Tecnológico de Monterrey

Aprendizaje automático

Clasificación

Dr. Omar Mendoza Montoya

19/08/2024

En aprendizaje automatizado, clasificación se refiere a la identificación de la categoría a la que una observación pertenece.

Clasificación

¿Es un perro o un gato?

Ejemplos de problemas de clasificación

- Una persona llega a la sala de emergencia presentando algunos síntomas, los cuales se pueden atribuir a tres condiciones médicas. ¿Cuál condición medica sería la que correspondería a los síntomas presentados?

- En un servicio de banca en línea se debe determinar si una transacción realizada fue fraudulenta, tomando como base la IP del usuario, el historial de transacciones, y otras variables.

- De acuerdo con la secuencia de ADN extraída de un grupo de sujetos, de los algunos presentan una enfermedad, un biólogo quisiera determinar qué mutaciones en el ADN son perjudiciales al causar la enfermedad y cuales no.

Clasificación

- Formalmente, un modelo de clasificación es una función que mapea un vector de entrada a una categoría o etiqueta.

$$\hat{L} = f(X; \beta)$$

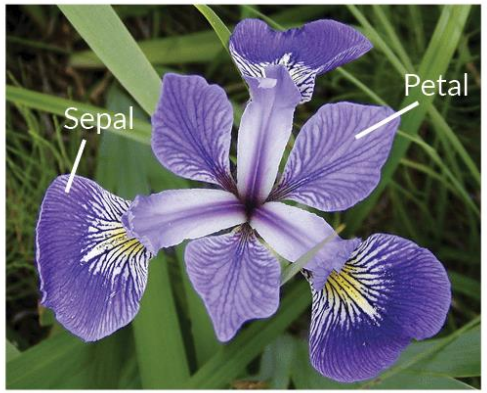
X = Vector de p variables (ordinales, booleanas o categóricas)

β = Parámetros del modelo

$\hat{L}$ = Etiqueta o categoría de la observación X

- Cada variable x_i en el vector X es llamada predictor, característica, dimensión, etc.
- $\hat{L}$ es la etiqueta predicha por el modelo y L es la verdadera etiqueta.
- La pareja (X, L) es una observación o muestra.

Ejemplo: El conjunto de datos Iris



Iris Versicolor

Iris Setosa

Iris Virginica

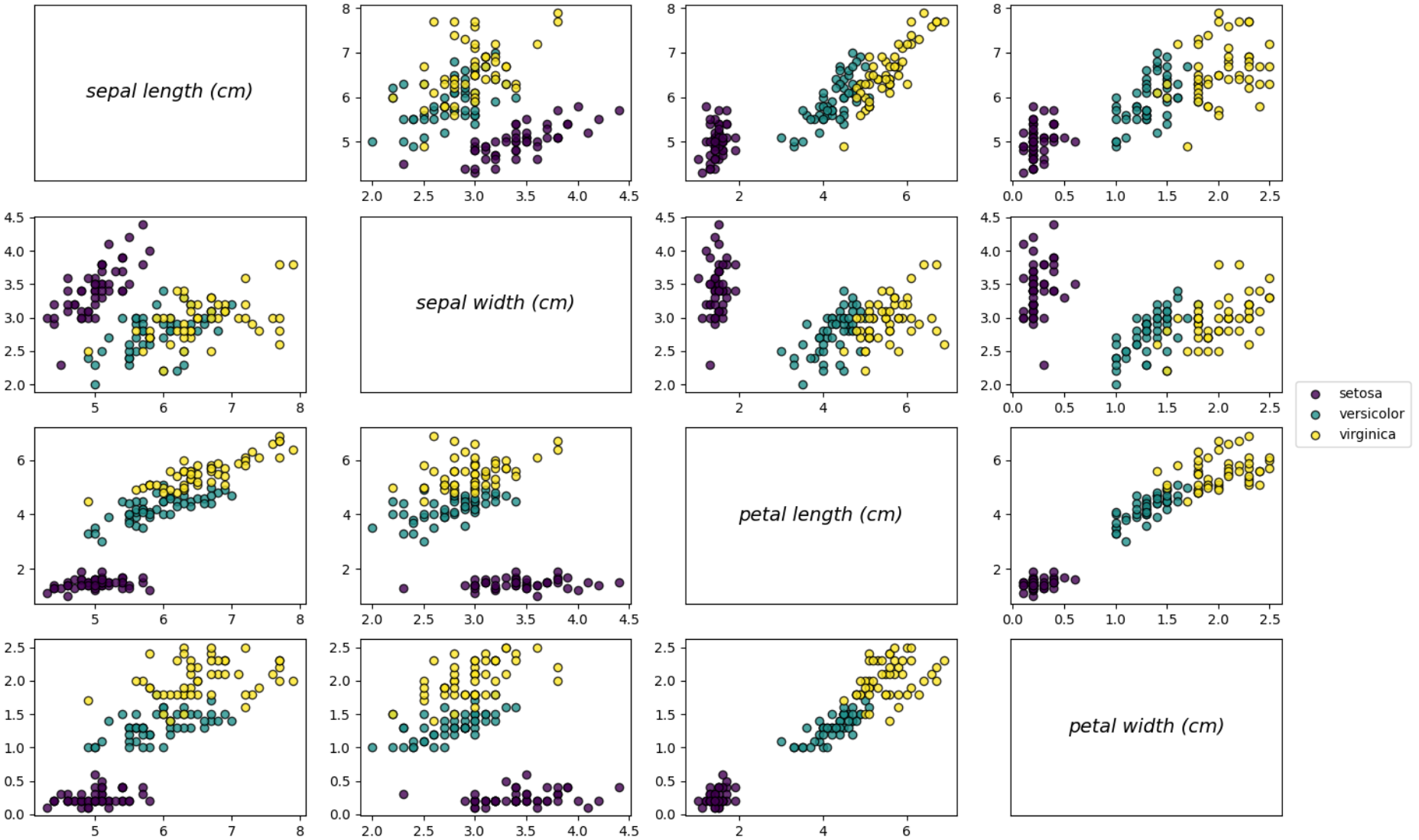
Etiquetas o clases

	Longitud del sépalo	Ancho del sépalo	Longitud del pétalo	Ancho del pétalo	Clase
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

Observaciones

Predictores, características, variables, espacio de características

Ejemplo: El conjunto de datos Iris



Ejemplo: El conjunto de datos Iris

$$f(X;\beta) = \begin{cases} Setosa & \text{si } X \text{ está cerca del grupo Setosa} \\ Versicolor & \text{si } X \text{ está cerca del grupo Versicolor} \\ Virginica & \text{si } X \text{ está cerca del grupo Virginica} \end{cases}$$

$X = [x_1, x_2, x_3, x_4]$ Vector de 4 variables

x_1 = Longitud del sépalo

x_2 = Ancho del sépalo

x_3 = Longitud del pétalo

x_4 = Ancho del pétalo

$L \in \{Setosa, Versicolor, Virginica\}$ Categorías

Independientemente del modelo de clasificación, requerimos de un conjunto de datos de entrenamiento para calcular los parámetros β del modelo.

El proceso de encontrar los parámetros del modelo de clasificación utilizando un conjunto de entrenamiento es lo que se conoce como aprendizaje supervisado.

Ejemplo: el conjunto de datos Iris

- Dado un conjunto de entrenamiento con N observaciones $\{(X_1, L_1), (X_2, L_2), \ldots, (X_n, L_n)\}$, el algoritmo de entrenamiento calcula el vector de parámetros β de tal forma que $f(X; \beta)$ se comporta lo mejor posible para los datos de entrenamiento.

- Por ejemplo, resolviendo el siguiente problema de optimización se obtendría el conjunto de pesos β^* que minimizarían el error de predicción del modelo:

$$\beta^* = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{N} I(f(X_i; \beta) \neq L_i)$$

donde $I(T)$ es la función indicadora, y regresa 1 si T es verdadero, o 0 si T es false.

Problemas en el ajuste de modelos de clasificación

- Es necesario seleccionar un modelo de entre muchas posibilidades (selección de modelo).
- Después de seleccionar un modelo, es necesario estimar los parámetros del modelo (entrenamiento del modelo).
- Una vez que se tiene un modelo candidato, es necesario determinar su habilidad para predecir correctamente las etiquetas de observaciones que no hayan sido utilizadas en el entrenamiento (evaluación del modelo).

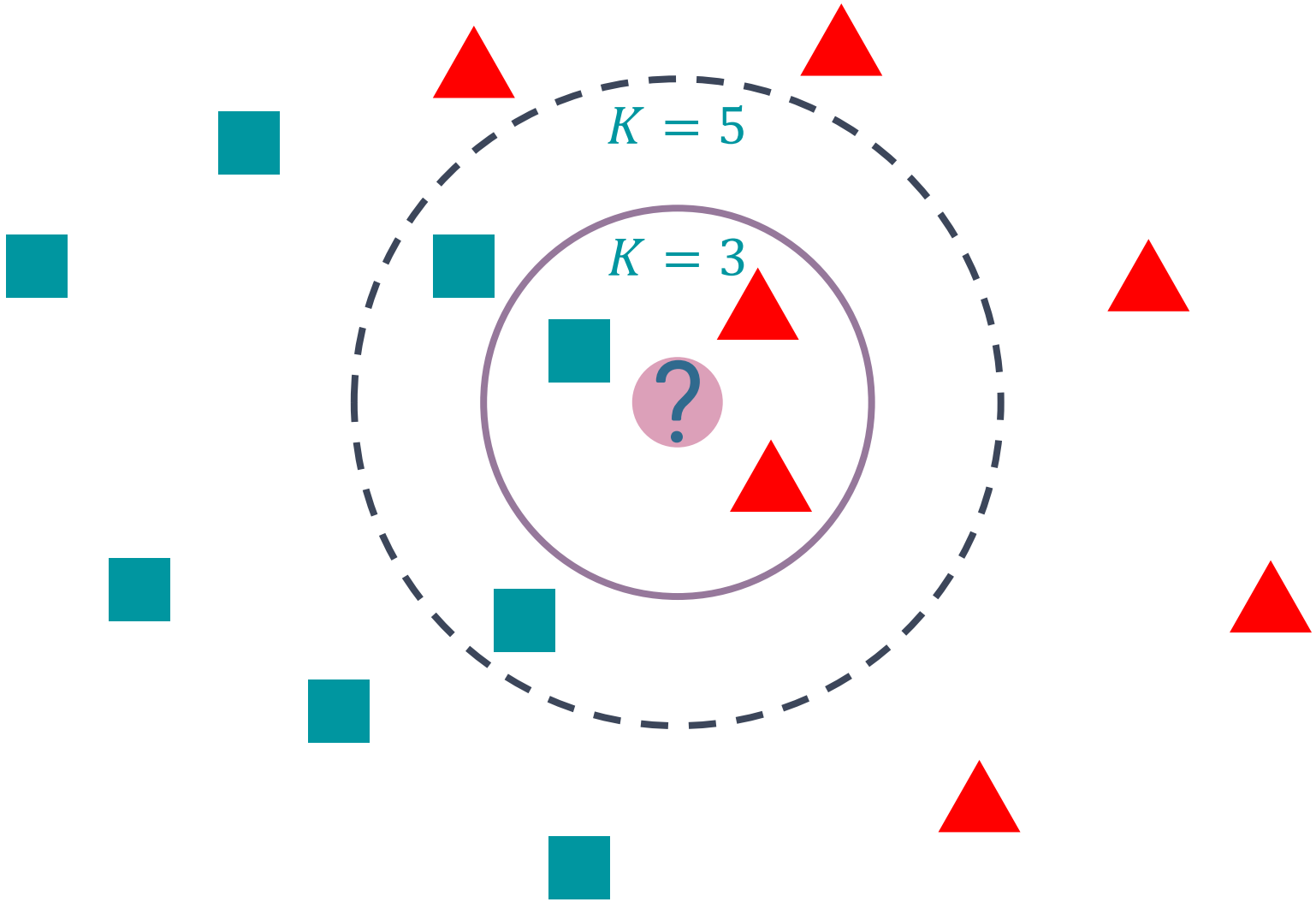
Ejemplos de modelos de clasificación

K-vecinos más cercanos (k-NN)

- Método multiclase, en el que se determinan las k observaciones del conjunto de entrenamiento más cercanas a la observación a clasificar.

