Métodos de Aprendizaje de Máquinas en Data Science

Clase 01: Introducción

Información General

Curso: ING559 - Métodos de Aprendizaje de Máquinas en Data Science

Profesor: Adrián Soto Suárez (adrian.soto@uai.cl)

Horario: W3-W4

Páginas: avisos por WebCursos y material por GitHub

Sobre mí

Academia / Investigación

- Ingeniero Civil de Industrias Mención Computación
- Terminando el doctorado en *Computer Science*
- Investigo en el área de manejo de datos desde el 2015
- Dicté el curso de Bases de Datos por 5 años en la PUC

Sobre mí

Fuera de la universidad

- Me gusta el fútbol (mucho)
- Me gusta producir música
- Fotógrafo frustrado
- Eventualmente transmito partidas de Age of Empires por Twitch (soy pocket)

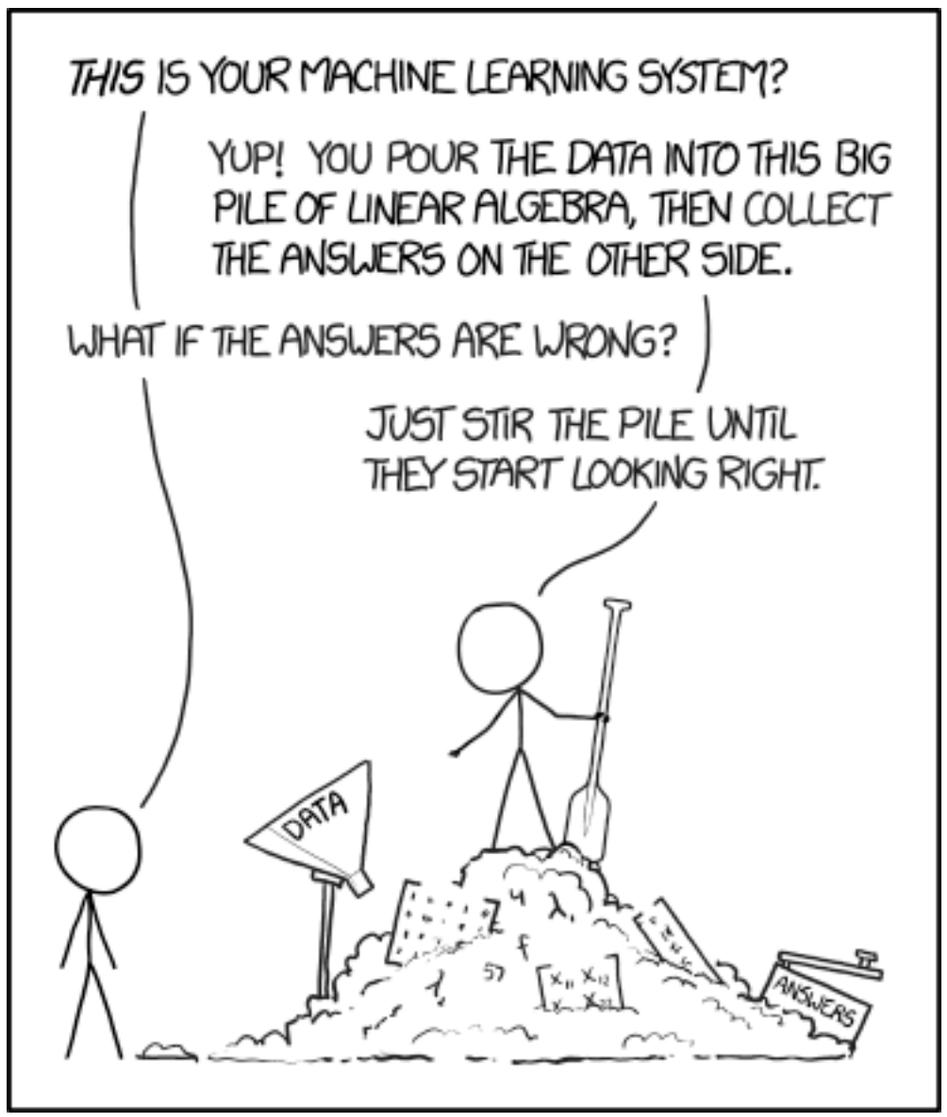
Machine Learning

Machine Learning

¿Qué es *Machine Learning*?

Machine Learning

¿Qué es *Machine Learning*?

