Taller de Aprendizaje Automático

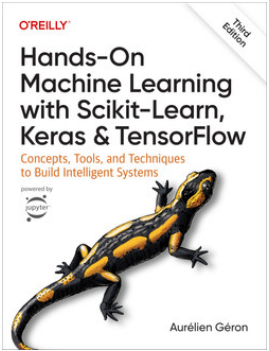
Introducción y conceptos fundamentales

Instituto de Ingeniería Eléctrica
Facultad de Ingeniería

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Descripción del curso

- herramientas conceptuales y metodológicas para desarrollar proyectos de aprendizaje automático
- modalidad taller, trabajo individual y en equipos
- técnicas de aprendizaje automático:
  - clásicas (p. ej. SVM, k-NN)
  - ensambles (p. ej. bagging, boosting)
  - aprendizaje profundo (p. ej. CNN, RNN)
- principalmente basado en un libro*
- 10 créditos
- 2 entregables, 2 proyectos
- previa *Fundamentos de Aprendizaje Automático*



*A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition*. O'Reilly Media, Inc., 2022

- se asume que el estudiante está bien familiarizado con:

- se introducen varias herramientas nuevas, como por ejemplo:

Actividades

Hacer cuestionario en la página del curso en **EVA**.

Crear usuario en **Kaggle** (`https://www.kaggle.com`).
Crear usuario en **Comet** (`https://www.comet.ml`).
Seguir el tutorial de **Pandas** en **Kaggle** (`https://www.kaggle.com/learn/pandas`).

Ir pensando en conformar grupos de tres estudiantes (se habilitará la elección de grupos la semana que viene en la página del curso en **EVA**).

Aprendizaje automático

Aprendizaje Automático

El aprendizaje automático es la ciencia (y el arte) de programar computadoras para que puedan aprender de los datos.

—Yaser Abu-Mostafa *

El aprendizaje automático es el campo de estudio que brinda a las computadoras la capacidad de aprender sin ser programadas explícitamente.

—Arthur Samuel, 1959

Una computadora aprende de la experiencia E con respecto a alguna tarea T y alguna medida de desempeño P, si su desempeño en T, medido por P, mejora con la experiencia E.

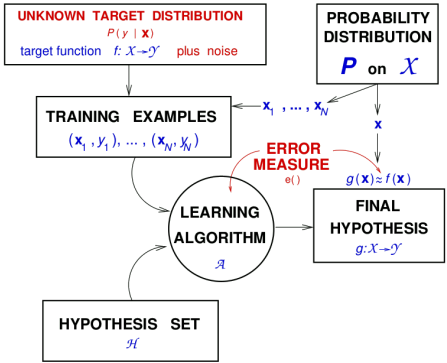
—Tom Mitchell, 1997

*Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AMLBook, 2012

El problema de aprendizaje

Componentes del problema de aprendizaje*

- función objetivo f
- N ejemplos de entrenamiento
- conjunto de hipótesis $\mathcal{H}$
- algoritmo de aprendizaje $\mathcal{A}$
- medida de error e
- hipótesis final $g \in \mathcal{H}$



*Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AMLBook, 2012

Aproximación versus generalización

Aproximar f fuera de muestra (E_{out})

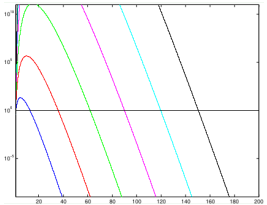
1. lograr un E_{in} suficientemente chico
2. asegurar que E_{out} y E_{in} sean cercanos

$\mathcal{H}$ más complejo $\Rightarrow$ mejora aproximación

$\mathcal{H}$ menos complejo $\Rightarrow$ mejora generalización

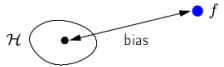
- Análisis *Vapnik-Chervonenkis*

$$E_{\text{out}} \leq E_{\text{in}} + \Omega(N, \mathcal{H}, \delta)$$

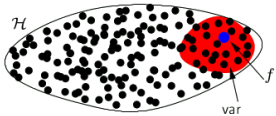


$$N \propto d_{\text{VC}}$$

- Análisis sesgo-varianza (*bias-variance*)