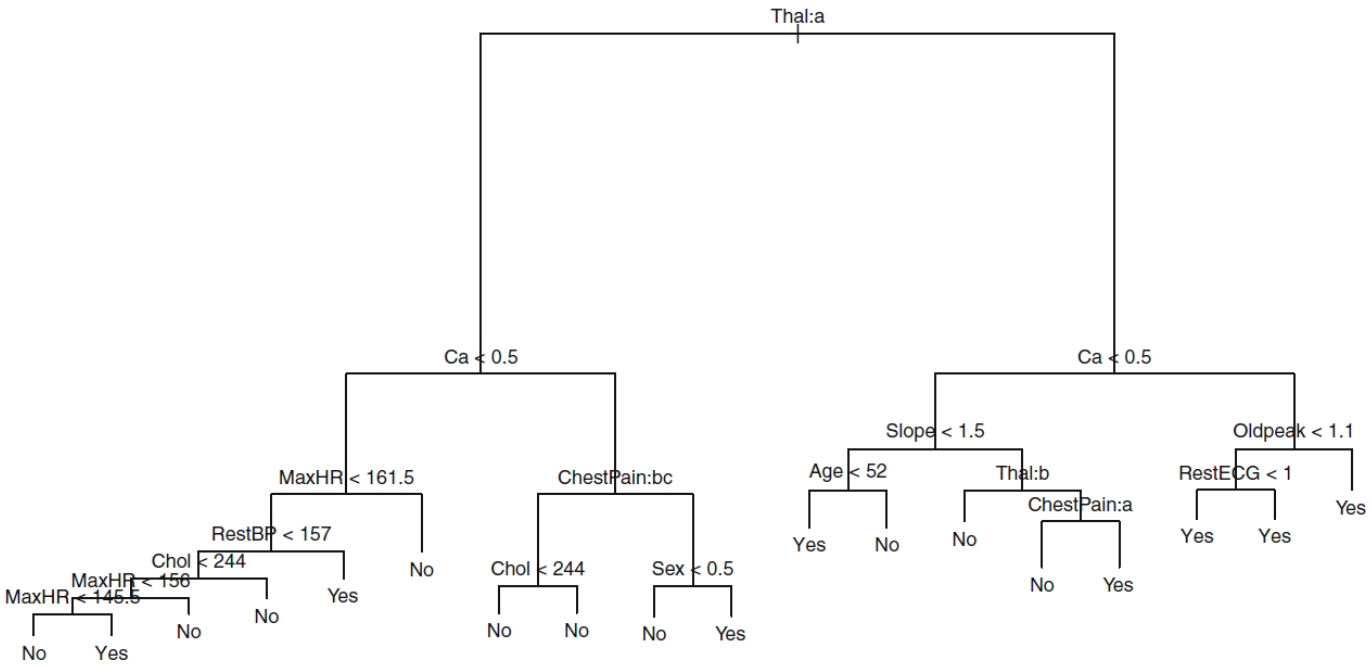
- Se realiza una votación entre las k elementos más cercanos para determinar la clase de la observación que se está evaluando. Si $k = 1$, la clase simplemente corresponde a la observación más cercana.

K-vecinos más cercanos (k-NN)



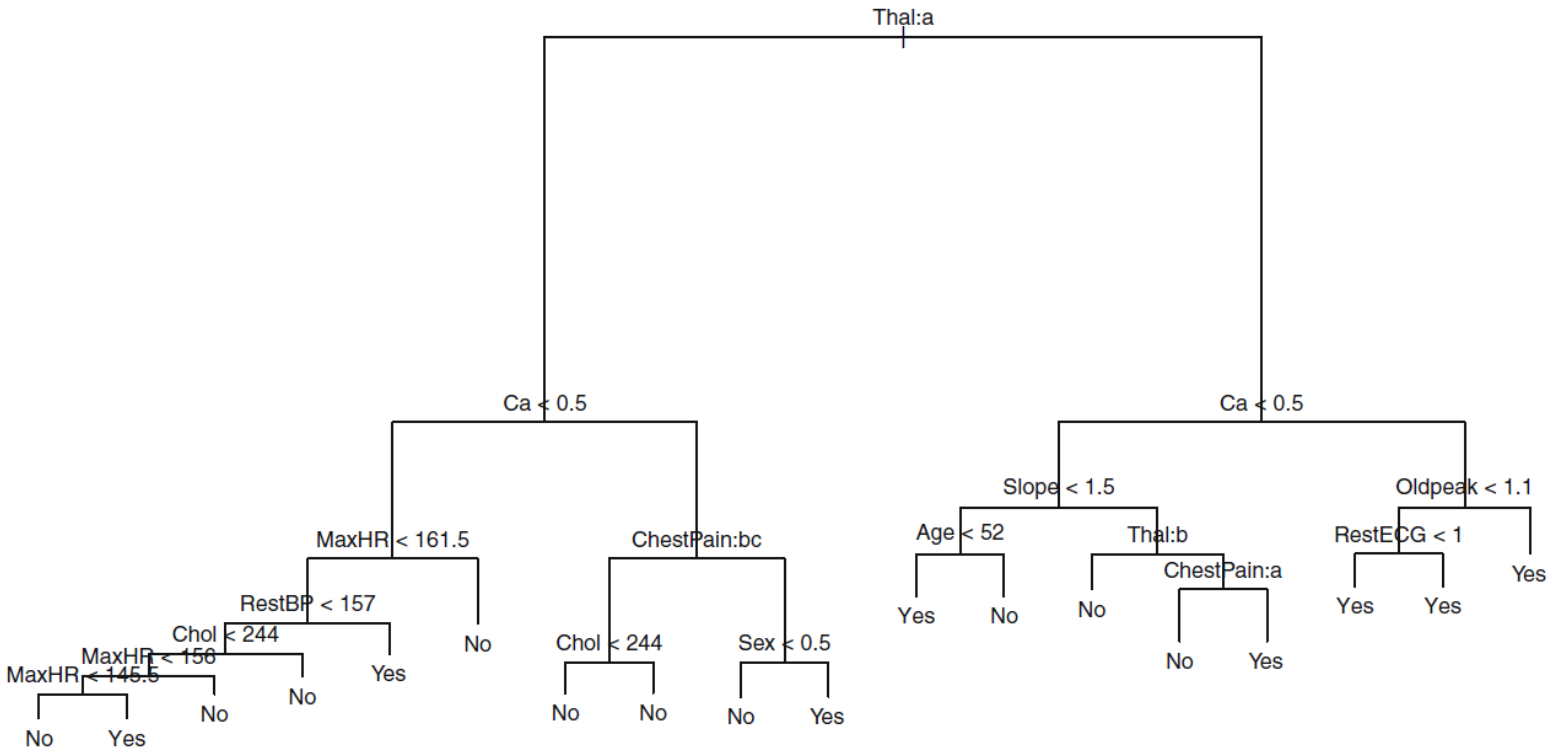
Árboles de decisión

- Los árboles de decisión son una estructura jerárquica en la cual cada nodo representa una prueba sobre un atributo o característica, cada rama representa el resultado de la prueba, y cada nodo hoja representa una etiqueta o resultado.



James, Witten, Hastie, & Tibshirani, 2023

Árbol de decisión



James, Witten, Hastie, & Tibshirani, 2023

- El árbol se construye buscando la variable o predictor para el cual se puede definir la regla que clasifica mejor los datos que se tienen en la rama correspondiente.
- Este proceso se extiende hasta llegar al caso que se tienen nodos con elementos con la misma clase.

Árboles de decisión

- Estos modelos aplican tanto a problemas de regresión como de clasificación.

- Por lo regular, por si solos no brindan una mejora respecto a otros modelos de aprendizaje automático. Sin embargo, tienen algunas propiedades que los hacen interesantes:
  - Permiten manejar de manera natural predictores que sean variables categóricas.
  - Se pueden combinar varios árboles de decisión fácilmente en un ensamble para generar clasificadores de mayor poder predictivo.

Clasificador logístico

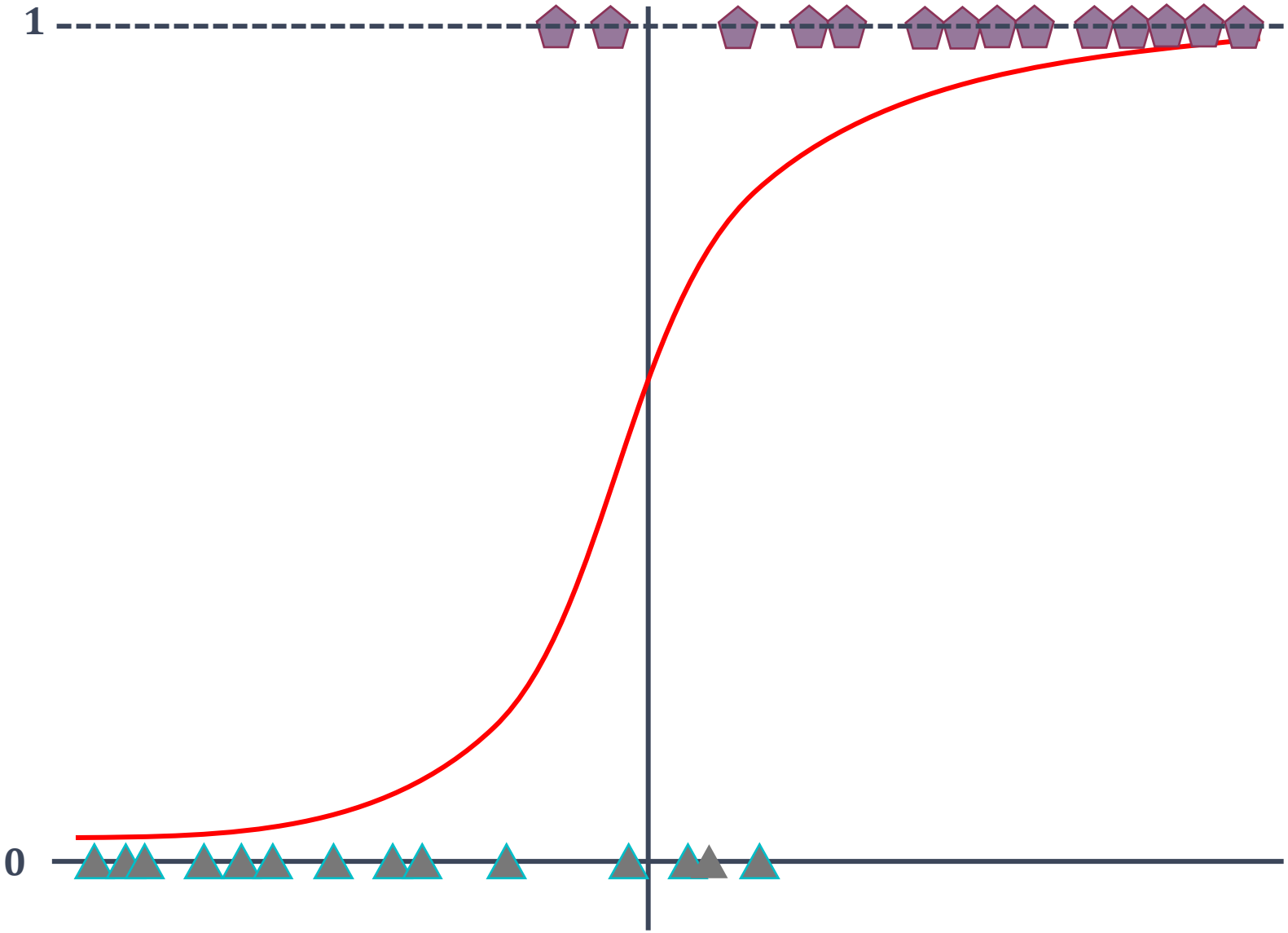
- Este clasificador está basado en el cálculo de las probabilidades condicionales $P(L = l|X = x)$ con la función logística:

$$P(L = 1|X = x) = \frac{1}{1 + e^{\beta^T x + \beta_0}}$$

$$P(L = 0|X = x) = \frac{e^{\beta^T x + \beta_0}}{1 + e^{\beta^T x + \beta_0}}$$

Donde $\beta = [\beta_1, \beta_2, \beta_3, \dots, \beta_p]$ es un vector de p valores, y β_0 es el intercepto del modelo.

Clasificador logístico



Clasificador logístico

- Para encontrar β y β_0, se resuelve el siguiente problema de optimización para un conjunto de N observaciones:

$$\arg\max_{\beta,\beta_0} \sum_{k=1}^{N} \ln\left(P(L = l_k | X = x_k)\right) - \lambda \|\beta\|_2^2$$

Donde $\|a\|_2$ es la norma l^2 del vector a, y λ es un factor de regularización que nos permite penalizar la función de costo cuando la norma del vector de coeficientes β es demasiado grande.

- Para clasificar un nuevo dato x, basta con determinar cuál probabilidad $P(L = 1|X = x)$ o $P(L = 0|X = x)$ es mayor.

Análisis discriminante lineal (LDA)

- LDA es un clasificador lineal, el cual se obtiene al encontrar las distribuciones de probabilidad condicional $P(L = l|X = x)$ para cada clase utilizando la regla de Bayes.

- La regla de decisión se determina maximizando la probabilidad condicional:

$$P(L = l|X = x)$$

asumiendo que las funciones de densidad $P(X = x|L = 0)$ y $P(X = x|L = 1)$ son ambas normales con la misma matriz de covarianza.