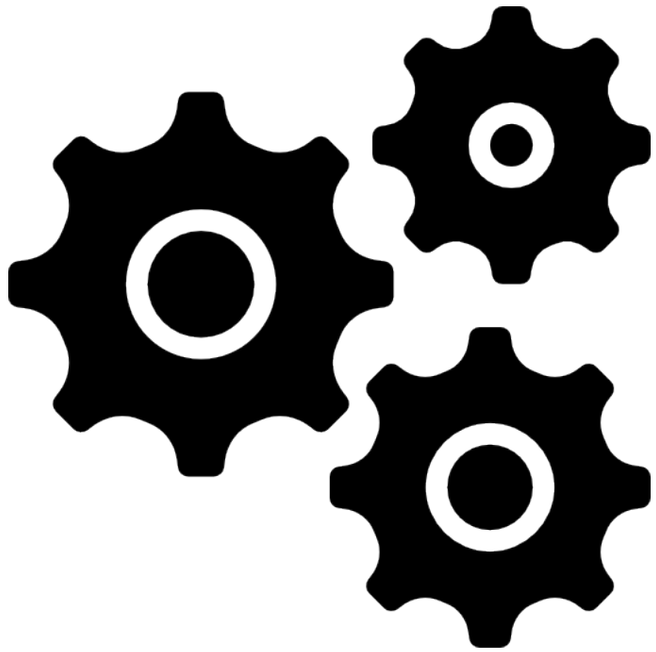


Machine Learning

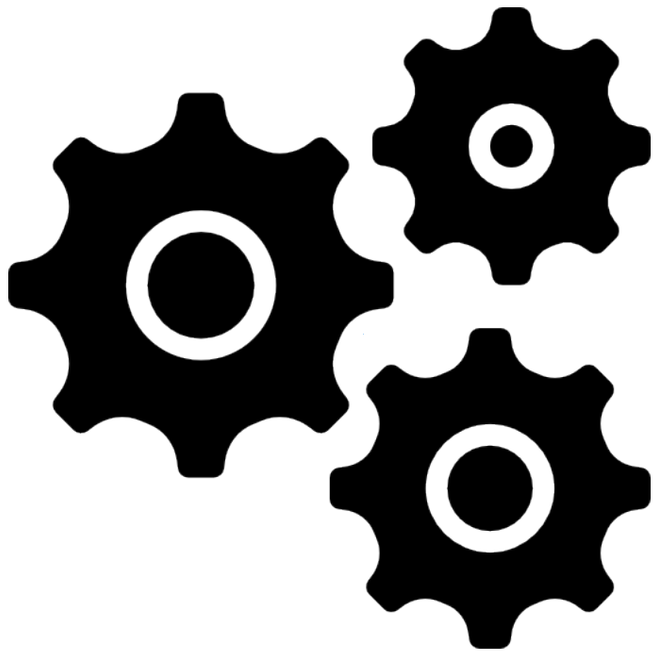
Machine Learning es la ciencia (y el arte) de programar computadores de manera tal que ellos puedan aprender de los datos sin haber dado instrucciones en concreto

