Laboratorio de Datos

Clustering: k-medias

Primer Cuatrimestre 2024

Turnos tarde y noche

Facultad de Ciencias Exactas y Naturales, UBA

Introducción a las Técnicas de Clasificación

- **Definición:** La clasificación es una técnica de **aprendizaje supervisado** donde el objetivo es predecir la categoría a la que pertenece una nueva observación, basada en un conjunto de datos de entrenamiento con categorías conocidas.
- **Aplicaciones Comunes:**
  - Diagnóstico médico (clasificación de enfermedades)
  - Reconocimiento de imágenes (clasificación de objetos)
  - Filtrado de correos electrónicos (spam vs no spam)
  - Detección de especies de pingüinos

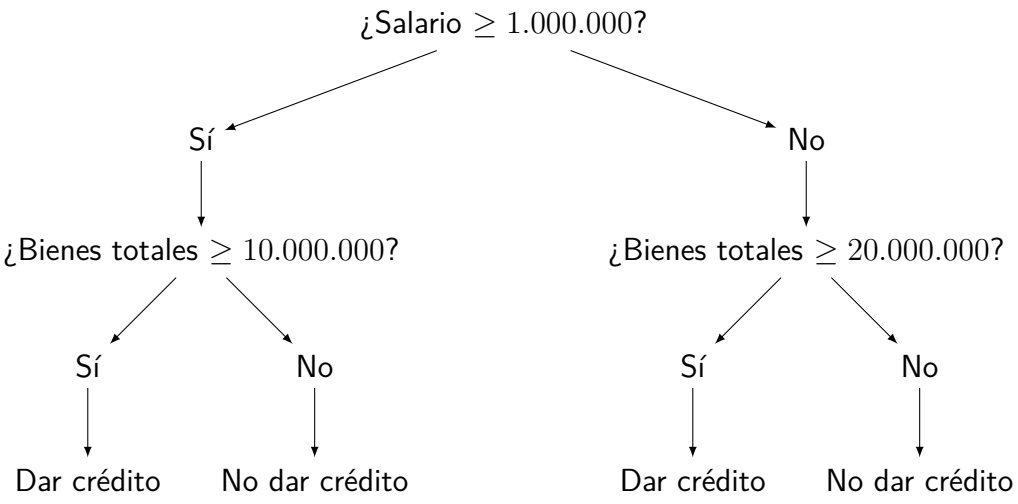
Técnicas Comunes de Clasificación

- **Regresión Logística:**
  - Método de regresión para modelar la probabilidad de una variable binaria.
  - Utiliza una función logística para limitar los valores predichos entre 0 y 1.
- **K-Vecinos Más Cercanos (K-NN):**
  - Clasifica basándose en la categoría de los vecinos más cercanos a un dato.
  - Sencillo pero puede ser computacionalmente costoso.

Otras Técnicas de Clasificación

- **Árboles de Decisión:**
  - Modelo basado en reglas de decisión en forma de árbol
  - Fácil de interpretar y visualizar
- **Máquinas de Vectores de Soporte (SVM):**
  - Encuentra el hiperplano que maximiza el margen entre las diferentes clases
  - Eficaz en espacios de alta dimensión
- **Redes Neuronales:**
  - Modelos inspirados en el funcionamiento del cerebro humano
  - Capaces de aprender representaciones complejas

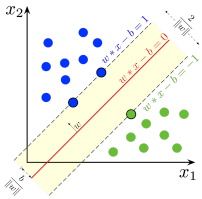
Árboles de decisión: ejemplo



Support vector machine: ejemplo

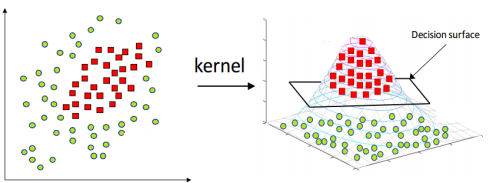
Caso lineal

Separamos los puntos por rectas o planos.



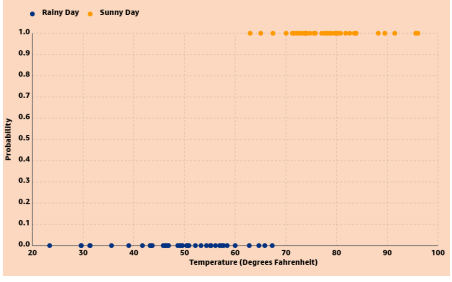
Caso no lineal

Antes de separar por un plano, aplicamos una transformación de los datos.



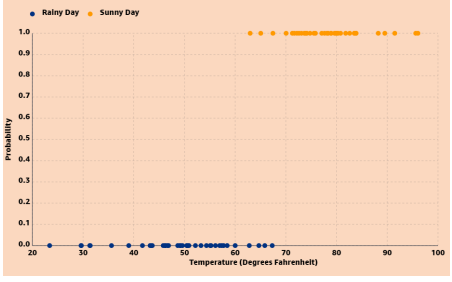
Regresión Logística en 1 minuto

- ¿Cómo podemos predecir una variable respuesta binaria $(0/1)$?