Análisis discriminante lineal (LDA)

- Si μ_1 y μ_2 son las medias de las distribuciones de $P(X = x|L = 0)$ y $P(X = x|L = 1)$, y Σ_1 y Σ_2 son sus matrices de covarianza ($\Sigma_1 = \Sigma_2 = \Sigma$), entonces:

$$P(L = l|X = x) = \frac{P(X = x|L = l)P(L = l)}{P(X = x)} = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} e^{-\frac{1}{2}(x-\mu_l)'\Sigma^{-1}(x-\mu_l)} \frac{P(L = l)}{P(X = x)}$$

- La regla de decisión queda expresada como:

$$P(L = 0|X = x) > P(L = 1|X = x)$$

$$\frac{1}{\sqrt{(2\pi)^p|\Sigma|}} e^{-\frac{1}{2}(x-\mu_0)'\Sigma^{-1}(x-\mu_0)} \frac{P(L = 0)}{P(X = x)} > \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} e^{-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)} \frac{P(L = 1)}{P(X = x)}$$

Análisis discriminante lineal (LDA)

- Sacando logaritmos y simplificando:

$$(\mu_0' - \mu_1')\Sigma^{-1}x - \frac{1}{2}(\mu_0'\Sigma^{-1}\mu_0 + \mu_1'\Sigma^{-1}\mu_1 0) + \ln\big(P(L=0)\big) - \ln\big(P(L=1)\big) > 0$$

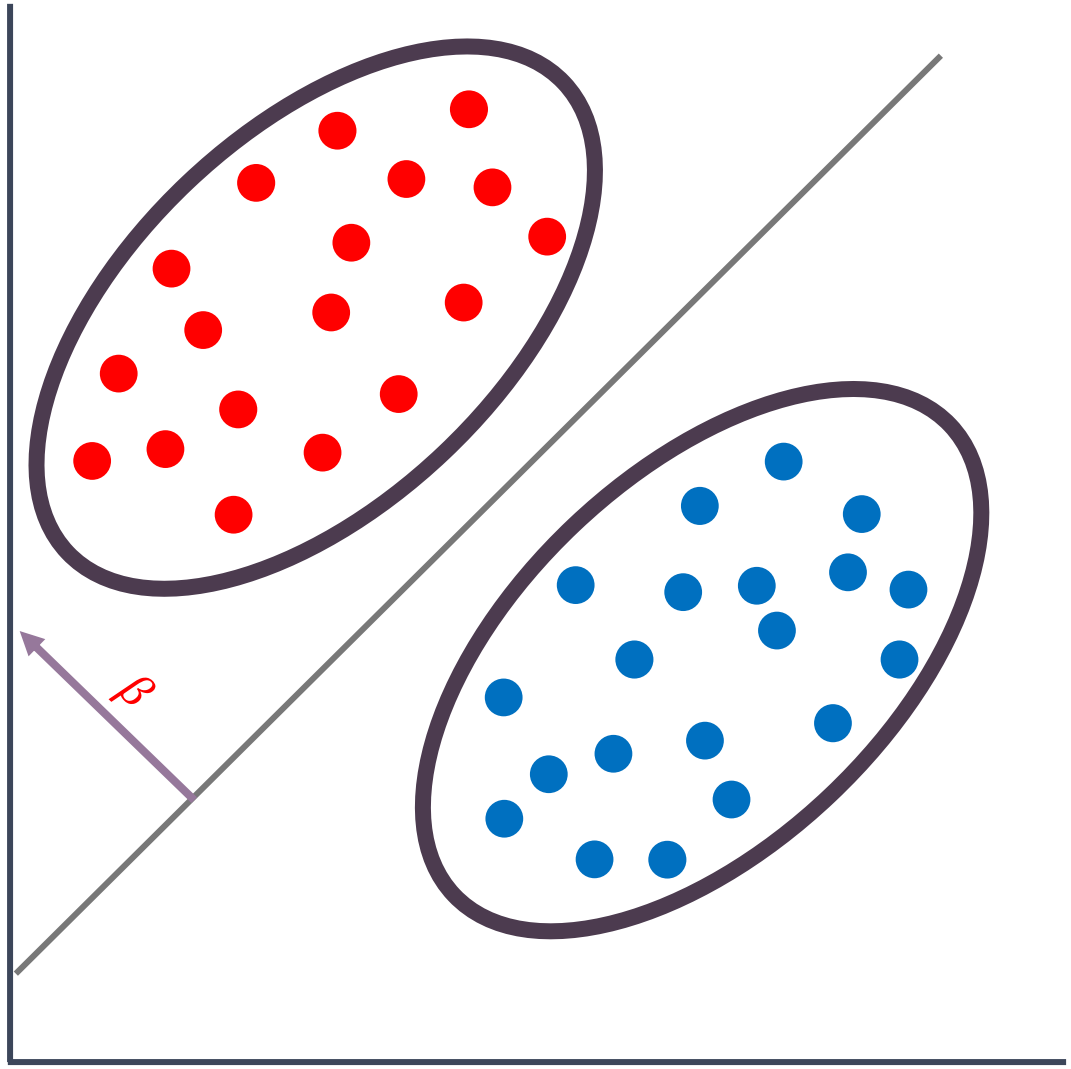
Es decir,

$$\beta' x + \beta_0 > 0 \quad \text{(clasificador lineal)}$$

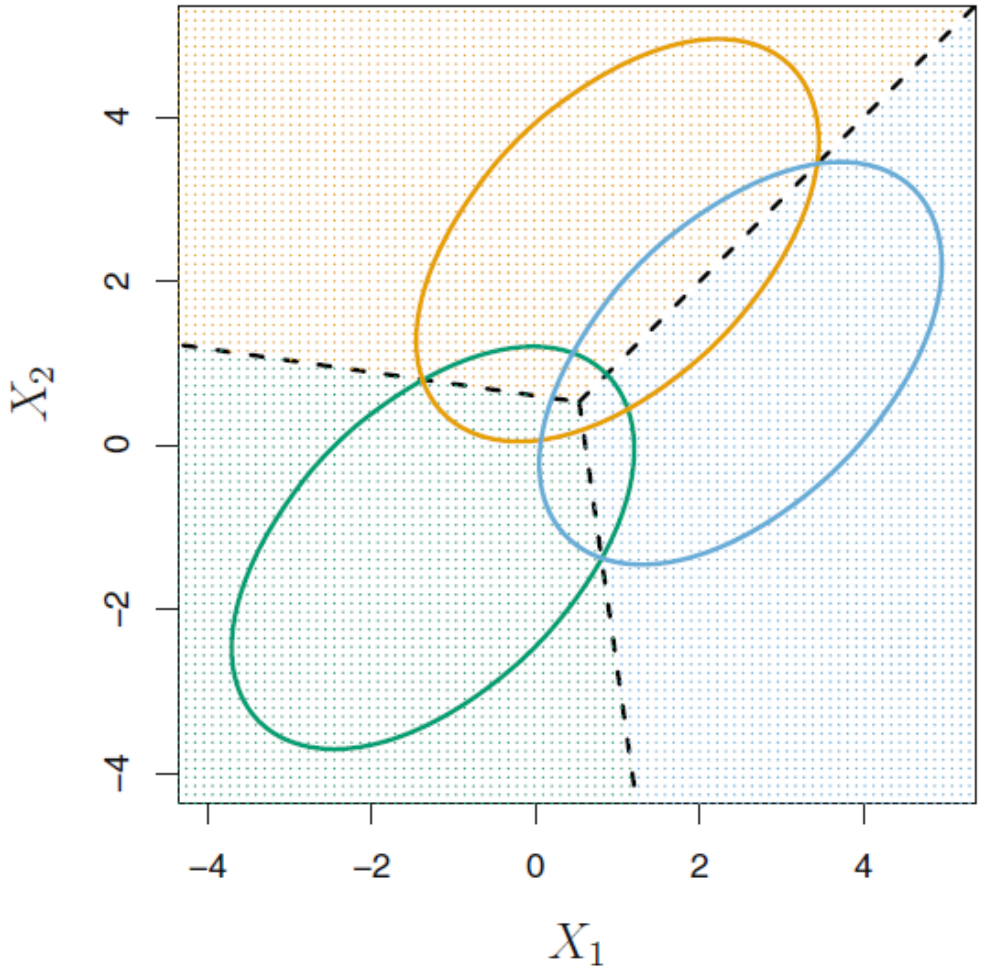
$$\beta = \Sigma^{-1}(\mu_0 - \mu_1)$$

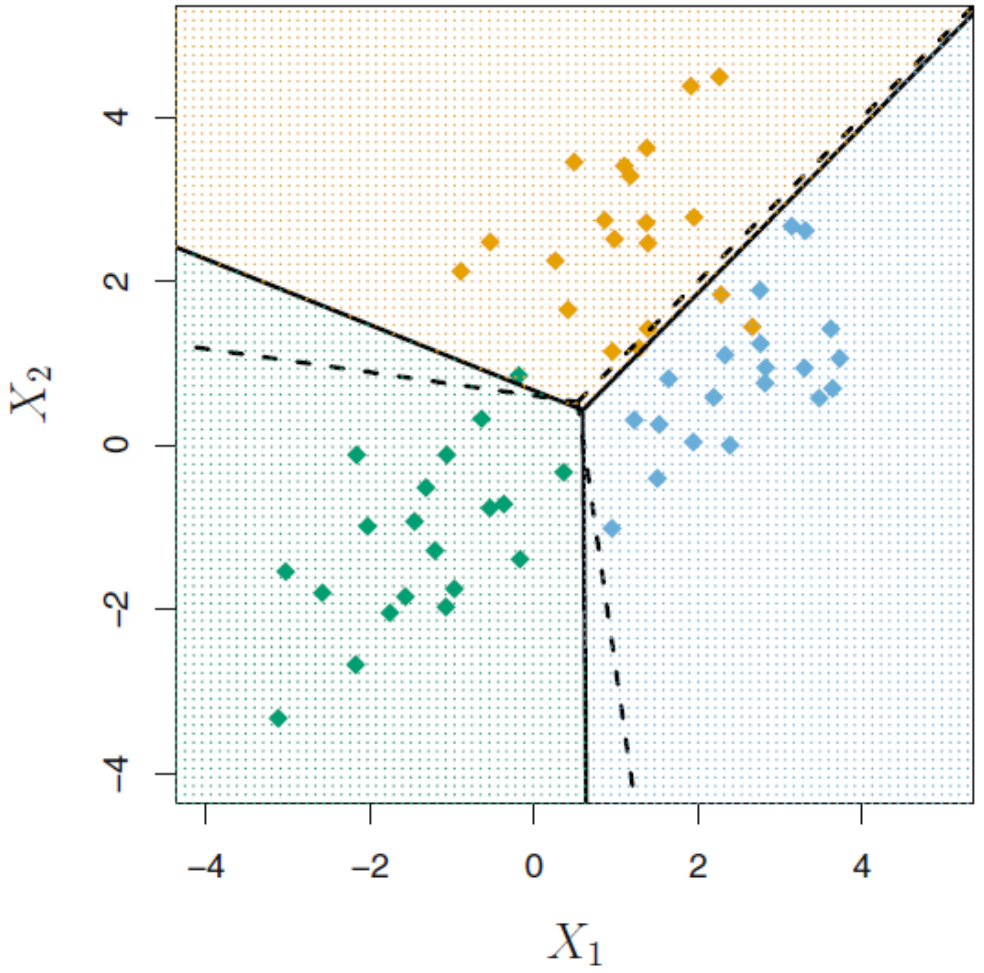
$$\beta_0 = \frac{1}{2}(\mu_1'\Sigma^{-1}\mu_1 - \mu_0'\Sigma^{-1}\mu_0) + \ln\big(P(L=0)\big) - \ln\big(P(L=1)\big)$$

Análisis discriminante lineal (LDA)



Análisis discriminante lineal multiclase





James, Witten, Hastie, & Tibshirani, 2023

Análisis discriminante cuadrático (QDA)

- QDA es un clasificador cuadrático, el cual se obtiene de manera similar a la formulación de LDA, pero sin considerar que $P(X = x|L = 0)$ y $P(X = x|L = 1)$ tienen la misma matriz de covarianza.