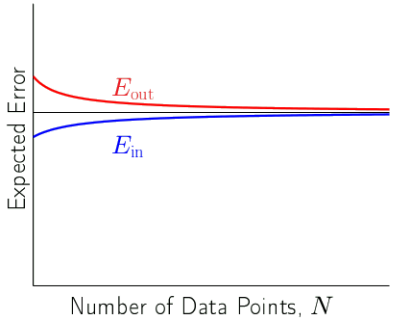


$\downarrow$

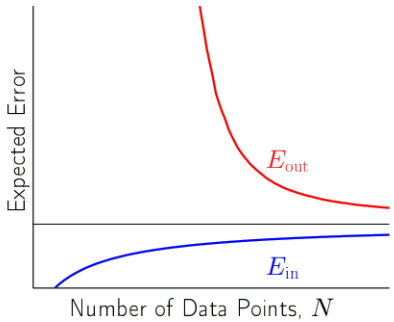


$\mathcal{H} \uparrow$ $\uparrow$

Aproximación versus generalización

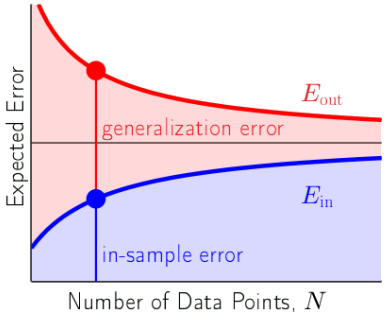


Simple Model

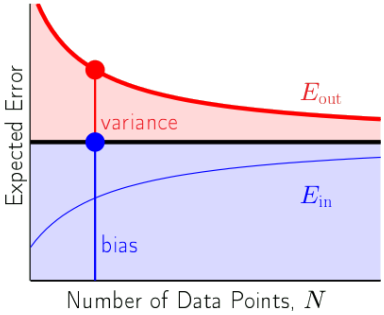


Complex Model

Aproximación versus generalización



VC analysis



bias-variance

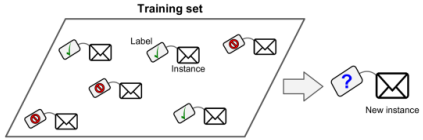
Tipos de aprendizaje

- aprendizaje supervisado
  - regresión lineal
  - regresión logística
  - redes neuronales
  - árboles de decisión
  - vecinos más cercanos (k-NN)
  - máquinas de vectores de soporte (SVM)

según el tipo de predicción

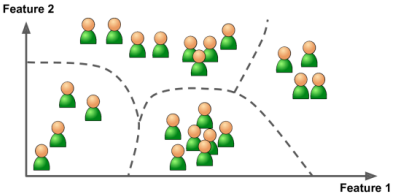
clasificación valor categórico de y

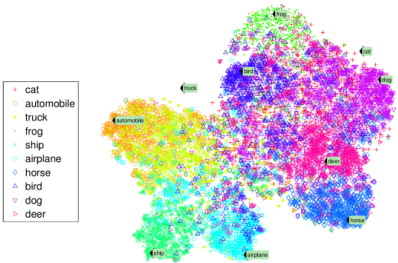
regresión valor numérico de y



Tipos de aprendizaje

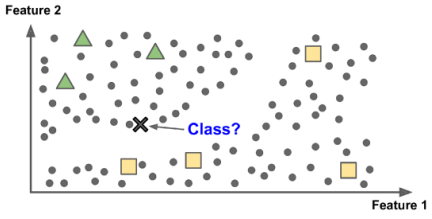
- aprendizaje no supervisado
  - K-means (agrupamiento)
  - PCA (reducción de dimensionalidad)
  - one–class SVM (detección de anomalías)