Machine Learning

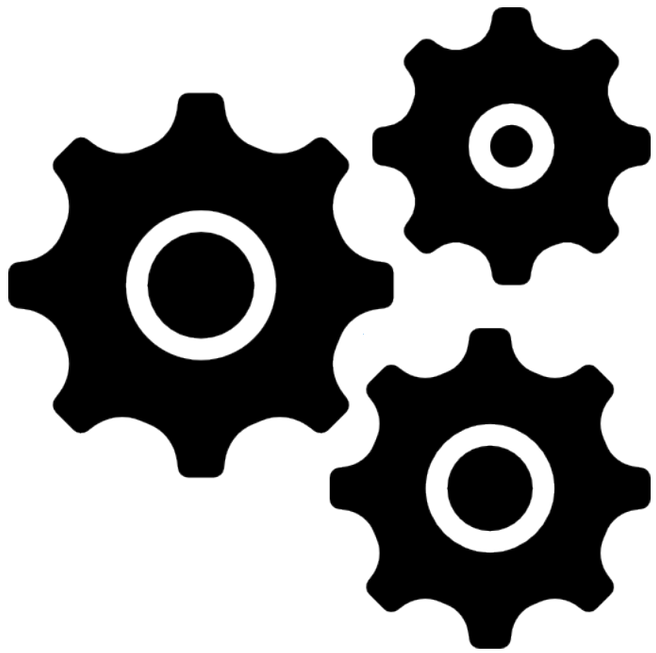
Machine Learning



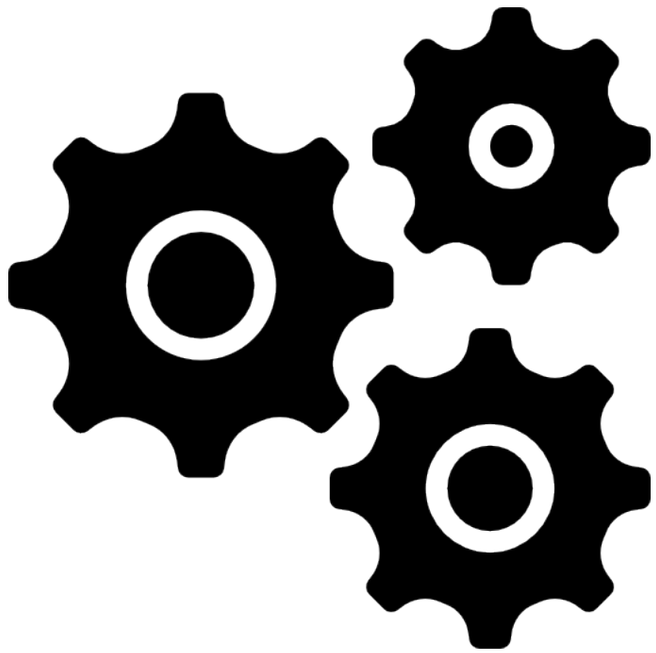
Machine Learning



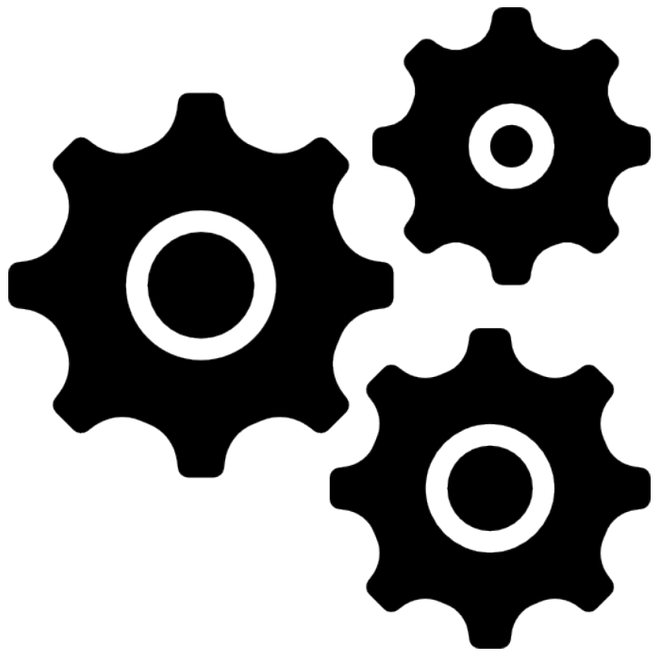
Machine Learning



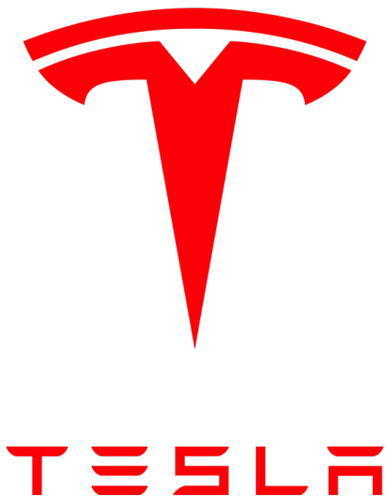
Machine Learning

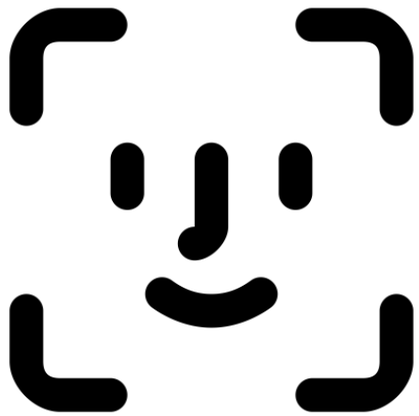


Machine Learning



Machine Learning





Machine Learning

NETFLIX

TESLA

Machine Learning

NETFLIX

TESLA

Cambridge Analytica

Pero profesor, ¿hacer este tipo de cosas está a nuestro alcance?

Machine Learning

En la actualidad existen muchas librerías y frameworks que ponen a nuestro alcance las herramientas del área de *Machine Learning*

Pero también existe un fundamento teórico importante que viene principalmente del campo de la estadística

Machine Learning

El flujo de trabajo en ML

Tenemos que aprender a realizar un proyecto desde el inicio hasta el final:

- Recolectar datos
- Limpiar datos
- Entender los datos (visualizar, correlaciones, ...)
- Entrenar el modelo (o los modelos)
- Entender su rendimiento
- Analizar errores y mejorar
- Llevar a producción

Machine Learning

El modelo

El modelo es el algoritmo de ML particular que vamos a utilizar y lo vamos a entrenar sobre algunos datos conocidos para hacer predicciones

Machine Learning

Ejemplo de modelo - clasificador de spam

Partimos con un clasificador que no hace nada o funciona mal

Machine Learning

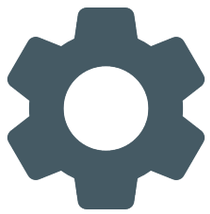
Ejemplo de modelo - clasificador de spam