- Con ello, la regla de decisión queda como sigue:

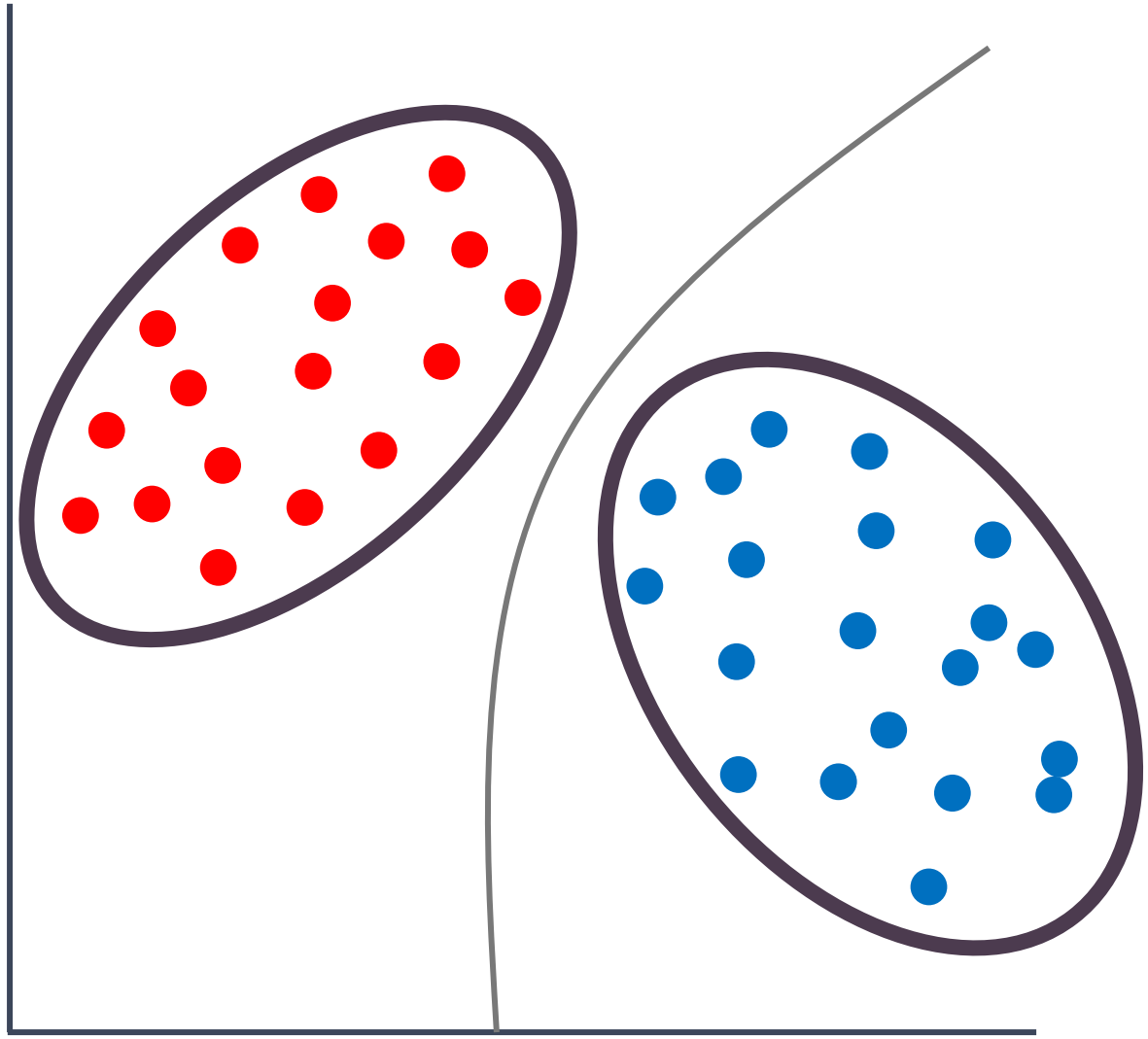
$$\frac{1}{2}x'Ax + \beta'x + \beta_0 > 0 \quad \text{(clasificador cuadrático)}$$

$$A = \Sigma_1^{-1} - \Sigma_0^{-1}$$

$$\beta = \Sigma_0^{-1}\mu_0 - \Sigma_1^{-1}\mu_1$$

$$\beta_0 = \frac{1}{2}\left(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_0'\Sigma_0^{-1}\mu_0 + \ln(|\Sigma_1|) - \ln(|\Sigma_0|)\right) + \ln\big(p(L = 0)\big) - \ln\big(p(L = 1)\big)$$

Análisis discriminante cuadrático (QDA)

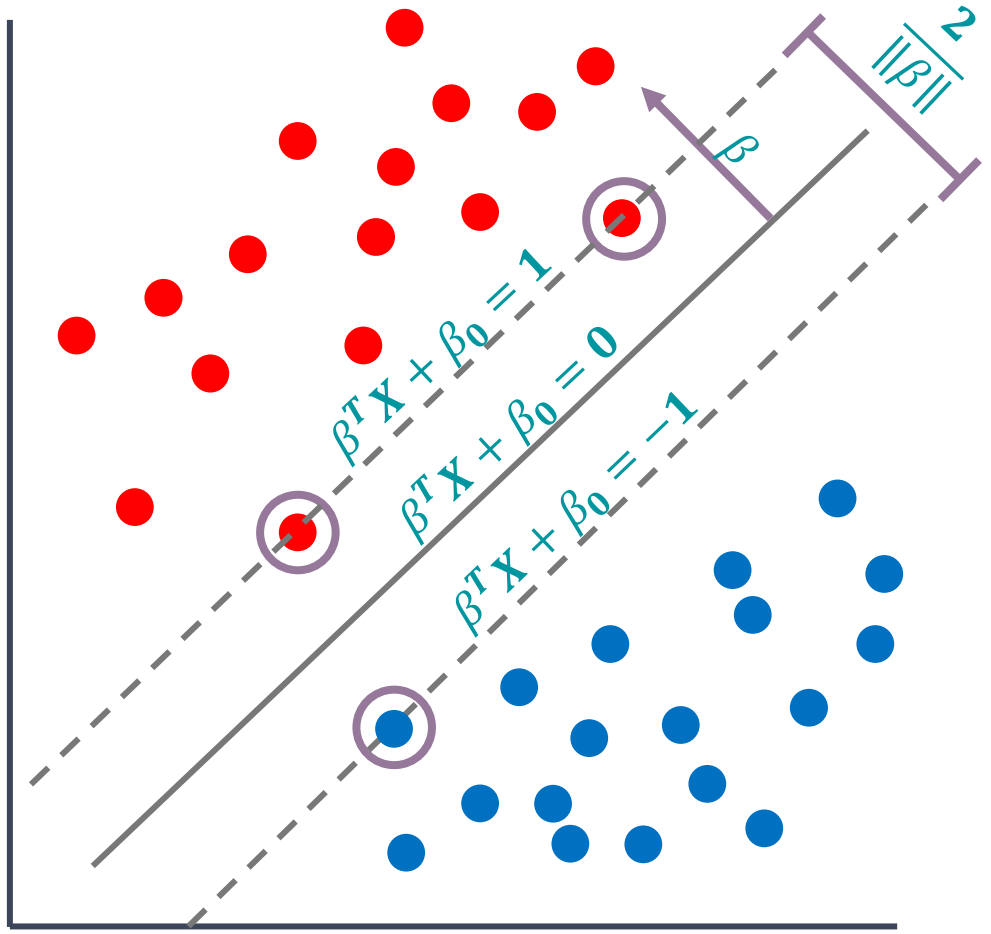


Máquinas de soporte vectorial (SVM)

- Este clasificador de dos clases encuentra el hiperplano separador que maximiza la distancia entre las observaciones más cercanas a dicho plano de ambas clases.

- Para un vector columna de características $X = [x_1, x_2, x_3, \dots, x_p]^T$, un vector de coeficientes $\beta = [\beta_1, \beta_2, \beta_3, \dots, \beta_p]^T$, y un intercepto β_0, este modelo es lineal de la forma $g(X) = \beta^T X + \beta_0$. Con ello, la regla de decisión está dada por:

$$f(x) = \begin{cases} 1 & \text{si } g(X) > 0 \\ -1 & \text{en otro caso} \end{cases}$$



A la distancia entre las observaciones de ambas clases cercanas al hiperplano separador se conoce como margen. La idea es maximizar el margen $\frac{2}{\|\beta\|}$ cuando los datos son linealmente separables (hard-margin).

Las observaciones que limitan al margen se conocen como vectores de soporte.

Maximizar el margen $\frac{2}{\|\beta\|}$ equivale a minimizar $\|\beta\|^2$, donde:

$$\|\beta\|^2 = \beta_1^2 + \beta_2^2 + \beta_3^2 + \cdots + \beta_p^2.$$

Es decir, el problema de maximizar el margen se plantea como el siguiente problema de minimización:

$$\arg\min_{\beta,\beta_0} \|\beta\|^2$$

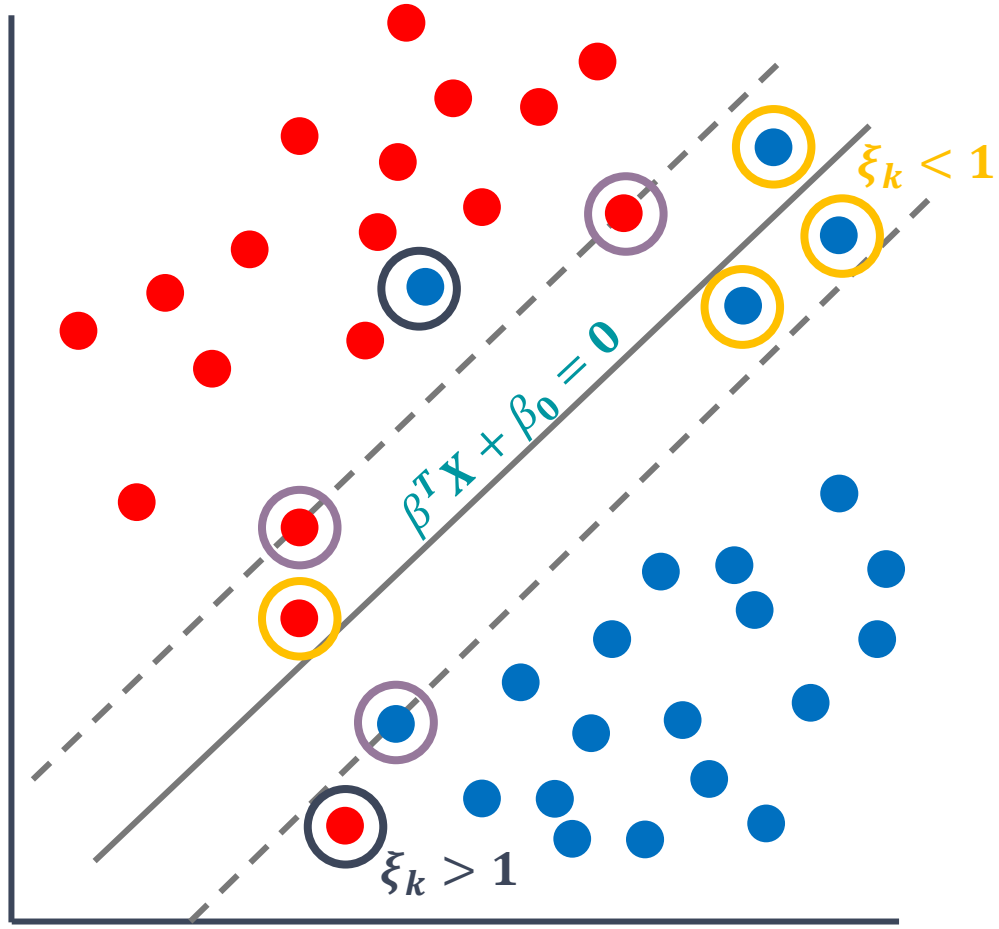
$$\text{sujeto a} \quad L_i(\beta^T X_i + \beta_0) \geq 1 \qquad i = 1,2,3,\ldots,n$$

Máquinas de soporte vectorial (SVM)

- Debido a que los datos usualmente no son linealmente separables, en la práctica se utiliza la versión soft-margin, en la cual se penaliza por cada elemento que se encuentre dentro del margen o esté mal clasificada.

- Para este caso, el problema de optimización queda de la siguiente forma:

$$\underset{\beta,\beta_0}{\arg\min} \ \|\beta\|^2 + C\sum_{k=1}^{N}\xi_k$$

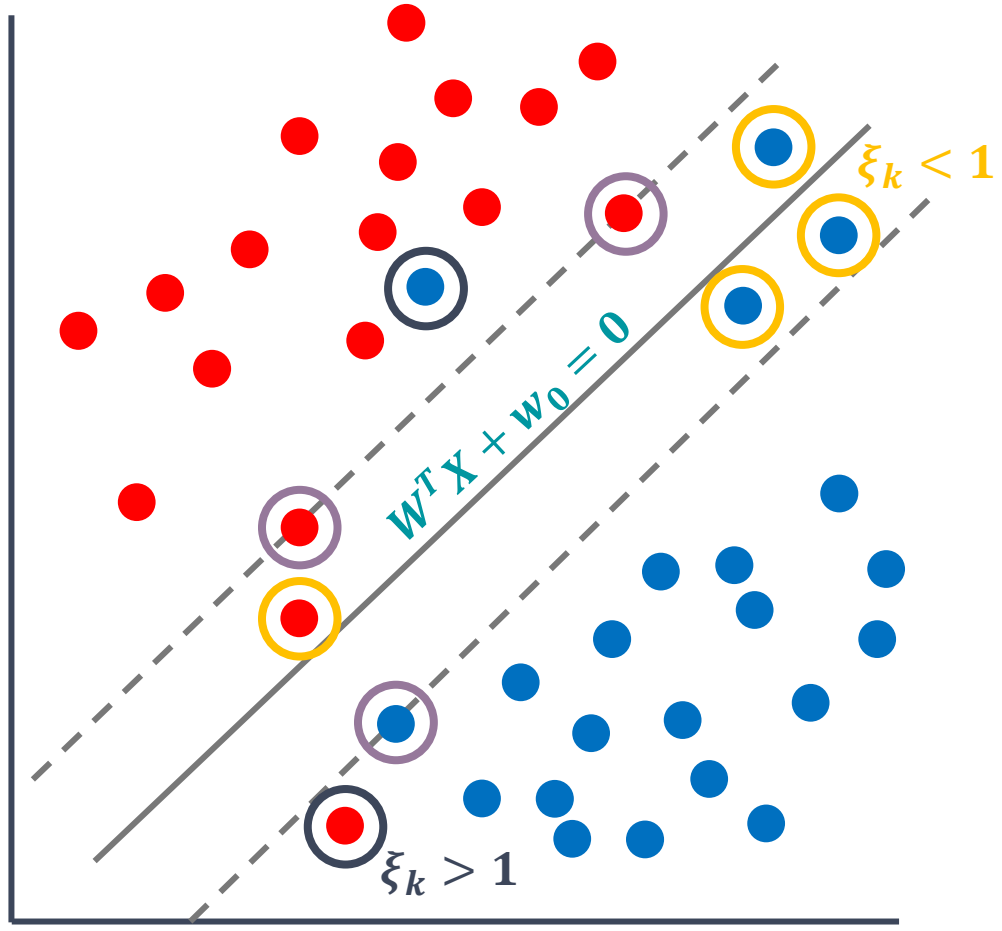


Máquinas de soporte vectorial (SVM)

- L_k es la etiqueta de la observación X_k, ξ_k indica qué tanto la observación x_k está violando la restricción de estar fuera del margen, y C es el parámetro que controla la relación del tamaño del margen (valor grande de C ocasiona un margen pequeño, y un valor grande de C conduce a un margen grande).

