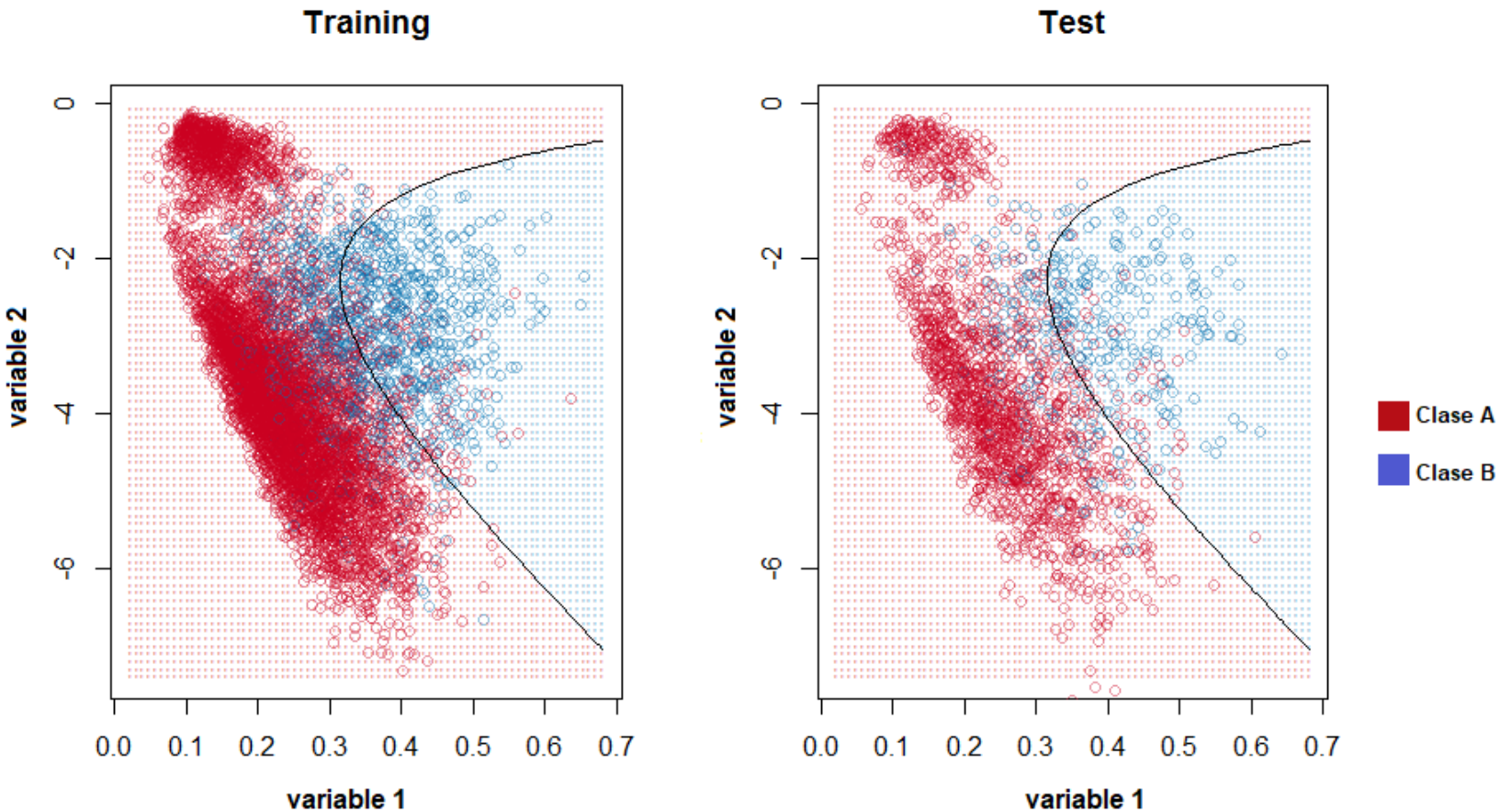


Red neuronal profunda



Extracción de características y clasificación

Ejemplos


Training
Test
variable 2
variable 1
0
-2
-4
-6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Clase A
Clase B

Temario

- Introducción a Tidyverse
- Introducción al aprendizaje automático
- Regresión lineal y logística (nomogramas)
- Dealing with Big Data in R (MapReduce)
- La librería `caret`
- Pasos previos a la creación de un modelo predictivo y medidas de validación
- Métodos de aprendizaje automático
  - K-vecinos más cercanos (KNN)
  - Análisis discriminante lineal (LDA)
  - Máquinas de soporte vectorial (SVM)
  - Árboles de decisión (clasificación, regresión, bagged trees, random forest)
  - Boosting (AdaBoost, GBM básico y estocástico, XGBoost)
- Respuesta no balanceada

Logística del curso

- Clases lunes de 15:00 a 17:00 (teoría) + 17:00 a 19:00 (prácticas)
- Clases presenciales aula C3/028
- Clases no presenciales (bookdown + algunas grabadas). Tendrá que hacerse las preguntas de autoevaluación el mismo día antes de las 21:00 horas (no negociable)
- Cada semana habrá una práctica similar a lo que se ha explicado en clase, pero con otros datos reales. También habrán "Data analysis Challenges"
- Cada semana se pondrá una noticia con lo que se va a hacer cada lunes.