Pero le comenzamos a mostrar ejemplos

Machine Learning

Ejemplo de modelo - clasificador de spam

Pero le comenzamos a mostrar ejemplos



Machine Learning

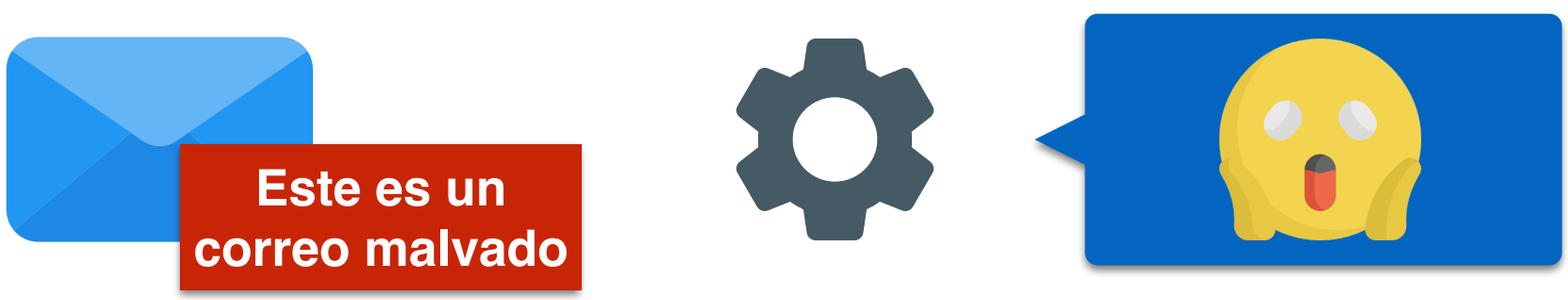
Ejemplo de modelo - clasificador de spam

Pero le comenzamos a mostrar ejemplos

Machine Learning

Ejemplo de modelo - clasificador de spam

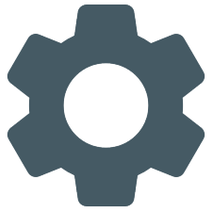
Pero le comenzamos a mostrar ejemplos



Machine Learning

Ejemplo de modelo - clasificador de spam

A medida que nuestro clasificador ve muchos ejemplos, comienza a predecir mejor



Machine Learning

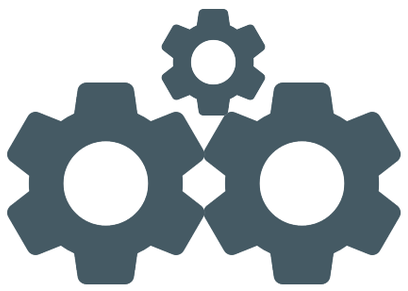
Ejemplo de modelo - clasificador de spam

A medida que nuestro clasificador ve muchos ejemplos, comienza a predecir mejor

Machine Learning

Ejemplo de modelo - clasificador de spam

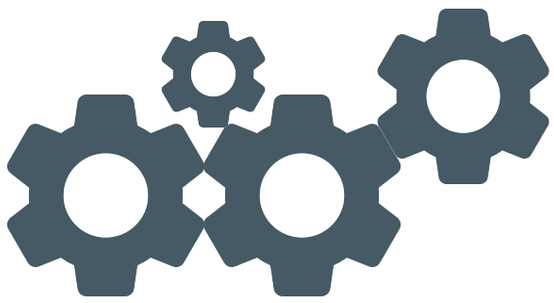
A medida que nuestro clasificador ve muchos ejemplos, comienza a predecir mejor



Machine Learning

Ejemplo de modelo - clasificador de spam

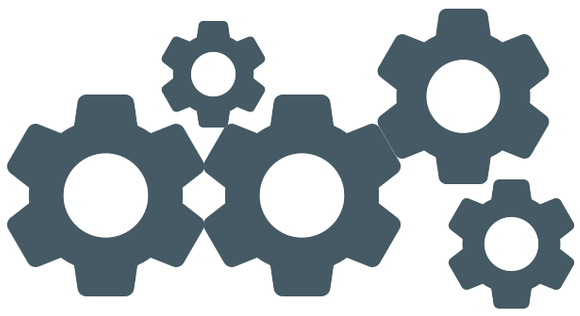
A medida que nuestro clasificador ve muchos ejemplos, comienza a predecir mejor



Machine Learning

Ejemplo de modelo - clasificador de spam

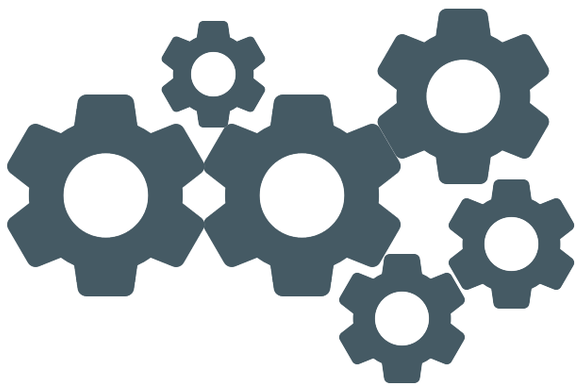
A medida que nuestro clasificador ve muchos ejemplos, comienza a predecir mejor