- Por otro lado

$$\xi_k = \max\left(0, 1 - L_k(\beta^T X_k + \beta_0)\right)$$



En la versión soft-margin, el problema de optimización a resolver queda de la siguiente manera:

$$\arg\min_{\beta,\beta_0} \|\beta\|^2 + C\sum_{k=1}^{N} \xi_k$$

$$\text{sujeto a} \quad L_i(\beta^T X_i + \beta_0) \geq 1 - \xi_i, \qquad \xi_i \geq 0$$
$$i = 1,2,3,\ldots,n$$

Formulación dual del SVM

- Se puede demostrar que:

$$\beta = \sum_{k=1}^{N} \alpha_k L_k X_k$$

donde $0 \leq \alpha_k \leq C$, y $\sum_{k=1}^{N} \alpha_k L_k = 0$.

- $\alpha_k = 0$ cuando X_k no es vector de soporte y está en el lado correcto en la clasificación, $0 < \alpha_k < C$ cuando X_k está dentro del margen (incluyendo las orillas), y $\alpha_k = C$ cuando X_k está fuera del margen y se clasifica incorrectamente.

Formulación dual del SVM

- Con ello

$$g(X) = \left(\sum_{k=1}^{N} \alpha_k L_k X_k \right)^T X + \beta_0 = \sum_{k=1}^{N} \alpha_k L_k X_k^T X + \beta_0$$

Y el problema a resolver queda como:

$$\arg\max_{\alpha} \sum_{k=1}^{N} \alpha_k - \frac{1}{2} \sum_{j=1}^{N} \sum_{k=1}^{N} \alpha_j \alpha_k L_j L_k X_j^T X_k$$

$$\text{sujeto a} \quad 0 \leq \alpha_k \leq C \; \forall \; k \quad \text{y} \quad \sum_{k=1}^{N} \alpha_k L_k = 0$$

La mayor ventaja de la formulación dual es que queda en términos de productos punto $X_j^T X_k$ entre observaciones. Dicho producto se puede remplazar por otra medida de similitud, la cual nos permite transformar los datos de manera indirecta sin tener que calcular dichas transformaciones.

A esto se le conoce como el truco del kernel. Los métodos basados en kernels mapean las observaciones originales a un espacio de mayor dimensión donde los datos se pueden separar con mayor facilidad.

Máquinas de soporte vectorial con kernels

- Al sustituir $X_j^T X_k$ por $k(X_j, X_k)$, queda la formulación dual de la siguiente manera:

$$g(X) = \sum_{k=1}^{N} \alpha_k L_k k(X_k, X) + w_0$$

$$\arg\max_{\alpha} \sum_{k=1}^{N} \alpha_k - \frac{1}{2} \sum_{j=1}^{N} \sum_{k=1}^{N} \alpha_j \alpha_k L_j L_k k(X_j, X_k)$$

$$\text{sujeto a} \quad 0 \leq \alpha_k \leq C \ \forall \ k \quad \text{y} \quad \sum_{k=1}^{N} \alpha_k L_k = 0$$

La función de kernel más utilizada es la de base radial:

$$k(X_i, X_j) = e^{-\gamma \|X_i - X_j\|_2^2}$$

donde γ es el parámetro que controla la dispersión del kernel.

Tipos de modelos de clasificación

Modelos de clasificación

- Modelos lineales. Análisis discriminante lineal (LDA), máquinas de soporte vectorial (SVM), clasificador de Fisher, clasificador bayesiano ingenuo lineal, clasificador logístico.

- Modelos cuadráticos. Análisis discriminante cuadrático (QDA), clasificador bayesiano ingenuo cuadrático.

- Modelos no lineales. k- vecinos más cercanos (k-NN), análisis discriminante generalizado, maquinas de soporte vectorial de base radial (RBSVM), ADA-boost, redes neuronales, árboles de decisión.

Modelos lineales

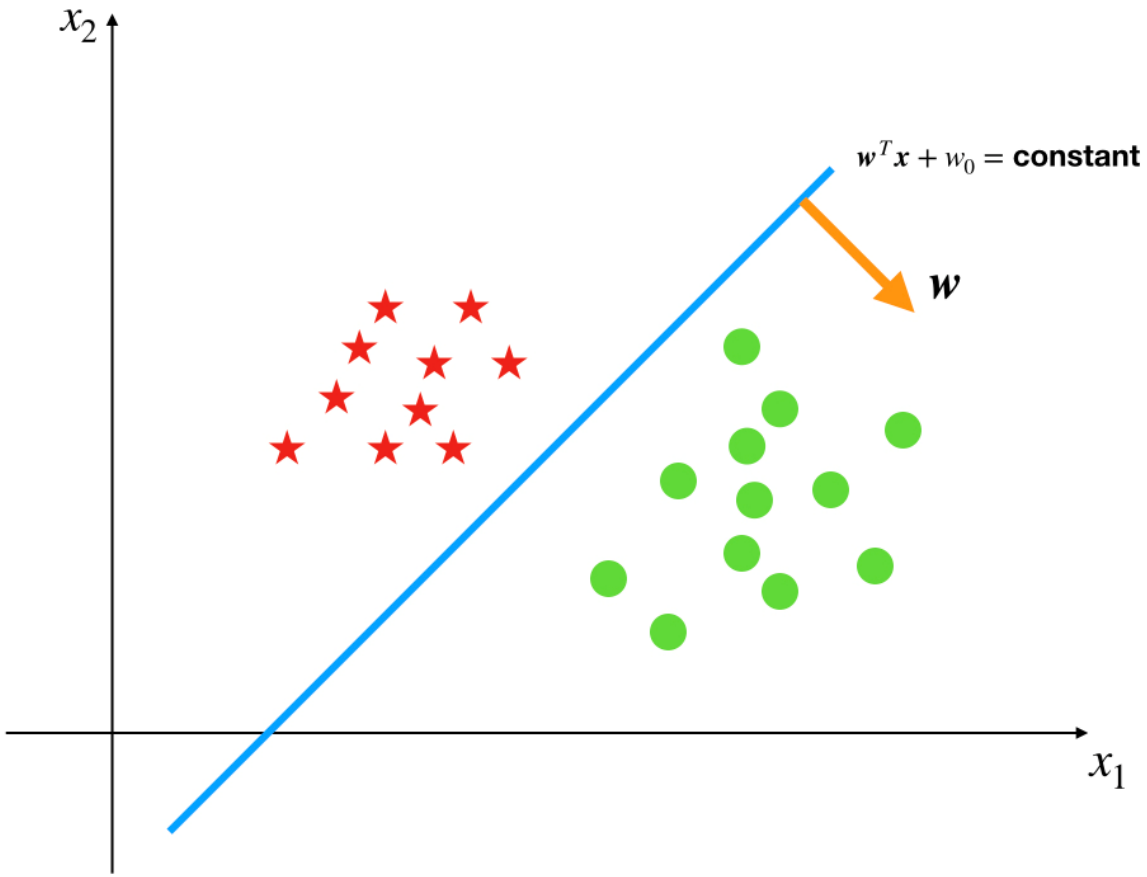
- El modelo de decisión se define en torno a una combinación lineal de las variables predictoras.
- Para el problema de dos clases, la forma general de un modelo de clasificación lineal está dada por la siguiente expresión:

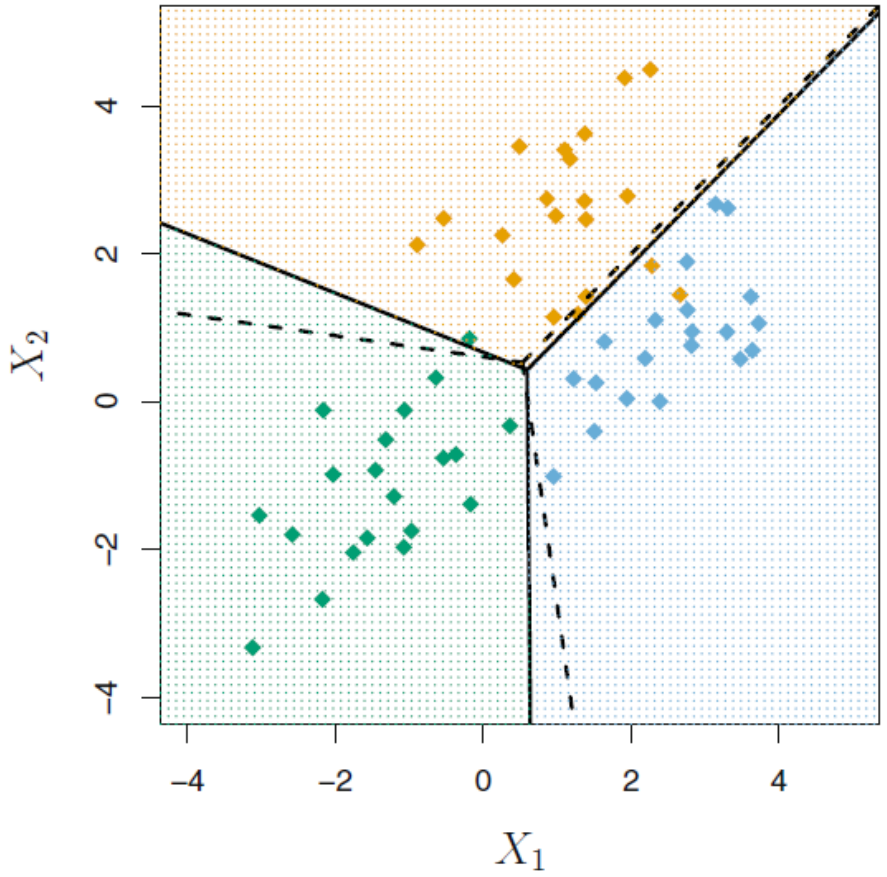
$$f(X;\beta) = \begin{cases} 1 & \text{if } g\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right) > Tr \\ 0 & \text{en otro caso} \end{cases}$$

donde Tr es un valor de umbral.

- La función $g(x)$ puede ser usada con una medida de confianza. Entre mayor sea su valor, hay más evidencia de que la clase sea 1.