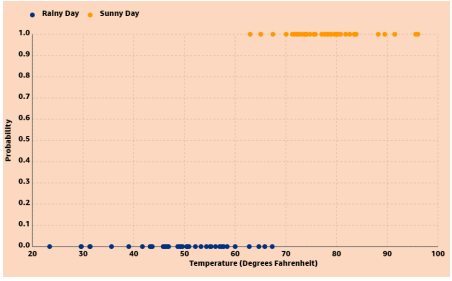
Regresión Logística en 1 minuto

- ¿Cómo podemos predecir una variable respuesta binaria $(0/1)$?
- Considerar a y como una variable numérica y ajustar una función lineal $y = \beta_0 + \beta_1 x$ no va a ser apropiado.



Regresión Logística en 1 minuto

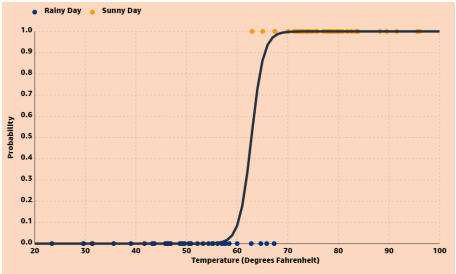
- ¿Cómo podemos predecir una variable respuesta binaria $(0/1)$?
- Considerar a y como una variable numérica y ajustar una función lineal $y = \beta_0 + \beta_1 x$ no va a ser apropiado.
- Queremos una función que para valores grandes positivos de x, el valor de y se acerque a 1 y para valores grandes negativos el valor de y se acerque a 0.



Regresión Logística en 1 minuto

Le aplicamos una transformación a la función lineal.
En regresión logística tomamos

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}.$$



Si $\beta_1 > 0$, cuanto mas grande es x, más cercano a 1 es y.

Si tenemos varias variables explicativas, usamos

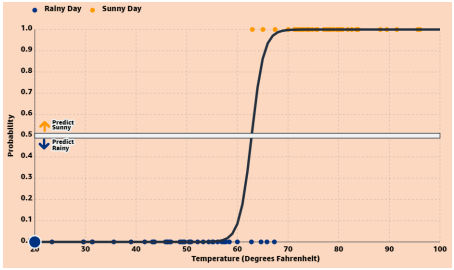
$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}.$$

Regresión Logística en 1 minuto

Finalmente usamos la curva para clasificar.

Tomamos los valores de y mayores que 0.5 como 1 y los menoes como 0 (o podemos fijar el corte en otro valor arbitrario).

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$



Métodos paramétricos vs. no paramétricos

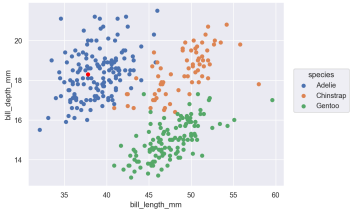
- Los modelos de regresión lineal o regresión logística son modelos *paramétricos*.
- Asumen una forma específica para la relación entre las variables independientes y la variable dependiente.
- Se define una ecuación paramétrica con un número fijo de parámetros (coeficientes) que determinan esta relación, por ejemplo $y = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$.
- Los métodos paramétricos hacen suposiciones sobre la distribución de los datos y la estructura del modelo, lo que puede proporcionar predicciones más eficientes y precisas si estas suposiciones son correctas.

¿Cómo sería un método no-paramétrico?

Modelo no paramétrico para clasificación

Tenemos un conjunto de pingüinos para los que sabemos su especie, y queremos saber la especie de un nuevo pingüino.

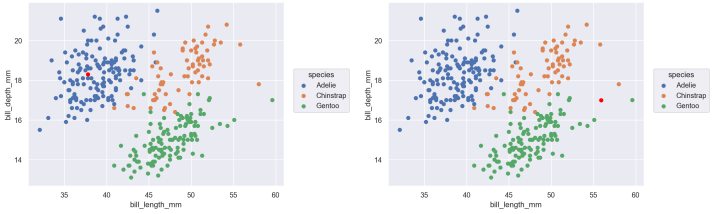
¿A qué especie pertenece el pingüino rojo?



Modelo no paramétrico para clasificación

Tenemos un conjunto de pingüinos para los que sabemos su especie, y queremos saber la especie de un nuevo pingüino.

