Universidad del Rosario

Analisis Avanzado de Datos

W12. Modelos Generalizados, regression logistica

FERNEY ALBERTO BELTRAN MOLINA
Escuela de Ingeniería, Ciencia y Tecnología
Matemáticas Aplicadas y Ciencias de la Computación

Flexibilidad, complejidad y sobreajuste

- En cualquier proyecto de ciencia de datos, asumimos que nuestros datos siguen un **patrón** y cualquier desviación de este patrón se considera **ruido**.
- Nuestros modelos deben ser lo suficientemente flexibles para **capturar la complejidad del patrón subyacente**, pero no demasiado flexibles para no memorizar detalles de **ruido irrelevantes** en nuestros datos.
- Una **metodología rigurosa** es crucial para evitar caer en trampas comunes, como el sobreajuste.

Si la metodología no la conoceremos estamos fallando, no importa el modelo.
No glorifiquemos los datos sino como se hay recogido eso datos

La encuesta Literary Digest de 1936

- The Literary Digest realizó una de las encuestas más grandes jamás hechas. **10 millones de personas.**
- Predijo que Landon obtendría el 57% de los votos y Roosevelt el 43%. Sin embargo, Landon terminó obteniendo el 38% de los votos y Roosevelt el 62%.
- ¿Qué sucedió? **Un muestreo deficiente**. Los nombres se tomaron de directorios telefónicos, listas de miembros de clubes, listas de suscriptores de revistas, etc.
- Las muestras **no fueron representativas** de la población.

Know your data!

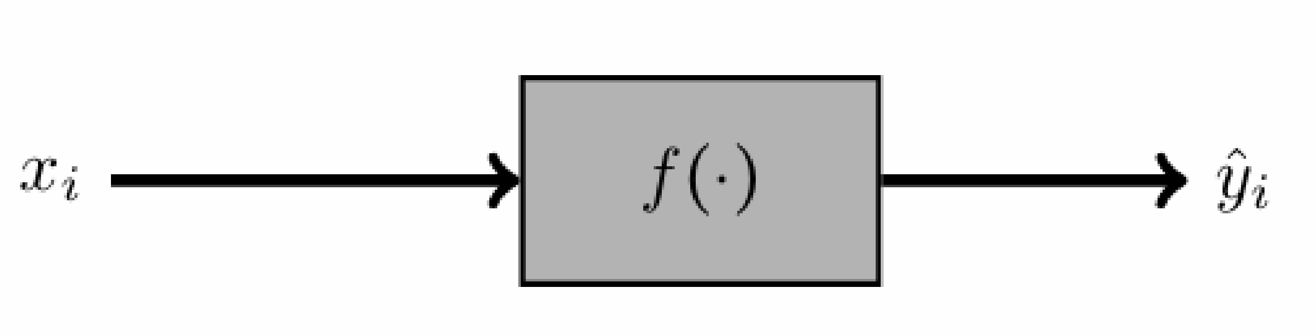
Formulación de problemas de clasificación

En **clasificación** (también conocida como **decisión** o **detección**):

- Tenemos un modelo que produce una **etiqueta** cuando se le muestran un conjunto de predictores.
- La etiqueta es discreta y sus valores se llaman clases.

En un **problema de clasificación:**

- Tenemos una noción de **calidad del modelo**
- Construimos un modelo $\hat{y} = f(x)$ utilizando dataset {(xi, yi) : 1 ≤ i ≤ N}.
- El par (xi, yi) puede interpretarse como “la muestra xi pertenece a la clase yi”, o “la etiqueta de la muestra xi es yi”.



En regresión Y es cuantitativa
En clasificación Y es cualitativa (categórica)

- Dado un objeto caracterizado por un vector de descriptores, el **objetivo de un clasificador** (modelo) es predecir a qué clase pertenece el objeto dentro de un conjunto de clases predefinidas.

Problemas de clasificación

El **análisis de sentimientos** permite identificar las opiniones humanas expresadas en fragmentos de texto. Se pueden considerar múltiples opiniones, pero en su forma más simple, se definen dos, positiva y negativa.