Machine Learning

Ejemplo de modelo - clasificador de spam

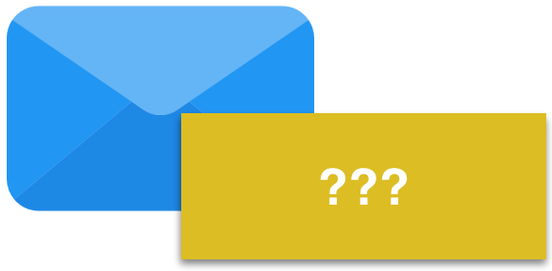
A medida que nuestro clasificador ve muchos ejemplos, comienza a predecir mejor

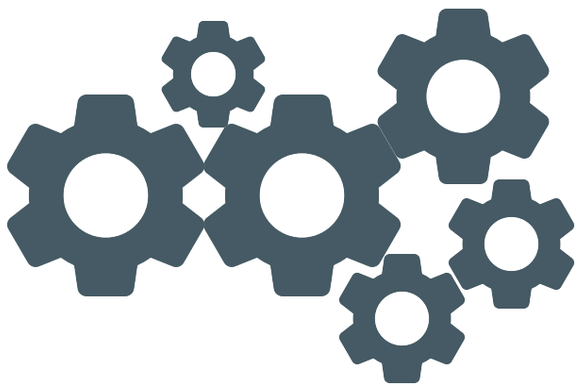


Machine Learning

Ejemplo de modelo - clasificador de spam

Y ahora cuando ve un correo desconocido:

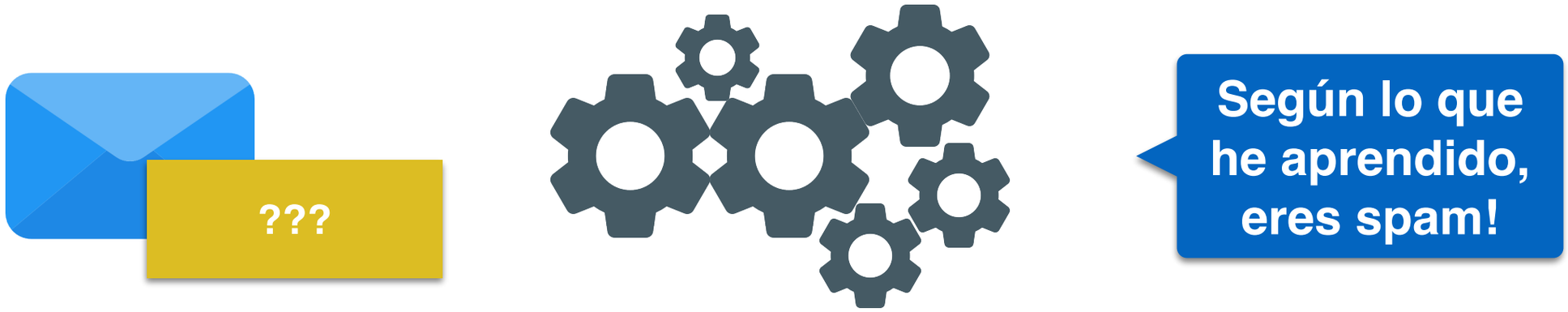




Machine Learning

Ejemplo de modelo - clasificador de spam

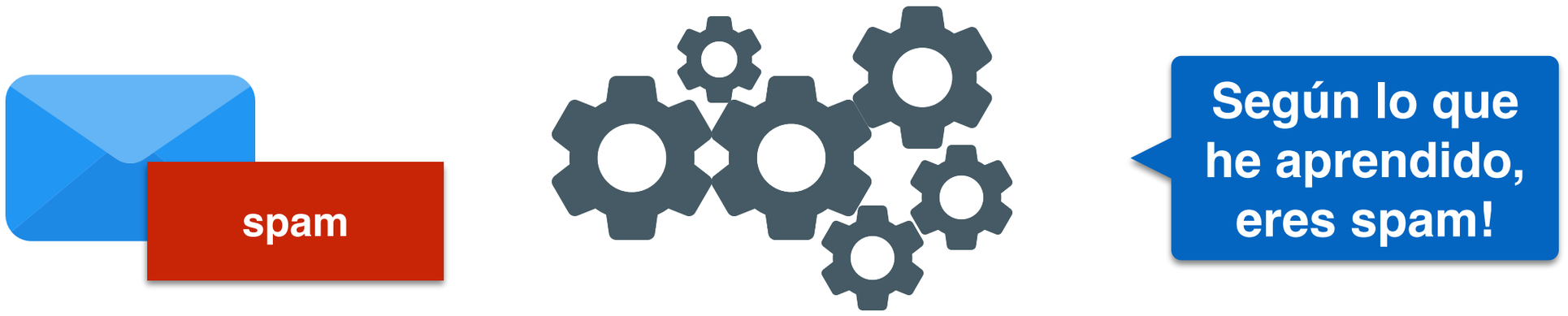
Y ahora cuando ve un correo desconocido:



Machine Learning

Ejemplo de modelo - clasificador de spam

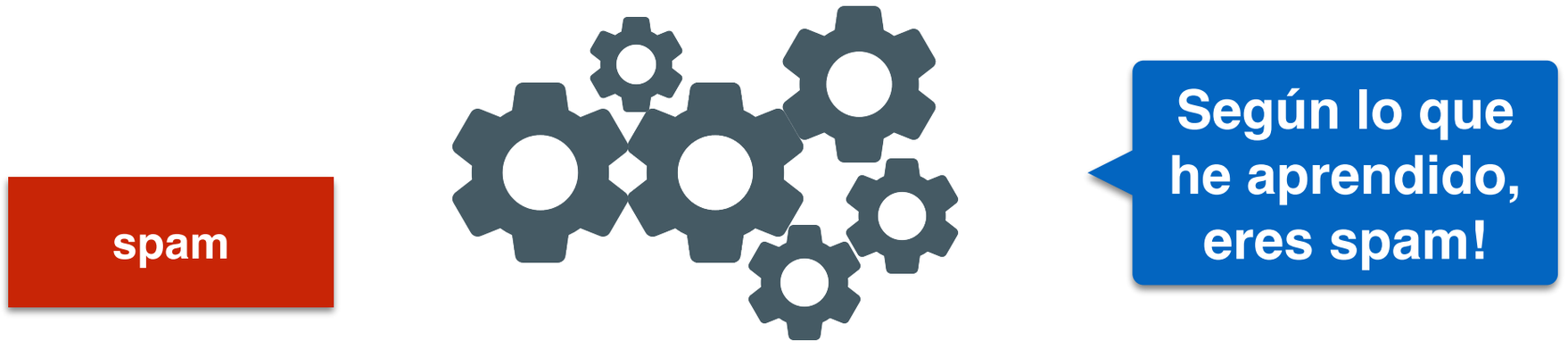
Y ahora cuando ve un correo desconocido:



Machine Learning

Ejemplo de modelo - clasificador de spam

Y ahora cuando ve un correo desconocido:



Machine Learning

Clasificación

En este caso estamos decidiendo si un correo que no hemos visto pertenece a alguna de estas dos clases:

- Clase 1: correo deseado
- Clase 2: correo no deseado

Pero también podemos hacer más cosas!

Machine Learning

Algunas tareas

- Clasificación binaria: clasificamos entre dos clases

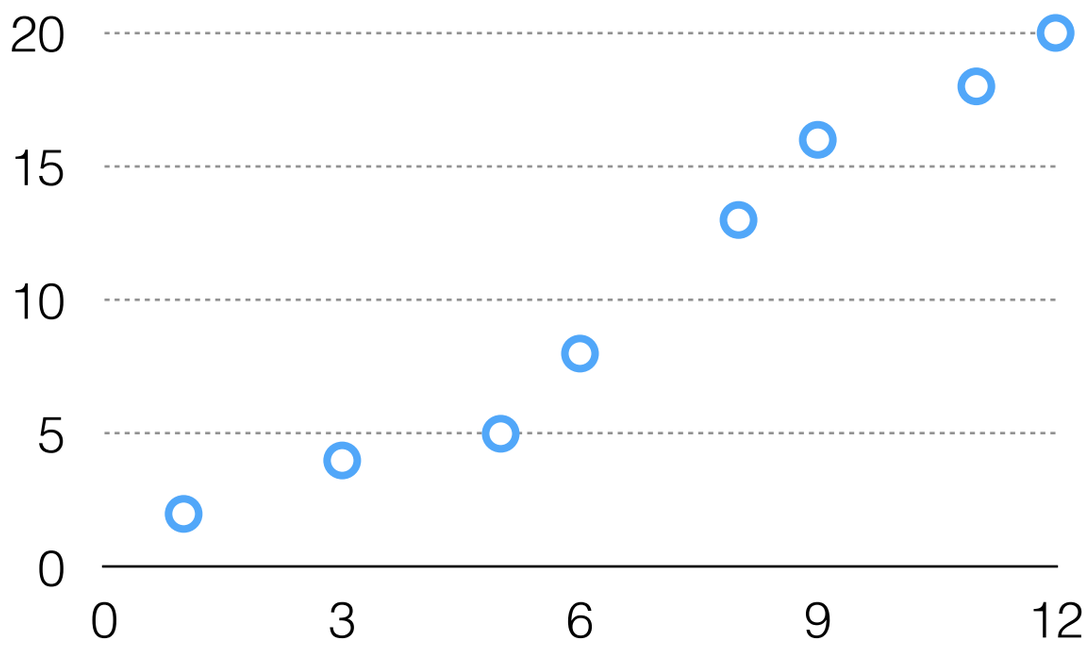
- Clasificación multiclase: clasificamos entre varias clases