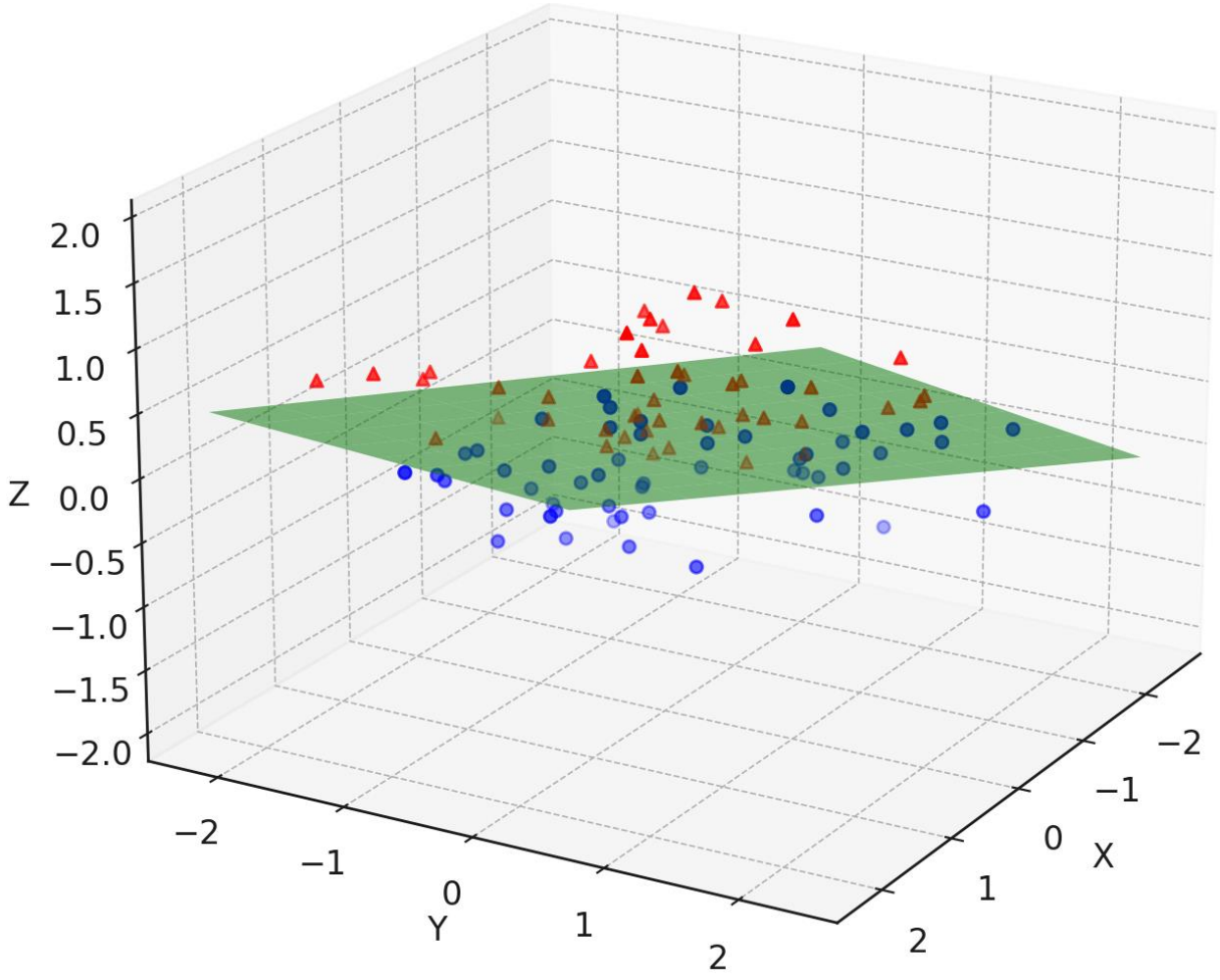
Modelos lineales





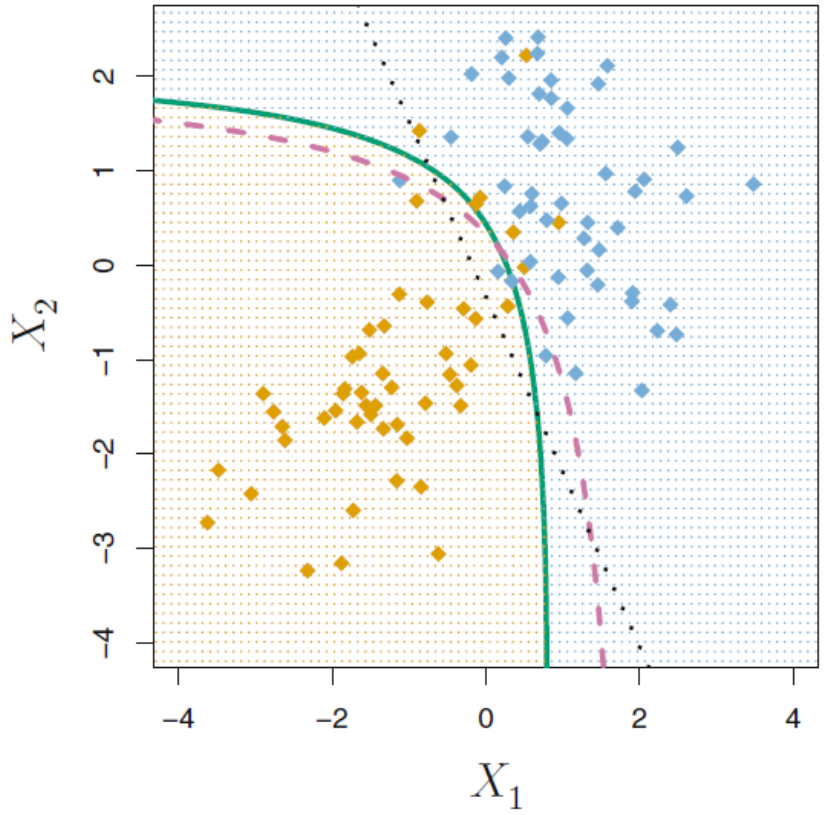
James, Witten, Hastie, & Tibshirani, 2023

Modelos lineales



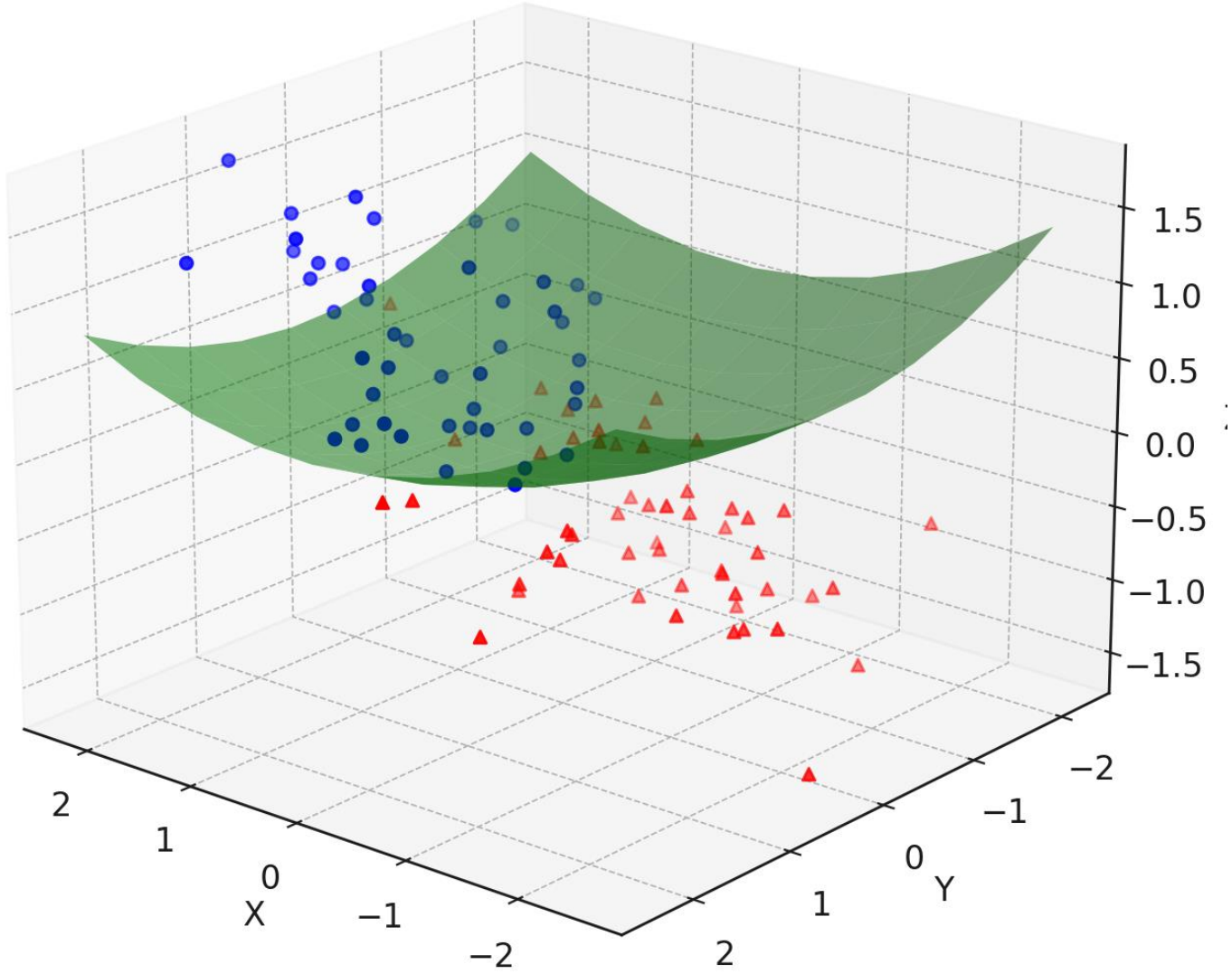
Modelos cuadráticos

- En estos clasificadores, el límite entre clases está dado por una función cuadrática.



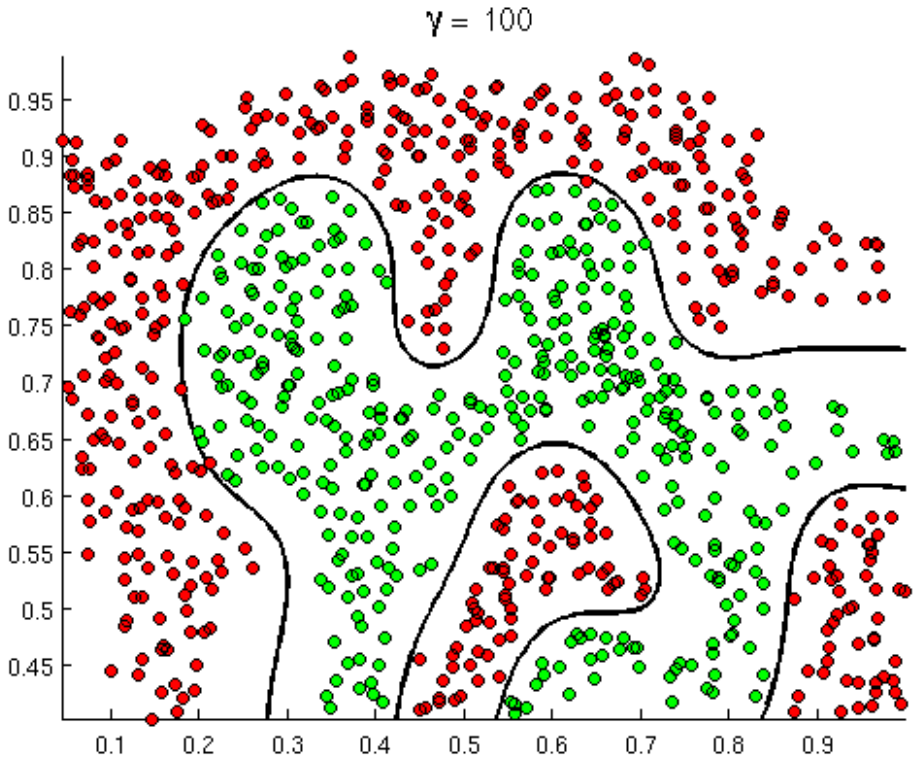
James, Witten, Hastie, & Tibshirani, 2023

Modelos cuadráticos



Clasificadores no lineales

- Los límites de decisión no son funciones lineales ni cuadráticas.



Open classroom – Non-linear SVM classification with kernels

Modelos de clasificación (geométrico vs lógico vs probabilístico vs ensamble)

- Modelos lógicos. Basados en secuencias de reglas lógicas AND-OR. Árboles de decisión, clasificadores basados en reglas lógicas.

- Modelos geométricos. Separan el espacio en regiones para cada clase, o comparan distancias entre observaciones de cada clase con el elemento a clasificar. Máquinas de soporte vectorial (SVM), Perceptrón, K-NN.