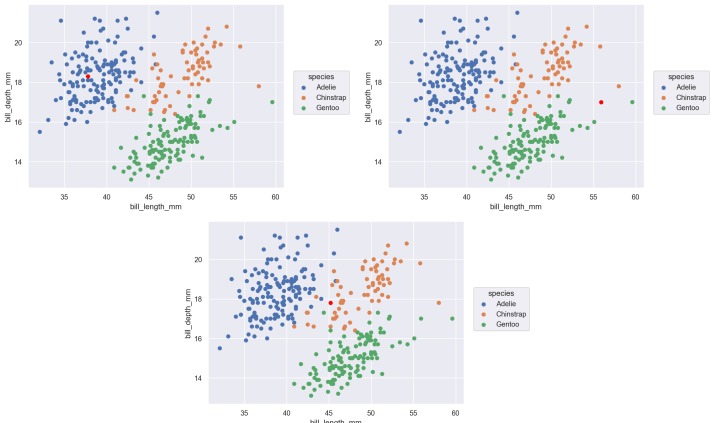
¿A qué especie pertenece el pingüino rojo?



Modelo no paramétrico para clasificación

Tenemos un conjunto de pingüinos para los que sabemos su especie, y queremos saber la especie de un nuevo pingüino.

¿A qué especie pertenece el pingüino rojo?



Algoritmo K-Vecinos Más Cercanos (K-NN)

- **Definición:**
  - K-NN (K nearest neighbors) es un algoritmo de aprendizaje supervisado utilizado para clasificación (y regresión).
  - Se basa en la idea de que objetos similares están cerca unos de otros.
- **Funcionamiento:**
  - Para clasificar un nuevo punto, el algoritmo busca los K puntos más cercanos en el conjunto de entrenamiento.
  - La categoría más común entre estos K vecinos se asigna al nuevo punto.
- **Definición:** En estadística, el valor que se repite más veces en un conjunto de valores se denomina *moda*.

Selección del K óptimo

Como es un modelo de aprendizaje supervisado, podemos elegir el K óptimo con alguna estrategia de validación.

Esquema posible de validación:

1. Separamos el conjunto de todos los datos en 80 % entrenamiento y 20 % testeo
2. Elegimos el valor de K por validación cruzada en la muestra de entrenamiento.
3. Probamos la perfomance de nuestro modelo en la muestra de testeo.

¿Ajuste del modelo?

Preguntas:

- ¿Cuál es el valor a optimizar en la validación? Típicamente, elegimos el valor de K que nos de el mayor porcentaje de aciertos.
- ¿Qué significa entrenar nuestro modelo? Como es un modelo no-paramétrico, no hay que aprender ningún parámetro. Entrenar es simplemente guardar los datos.

Validación por pliegos (K-fold pero es otro K!)

Podemos hacer validación cruzada en pliegos:

- Dividimos nuestra muestra en 5 pliegos (o cualquier otra cantidad).
- Seleccionamos 4 pliegos para entrenamiento y 1 pliego para validación.
- Para cada uno de los datos de validación, buscamos los K vecinos más cercanos en los pliegos de entrenamiento.
- Elegimos el valor más votado entre los vecinos.

Validación “leave-one-out” (dejar uno afuera)

Validación leave-one-out.

- Es el caso extremo de validación en pliegos, donde cada observación es un pliego.
- Para cada observación, buscamos los K vecinos más cercanos entre todos los demás puntos.
- Elegimos el valor más votado entre los vecinos.

En modelos de regresión, validación leave-one-out puede ser costoso porque debemos entrenar nuevamente el modelo para cada observación.

En KNN, como no hay que ajustar parámetros, este método no tiene costo adicional comparado con usar pocos pliegos.

Características y Aplicaciones del Algoritmo K-NN

- **Ventajas:**
  - Fácil de entender e implementar.
  - No requiere un modelo paramétrico, lo que lo hace flexible.
  - No se ve afectado por la presencia de outliers.
- **Desventajas:**
  - Computacionalmente costoso para grandes conjuntos de datos debido a la necesidad de calcular distancias.
  - El rendimiento puede verse afectado por la elección del valor de K y la escala de las características.
  - Si tenemos muchas variables explicativas, sufriremos la maldición de la dimensionalidad.

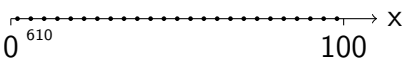
La maldición de la dimensionalidad

La maldición de la dimensionalidad se refiere a diversos problemas que surgen cuando el análisis se realiza en espacios de alta dimensión.

- **Efecto en K-NN:**
  - **Densidad:** Los datos se vuelven más dispersos y la densidad de los puntos de datos disminuye rápidamente.
  - **Relevancia de características:** En alta dimensión, muchas características pueden ser irrelevantes o ruidosas, afectando negativamente el rendimiento del modelo.
- **Consecuencias:**
  - La eficiencia y precisión del K-NN disminuyen en espacios de alta dimensión.
  - Se necesitan técnicas de reducción de dimensionalidad como PCA para mitigar estos efectos.

La maldición de la dimensionalidad...

Si colocamos 25 nodos en un rango de 0 a 100m, la máxima distancia de un punto cualquiera a un nodo será de 2m.



Si los colocamos en un cuadrado de $100m \times 100m$, la máxima distancia será $\sqrt{200} = 14.14\ldots$.