- Regresión: buscamos un valor numérico

Machine Learning

Ejemplo - regresión lineal

Tenemos datos de los valores de viviendas en base a los m^2 de las mismas



Machine Learning

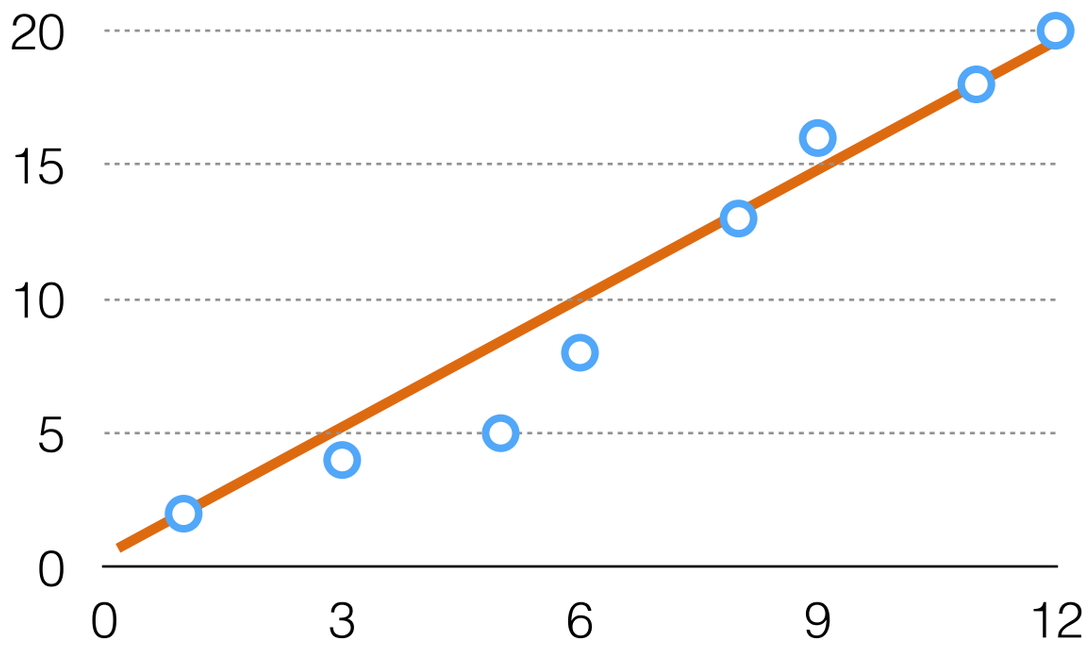
Ejemplo - regresión lineal

Si alguien nos entrega los m^2 de una vivienda que no conocemos, ¿cómo podemos calcular su valor?

Machine Learning

Ejemplo - regresión lineal

Si alguien nos entrega los m^2 de una vivienda que no conocemos, ¿cómo podemos calcular su valor?



Machine Learning

Modelos

Ahora estamos explicando esto como una caja negra: mostramos ejemplos y el modelo aprende

¿Qué tanto se aleja esto de la realidad?

Frameworks de ML

Scikit Learn

```
# Create linear regression object
regr = linear_model.LinearRegression()

# Train the model using the training sets
regr.fit(X_train, y_train)

# Make predictions using the testing set
y_pred = regr.predict(X_test)
```

Frameworks de ML

Existe una amplia variedad de *frameworks* que tienen muchas soluciones implementadas:

- Scikit Learn
- Tensorflow
- Keras
- Pytorch
- ...

La teoría detrás

Todos los modelos están basados en sólidos fundamentos teóricos, que provienen del campo de estadística y álgebra lineal

Entender estos fundamentos nos permite comprender cuando un modelo nos conviene más que el otro