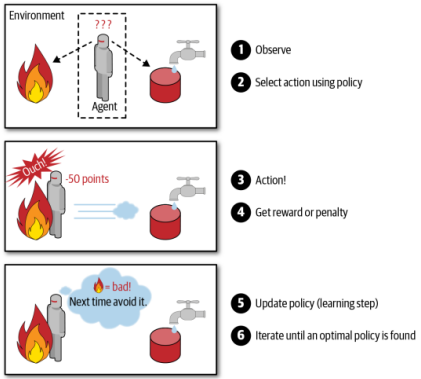


Tipos de aprendizaje

- aprendizaje semi supervisado
  - deep belief networks (DBNs)

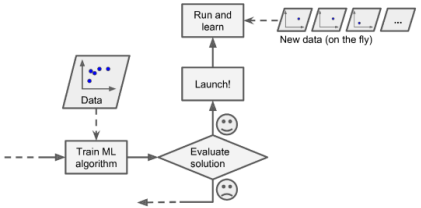


- aprendizaje por refuerzos

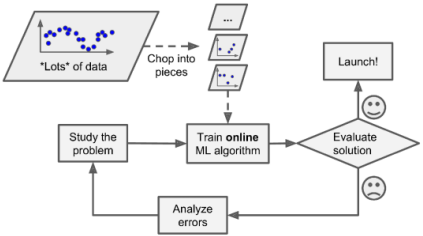


Tipos de aprendizaje

- aprendizaje por lotes (*batch*)
  - no puede aprender de forma incremental
  - actualizarlo implica entrenarlo de nuevo



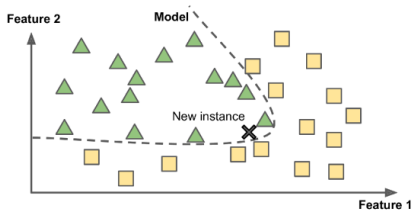
- aprendizaje incremental (*online*)
  - instancias individuales
  - pequeños grupos (*mini-batch*)



Tipos de aprendizaje

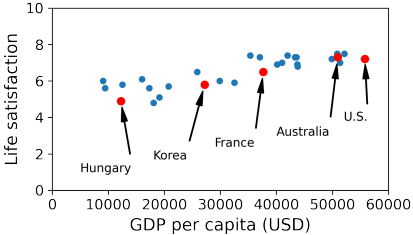
- aprendizaje basado en instancias
  - datos de entrenamiento
  - y medida de similitud

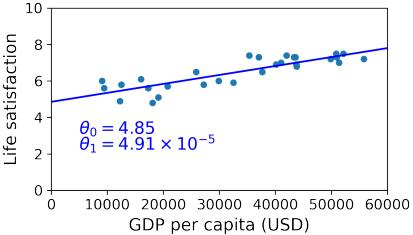
- aprendizaje basado en modelo
  - derivar modelo de los datos
  - usarlo para hacer predicciones

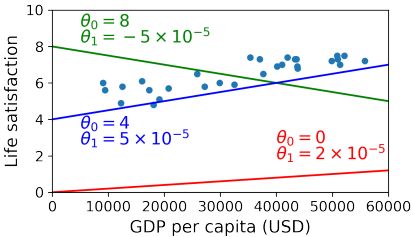


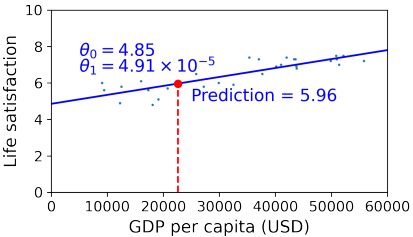
¿El dinero hace la felicidad?

$$\texttt{life\_satisfaction} = \theta_0 + \theta_1 * \texttt{GDP\_per\_capita}$$