Material del curso

- No seguiremos ningún libro de texto porque estamos tratando muchos temas que están muy bien explicados en varios libros y, sobre todo, en materiales públicos. He creado un bookdown accesible en https://isglobal-brge.github.io/Aprendizaje_Automatico_1/ que describe el contenido del curso
- Habrán presentaciones en diapositivas que estarán disponibles en el Moodle de la asignatura
- Habrán link a material opcional
- Las actividades en el Moodle estarán clasificadas como:
  - P: puntuable (se evaluará según el contenido)
  - P2: puntuable (no se evaluará si está bien - se dará la solución: 10 entregado, 5 entregado no completo, 0 no entregado)
  - I: información
  - O: opcional

Metodología

- Clase de teoria: se definen y se explican los diferentes métodos con sus características particulares y se muestran ejemplos concretos.
- Clase de prácticas: se trabajan los métodos explicados en clase de teoria con diversos conjuntos de datos utilizando el lenguaje de programación R.

Se considera que, para cada hora de teoria y prácticas, el alumno deberá dedicar una hora adicional para la preparación y/o finalización de la sesión.

- Preguntas de auto-evaluación que se llevarán a cabo en el Moodle que servirá para consolidar los conocimientos aprendidos en las sesiones teóricas y que en las sesiones no presenciales se tendrá que hacer el mismo día antes de las 21:00.

Evaluación

La nota final se basará en las siguientes notas ponderadas

- 50% Nota Examen final (tipo test - conceptos)
- 40% Nota Prácticas (compuesta por prácticas semanales más una práctica final)
- 10% Nota preguntas de Auto-Evaluación

Se requerirá tener un 5 en el Examen final para aprobar la asignatura, en caso contrario el alumno deberá presentarse al examen de recuperación. Existe la posibilidad de aprobar el examen con un 4.5 si el alumno ha participado de forma activa en el foro de la asignatura.

Fechas de Evaluación

- Examen parcial (no hay)
- Examen final (fecha por saber)
- Recuperación Examen 70% (fecha por saber) + Notas de prácticas 30%

Cronograma

Fecha	Teoria	Tipo	Tareas
13-sep	Intro + Tidyverse (manejo de datos)	Presencial	P - Autoevaluación tidyverse; P2 - Entrega ejercicios (data transform)
20-sep	Tidiverse (visualización) + Regresión lineal	No presencial	P - Autoevaluación Regresión Lineal; P2 - Entrega ejercicios (visualización); P2-Regresión lineal función; P - Práctica Regresión Lineal
27-sep	Creación de modelos + CV + Missing imputation	Presencial	P- Autoevaluación creación de modelos; P- Práctica creación de modelos; P2 - LOOCV; P2 - Kold; P2 - bootstrap
04-oct	Regresión logística + Nomogramas	Presencial	P- Autoevaluación regresión logistica; P - Práctica regresión logística; P2 - Precisión
12-oct	Paralelización + Regresión con Big Data	Presencial	P-Autoevaluación intro Big Data; P2-MapReduce
18-oct	Librería caret + KNN	Presencial	P-Autoevaluación AUC y método; P-Preproceso cáncer de cervix; P2-KNN; P2-Cáncer de Cervix con KNN; Concurso FitBit
25-oct	LDA + Máquinas de soporte vectorial	Presencial	P2-LDA
01-nov	No hay clase		
08-nov	No hay clase		
15-nov	Datos no balanceados	Presencial	P2-Cáncer de Cervix con LDA, KNN y SVM
22-nov	Árboles (I): Clasificación, regresión, bagged	Presencial	P2-Bagged; P-Bagged Breast Cancer; P-Autoevaluación CART y otros
29-nov	Árboles (II): Random forest	Presencial	P2-Plot RF; P2-RF Breast Cancer; Concurso EDY
06-dic	No hay clase		
13-dic	Árboles (III): Boosting, GBM	Presencial	P2-Plot RF; P2-RF Breast Cancer; Concurso EDY
20-dic	Árboles (IV): XGBoost	Presencial (2h)	P2-AdaBoost; Concurso EDY
27-oct	No hay clase		
03-ene	No hay clase		
10-ene	Repaso - Margen por si retraso	Presencial	Tipo test - Similar a preguntas de autoevaluación

Session info

```
sessionInfo()
```

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 18362)

Matrix products: default

locale:
[1] LC_COLLATE=Spanish_Spain.1252  LC_CTYPE=Spanish_Spain.1252
[3] LC_MONETARY=Spanish_Spain.1252 LC_NUMERIC=C
[5] LC_TIME=Spanish_Spain.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
 [1] compiler_4.0.2  magrittr_1.5    tools_4.0.2     htmltools_0.5.0
 [5] yaml_2.2.1      stringi_1.4.6   rmarkdown_2.3   knitr_1.29
 [9] stringr_1.4.0   xfun_0.16       digest_0.6.25   rlang_0.4.7
[13] evaluate_0.14
```