Modelos de clasificación (geométrico vs lógico vs probabilístico vs ensamble)

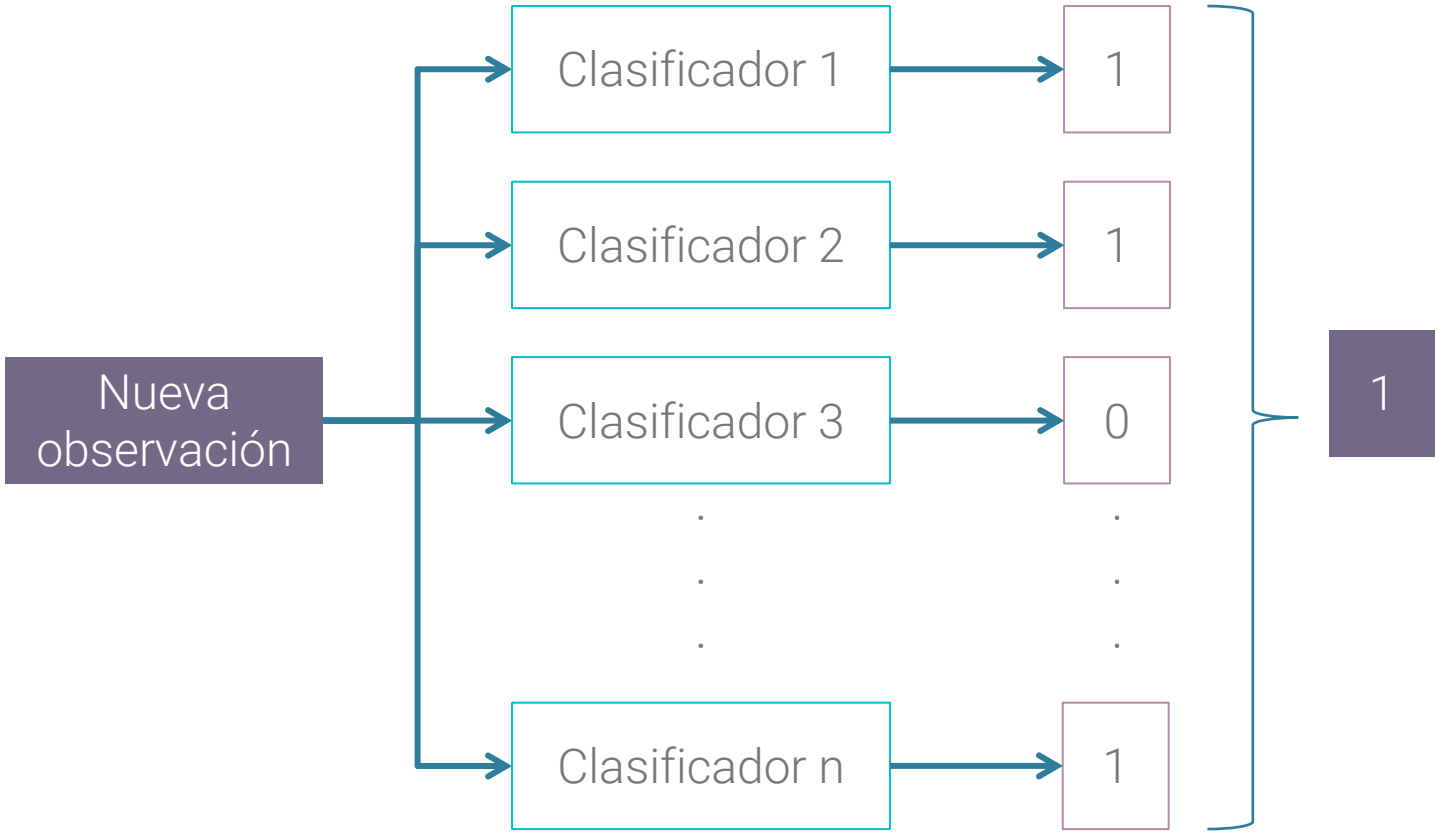
- Modelos probabilísticos. Calculan las probabilidades de que una observación pertenezca a cada clase. Naïve Bayes, Clasificador logístico, análisis discriminante lineal (LDA), análisis discriminante cuadrático (QDA).

- Modelos de ensamble. Combinan varios clasificadores para generar reglas con mejor poder predictivo o reducir errores y sesgos. Voting, Bagging, Boosting, Stacking.

Métodos de ensamble (voting)

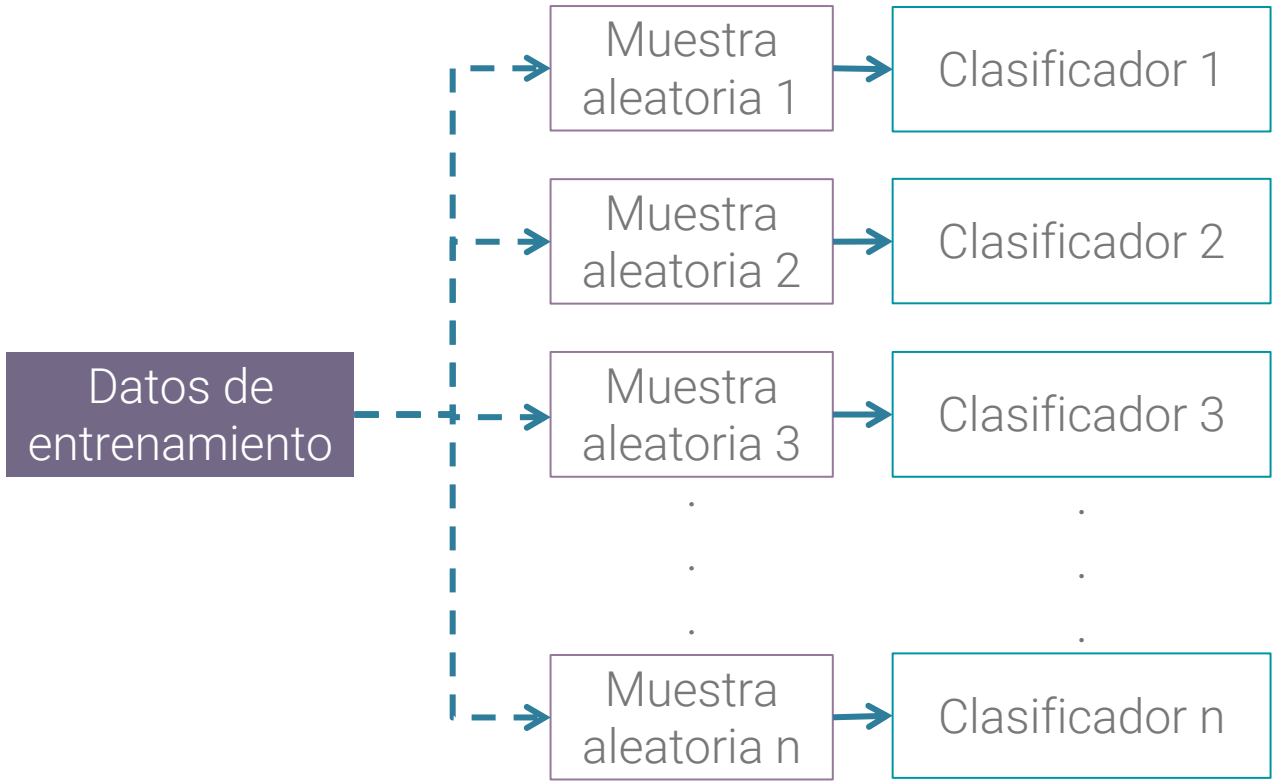
- En Voting se entrenan varios modelos de diferente tipo.

- Al momento de clasificar una nueva observación, se hace un consenso sobre los resultados obtenidos con los clasificadores.



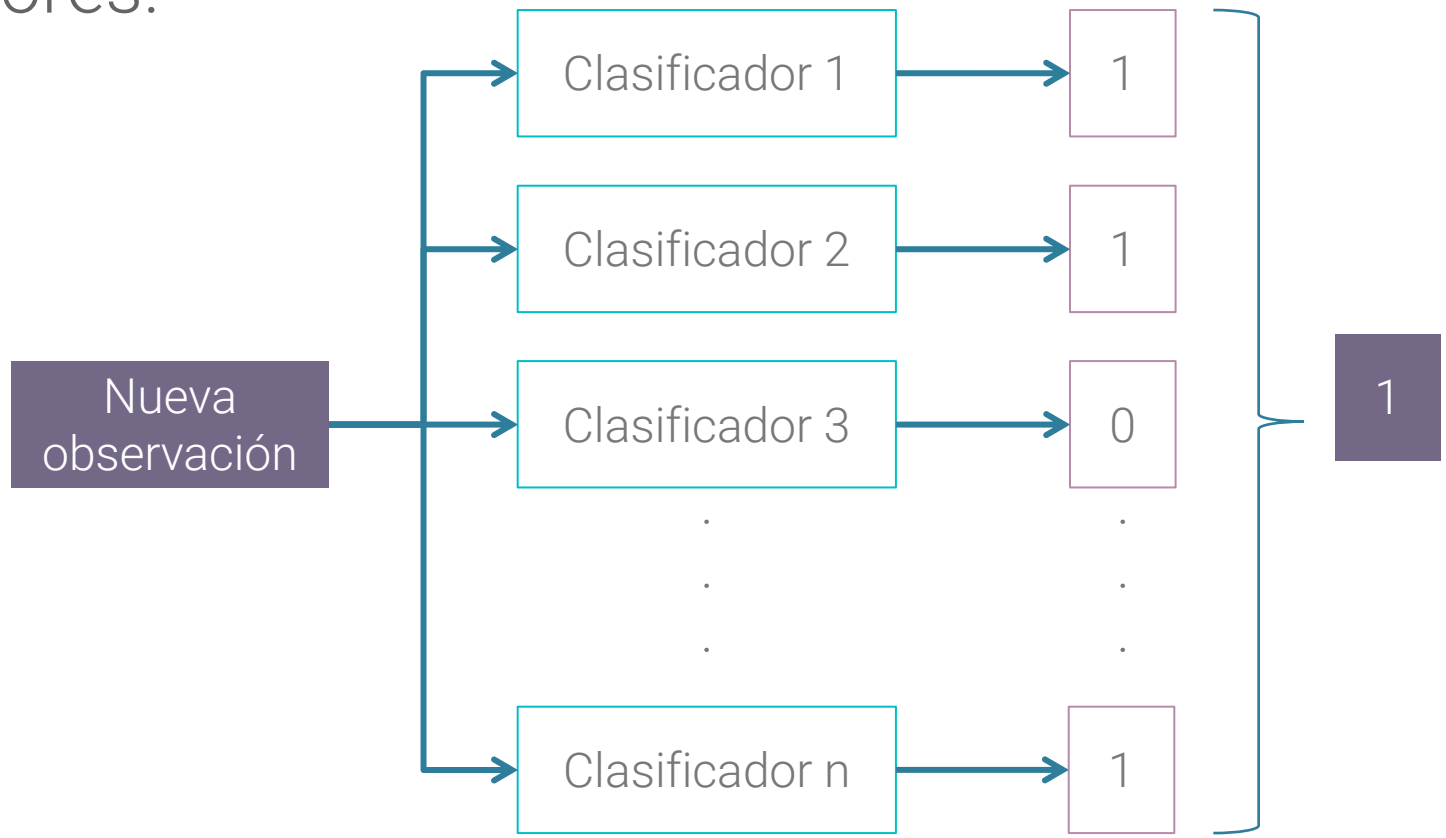
Métodos de ensamble (bagging)

- En Bagging se muestrea con remplazo el mismo conjunto de datos de entrenamiento para generar nuevos conjuntos de entrenamiento, con los que se ajustan varias modelos del mismo tipo.



Métodos de ensamble (bagging)

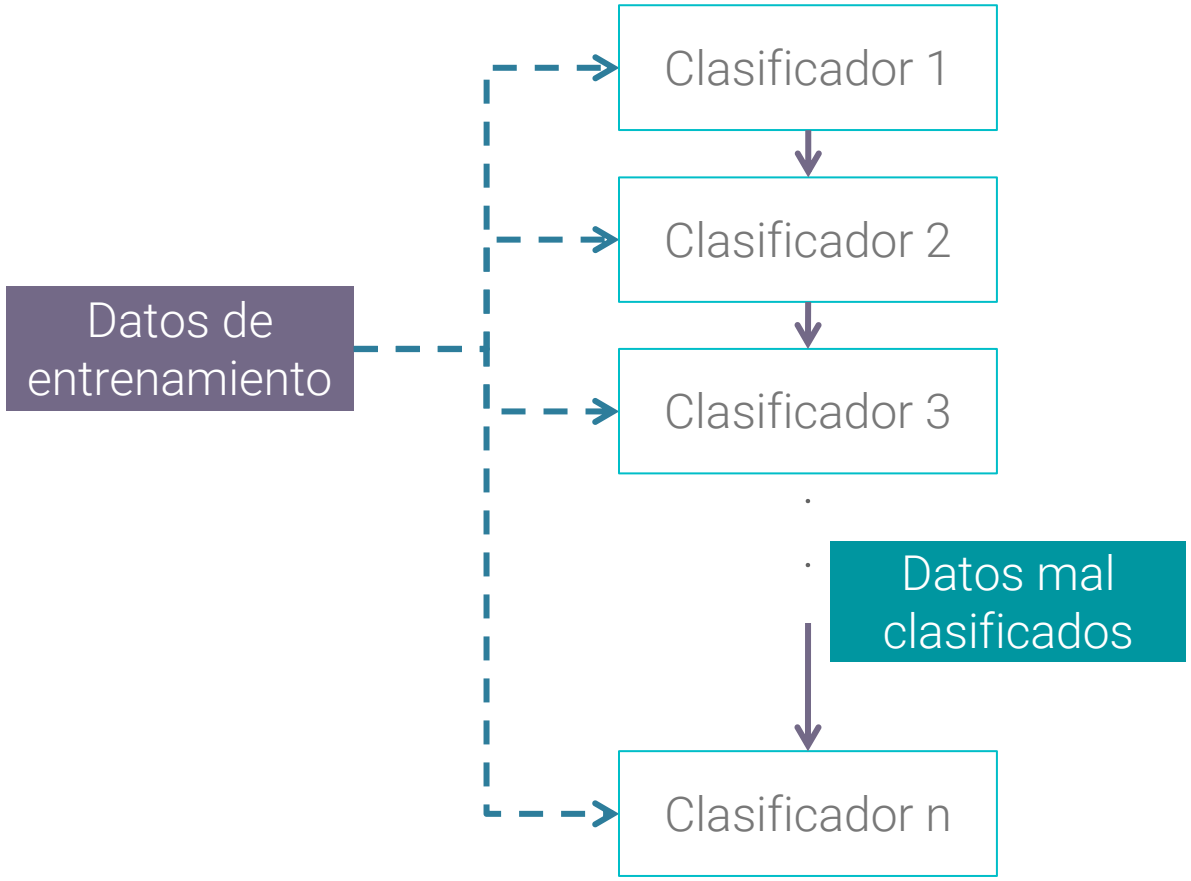
- Al momento de clasificar una nueva observación, se hace un consenso sobre los resultados obtenidos con los diferentes clasificadores.