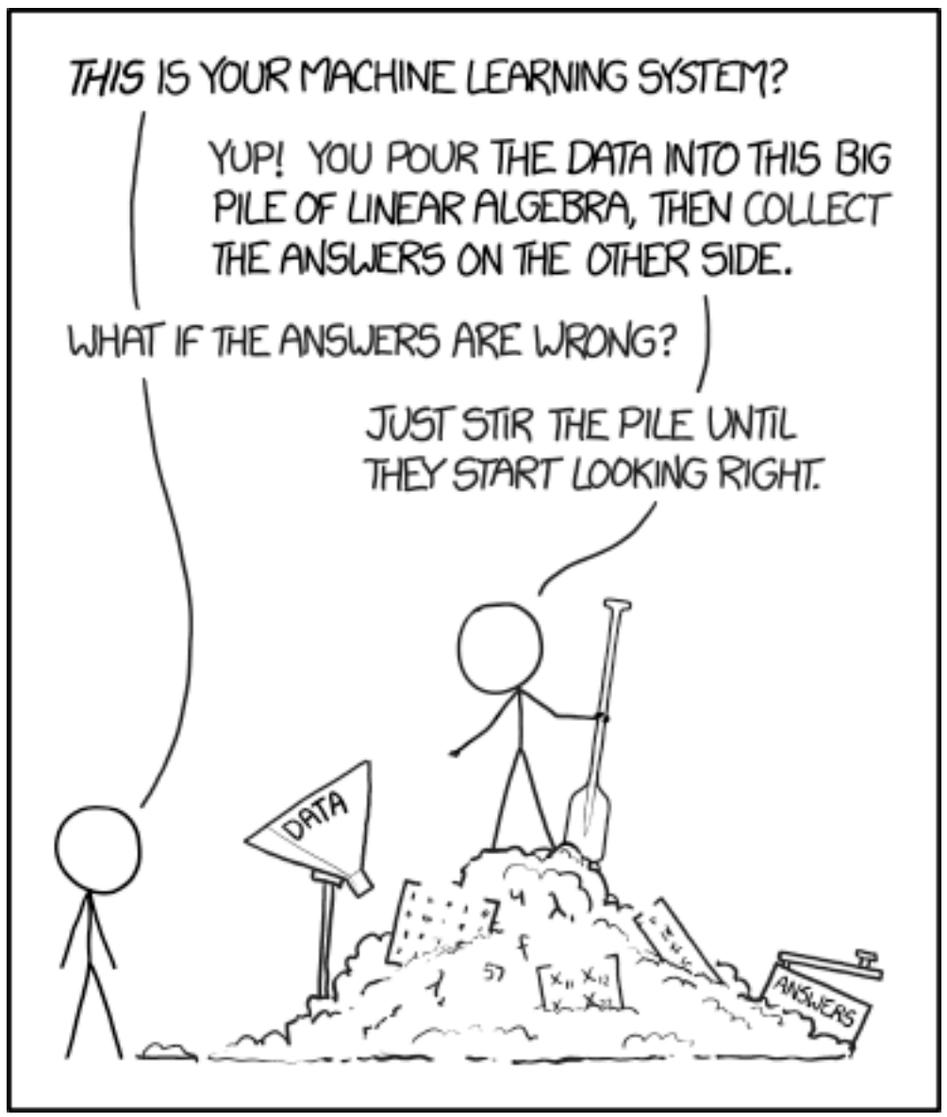
E incluso a veces queremos usar variaciones de los modelos estándar!

Machine Learning

¿Se entiende mejor?

Machine Learning

¿Se entiende mejor?



En este curso

Vamos a necesitar

Saber Python (repaso en primeras dos clases)

Tener conocimiento de probabilidades, estadística y álgebra lineal (repaso a medida que sea necesario)

Saber usar Jupyter, Google Colab y ojalá Git

Saber buscar bien en Google

Vamos a aprender

Herramientas de análisis de datos y visualización:

- Pandas
- Matplotlib
- Altair

Herramientas de análisis numérico:

- Numpy
- Scipy

Vamos a aprender

Fundamentos teóricos de cada uno de los modelos y técnicas

Aprender a usar todos los modelos y programar algunos nosotros mismos

La formalización del problema de aprendizaje

Vamos a aprender

Modelos de aprendizaje supervisado y no supervisado

Conceptos clave del área de *machine learning*

Cómo medir el rendimiento de nuestros programas

Vamos a aprender

Modelos de clasificación y regresión:

- Naive Bayes
- Regresión lineal y polinomial
- Regresión logística
- Regresión logistica
- Árboles de decisión y *random forest*
- SVM
- Redes neuronales

Vamos a aprender

Otros tópicos:

- Reducción de dimensionalidad
- Aprendizaje no supervisado
- Boosting
- Feature engineering

Metodología

Primer módulo: segmento de teoría en el que entendemos los modelos y discutimos el algoritmo detrás

Segundo módulo: manos al código! vamos a programar algunos modelos desde 0 y vamos a utilizar otros

Metodología

La filosofía del curso es que lo que aprendamos, lo tenemos que aprender bien, aunque esto implique cubrir menos contenidos

Además, dependiendo de la situación, pueden existir clases exclusivas para ponernos al día

Y probablemente tengamos invitados que darán charlas súper interesantes

Evaluaciones

Varias actividades del segmento práctico van a llevar nota (idealmente 7) y se borran las dos peores notas

Controles (entre 2 y 4) para la casa que evalúan la teoría y no se borra ninguno

Un proyecto final donde aplican, más o menos libremente, todo lo que han aprendido

Evaluaciones

La nota de controles es el promedio de todos los controles y la nota de actividades es el promedio de todas las notas

El promedio final es:

$$0.45 \cdot \text{Actividades} + 0.25 \cdot \text{Controles} + 0.3 \cdot \text{Proyecto}$$

Y se necesita un promedio mayor o igual a 4 para aprobar

Ahora veamos un ejemplo