Código python para regresión lineal

```
import numpy as np
import pandas as pd
import sklearn.linear_model

# load the data
oecd_bli = pd.read_csv(datapath + "oecd_bli_2015.csv", thousands=',')
gdp_per_capita = pd.read_csv(datapath + "gdp_per_capita.csv",thousands=',',delimiter='\t',
                             encoding='latin1', na_values="n/a")

# prepare the data
country_stats = prepare_country_stats(oecd_bli, gdp_per_capita)
X = np.c_[country_stats["GDP per capita"]]
y = np.c_[country_stats["Life satisfaction"]]

# select a linear model
model = sklearn.linear_model.LinearRegression()

# train the model
model.fit(X, y)

# make a prediction for Cyprus
X_new = [[22587]]  # Cyprus' GDP per capita
print(model.predict(X_new)) # outputs [[ 5.96242338]]
```

Actividades

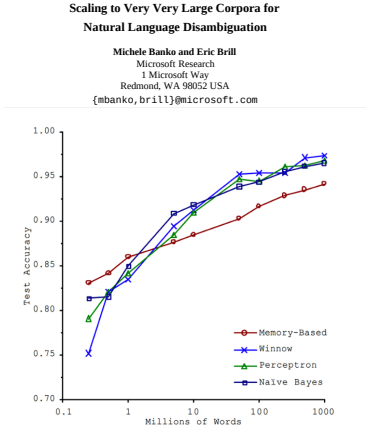
Correr el notebook del Capítulo 1 del libro.

- Utilizar un modelo de vecinos más cercanos.
- Probar modelo entrenado con datos de test.
- Utilizar modelo de regresión polinómica.

Tamaño del conjunto de entrenamiento

- cantidad de datos N
  - de miles a millones de ejemplos
  - mejores algoritmos versus más datos*†
  - aún existen problemas con pocos datos

- *la complejidad del modelo debe estar dada por la cantidad de datos disponibles y no por la complejidad del problema* ‡



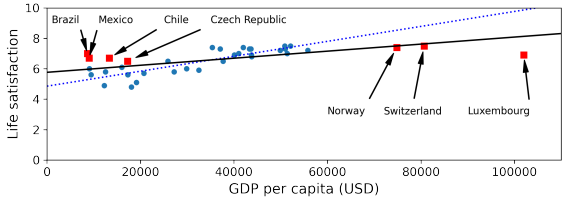
*M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proceedings of the 39th ACL 2001*, 2001

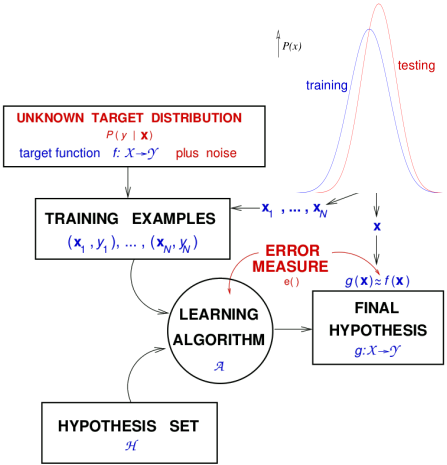
†A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009

‡Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AMLBook, 2012

Desajuste en los datos

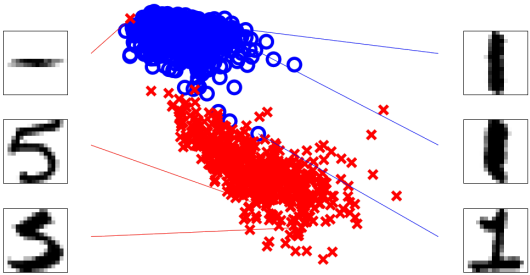
- datos no representativos
  - impide generalización
  - *sampling bias*