Métodos de ensamble (boosting)

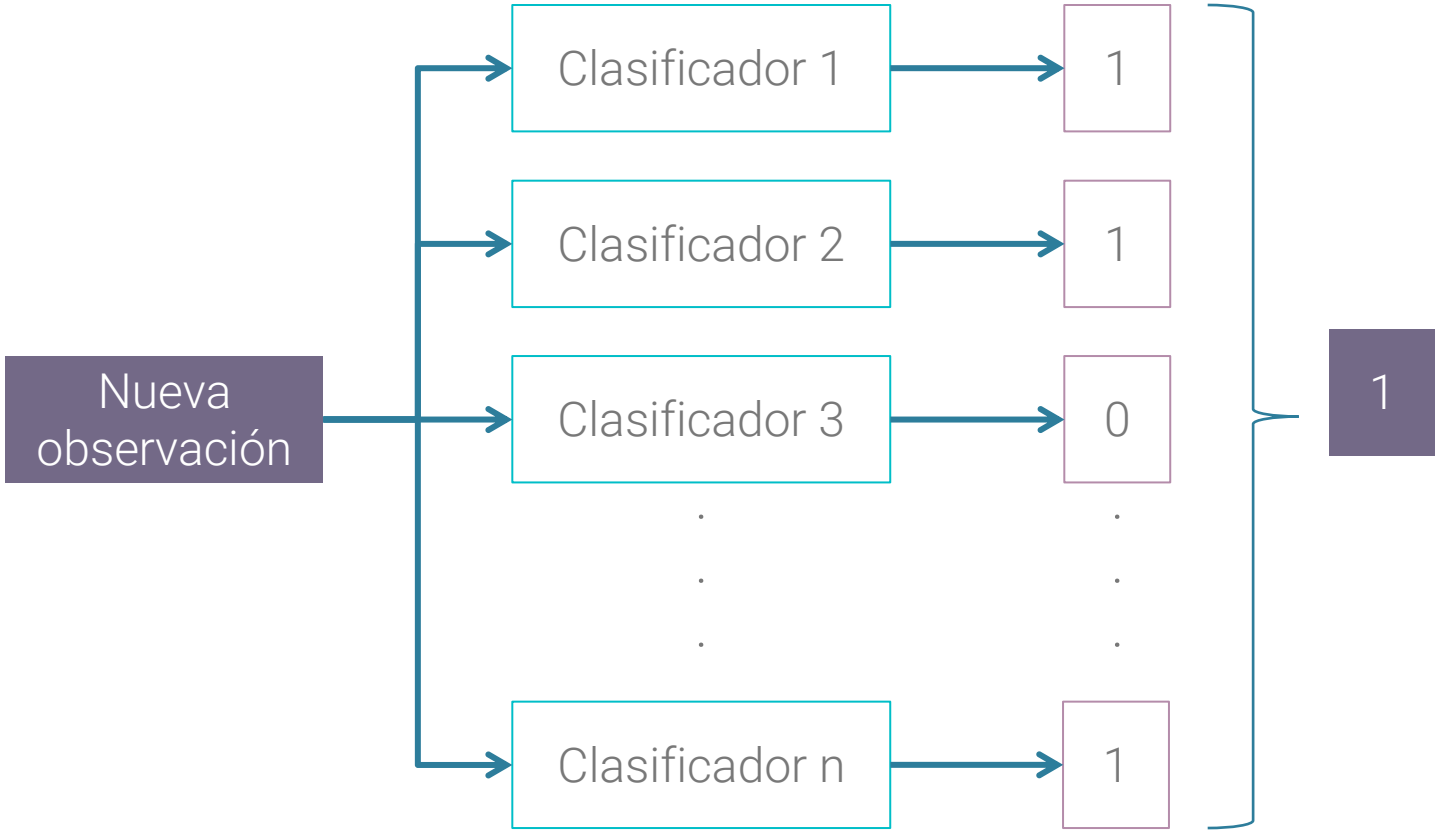
- En Boosting, se entrena de manera secuencial clasificadores “débiles”, de tal manera que los errores de uno se intentan compensar en los siguientes modelos.

- Esto se logra con identificar las muestras mal clasificadas por un clasificador, y darles mas importancia al momento de entrenar el siguiente modelo.



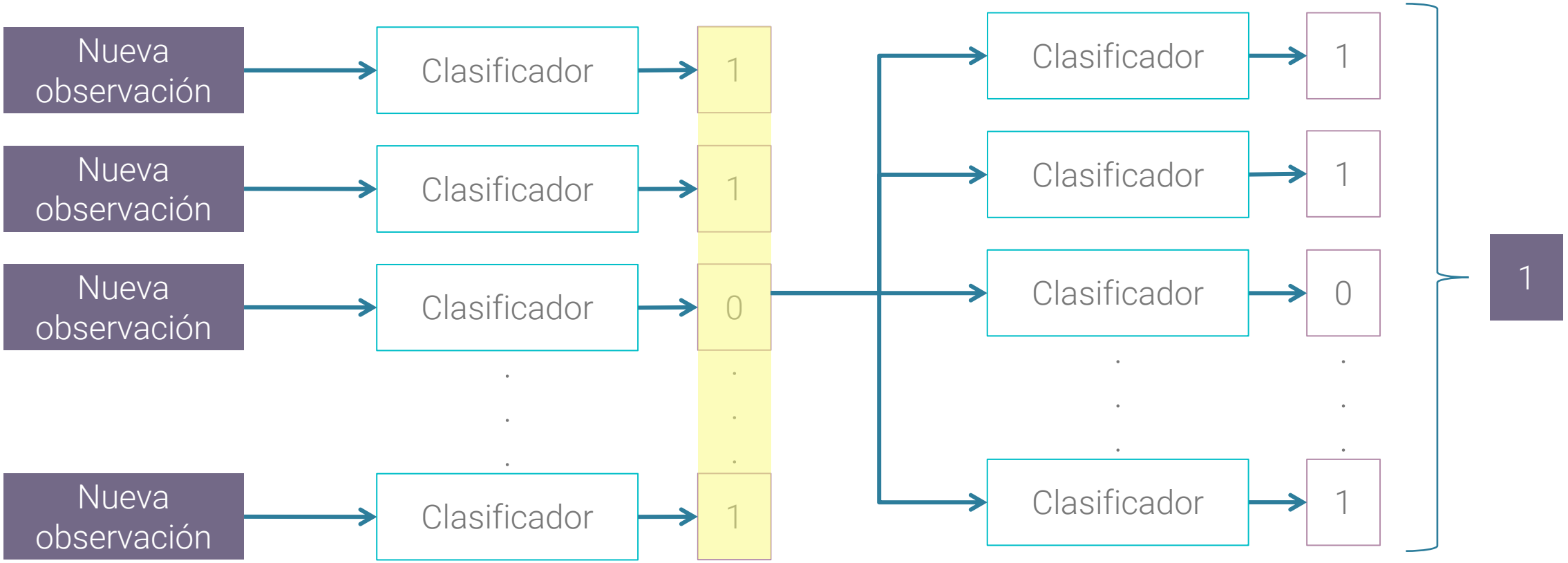
Métodos de ensamble (boosting)

- Para clasificar, de nuevo se utiliza el consenso (votación) entre el conjunto de clasificadores entrenado.



Métodos de ensamble de modelos

- En Stacking, la salida de un conjunto de clasificadores es la entrada de otro conjunto.



Clasificación multiclase

Estrategias de clasificación multiclase

- Modelos multiclase. Clasificadores que de manera natural pueden resolver problemas multiclase. k-NN, LDA, redes neuronales, árboles de decisión.

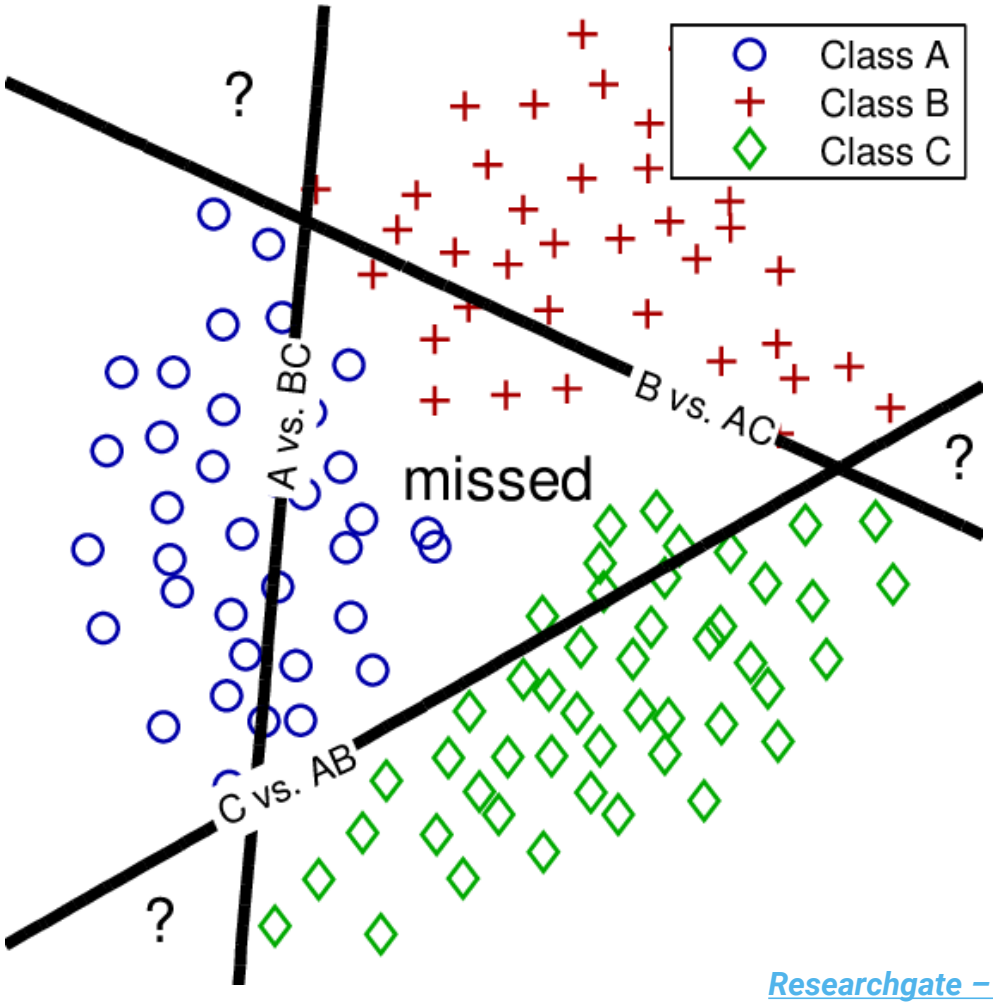
- Extensión de modelos de dos clases. Se entrenan clasificadores de dos clases y se hace uso de un esquema de votación para determinar la clase se una nueva observación.
  - Estrategia uno contra el resto.
  - Estrategia uno contra uno.

Estrategia uno contra el resto

- En la estrategia uno contra el resto, se entrenan C clasificadores (donde C es la cantidad de clases), de tal forma que cada clasificador compara las observaciones de una clase contra el resto de las observaciones.

- Para determinar la clase de una nueva observación, se evalúan todos los clasificadores. Si sólo uno detecta de manera positiva la clase de la observación, y el resto indica que pertenecen a la clase “resto”, entonces la clase es la del clasificador que dio positivo.

Estrategia uno contra el resto



Researchgate – One-versus-one ensamble

Estrategia uno contra uno

- En la reducción uno contra uno, se entrenan $C(C-1)/2$ clasificadores binarios para cada pareja posible de clases ($C_1\ Vs\ C_2$, $C_1\ Vs\ C_3$, $C_2\ Vs\ C_3$, etc.).

- Para predecir la etiqueta de una nueva observación, se evalúan todos los clasificadores, y la clase que reciba más votos, es la que se le asigna a la observación.

¿Qué modelo debería utilizar?

Al igual que en regresión no existe una respuesta única a la pregunta sobre qué modelo es mejor de manera general.

Para encontrar un modelo adecuado, sólo se puede probar diferentes alternativas para el conjunto de datos dado utilizando las metodologías de validación y prueba apropiados.

Es una buena práctica probar modelos lineales y no lineales. A su vez, es buena idea tomar modelos geométricos, lógicos, probabilísticos y de ensamble para las pruebas.

Bibliografía

- James, G., Witten, D., Hastie, T. & Tibshirani, R. (2023). *An introduction to statistical learning: with applications in Python* (2da ed.). Springer.
  - Capítulo 4

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2da ed.). Springer.
  - Capítulo 4

- Russell, S. J. & Norvig, P. (2021). *Artificial intelligence: A modern approach (global edition)* (4ta ed.). Person.
  - Capítulo 19