Limpieza e ingeniería de datos

- calidad de los datos
  - casos atípicos (*outliers*)
  - información faltante
- ingeniería de características
  - selección de características
  - extracción de características

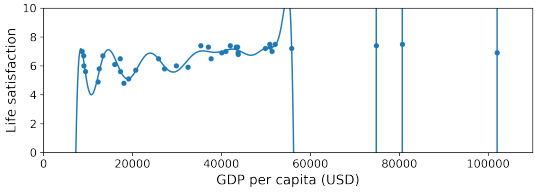


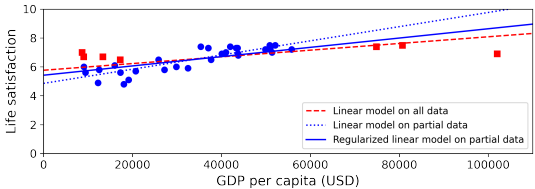
Sobreajuste a los datos

- el modelo no generaliza
  - buen desempeño en *train* (E_{in})
  - mal desempeño en *test* (E_{out})

modelo demasiado complejo!

- posibles soluciones
  - aumentar la cantidad de datos (N)
  - usar modelo menos complejo
  - regularizar el modelo

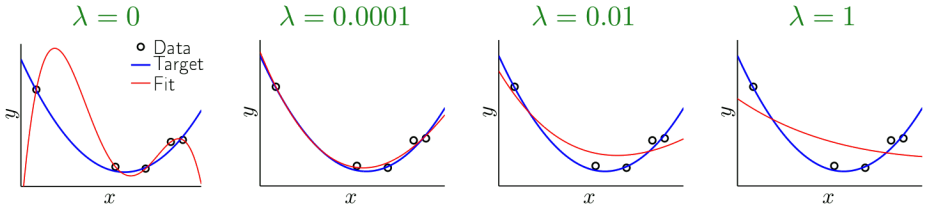




Regularización

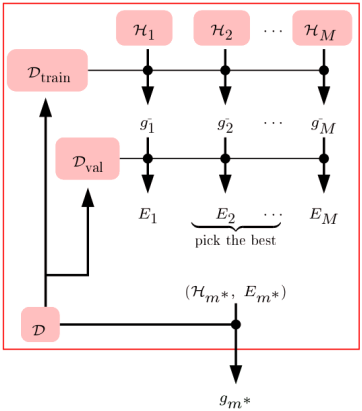
Restringir al modelo para producir una solución más simple.

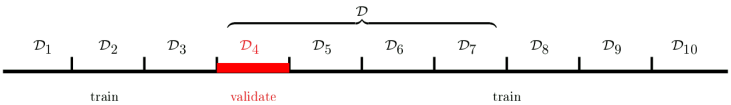
Minimizing $E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N}\,\mathbf{w}^{\mathsf{T}}\mathbf{w}$ for different λ's:



Validación y prueba

- entrenamiento y prueba (*training* y *test*)
  - permite estimar error de generalización
- validación (*holdout validation*)
  - ajuste de hiperparámetros (p. ej. λ)
  - selección de modelos
- validación cruzada (*cross–validation*)
  - partición en subconjuntos (*folds*)
  - valida con uno y entrena con el resto
  - se repite para todos y se promedia





RECORDAR: el modelo final siempre se entrena usando todos los datos.

Actividades

Problema: Al entrenar un sistema de aprendizaje automático encuentro que hay una diferencia de desempeño importante entre *train* y *validation*.

La diferencia podría deberse a **sobreajuste** a los datos de *train*.

O también a **desajuste** entre los datos de *train* y *validation*.

¿Cómo podría saber cuál es el caso?

Investigar en el Capítulo 1 del libro una estrategia para esto.
Responder cuestionario en la página del curso en EVA.

Referencias


A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition*. O'Reilly Media, Inc., 2022.


Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AMLBook, 2012.


M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proceedings of the 39th ACL 2001*, 2001.


A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.