El dataset “Large Movie Review Dataset” se creó para construir modelos que reconozcan sentimientos polarizados en fragmentos de texto:

- Contiene 25000 muestras para entrenamiento y 25000 muestras para pruebas.
- Cada instancia consta de un fragmento de texto utilizado como predictor y una etiqueta binaria (0 siendo opinión negativa y 1 opinión positiva).
- Puedes descargarlo desde http://ai.stanford.edu/~amaas/data/sentiment/

Un problema de clasificación multiclase

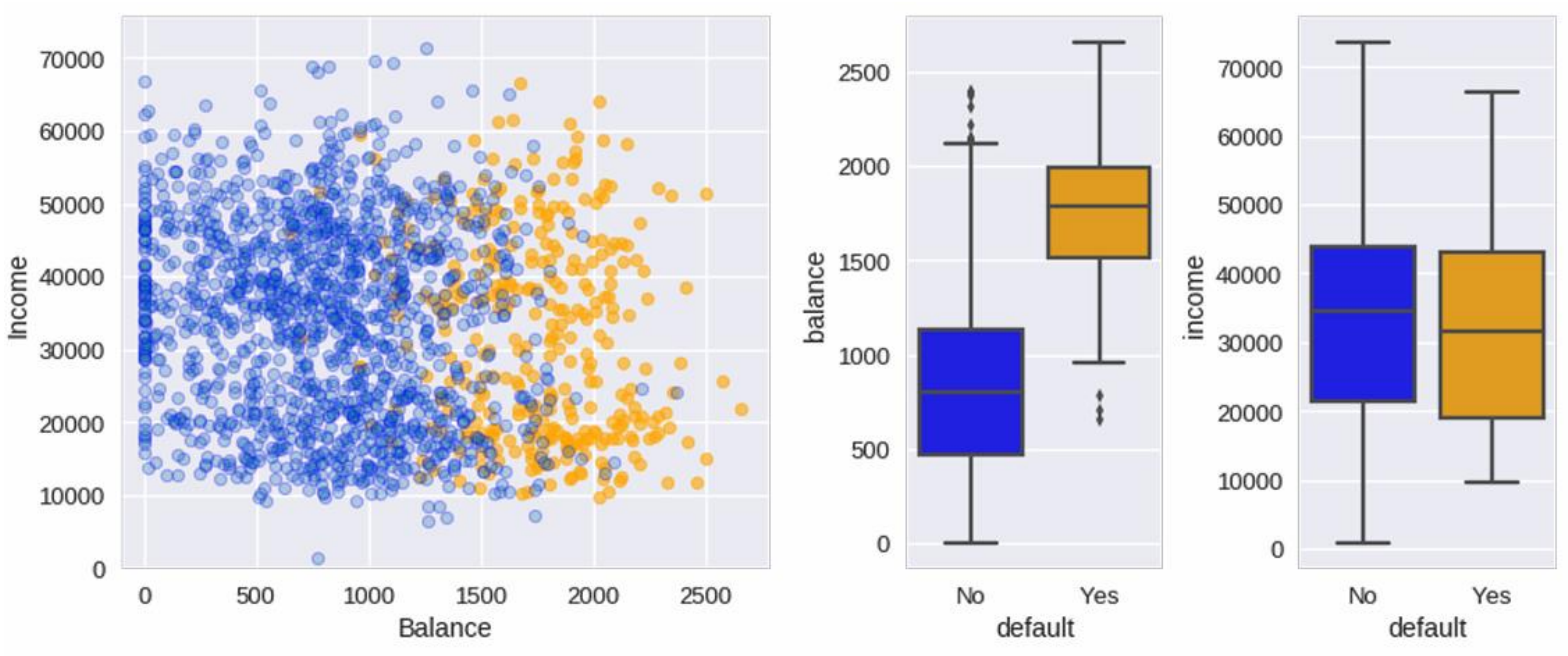
Reconocer dígitos en imágenes que contienen representaciones escritas a mano es un clásico problema de clasificación multiclase. El predictor es una matriz de valores (imagen) y hay 10 clases, es decir, 0 al 9.

En el aprendizaje automático, utilizamos conjuntos de datos de **imágenes etiquetadas**, es decir, pares de imágenes (**predictores**) y valores numéricos (**etiquetas**), para construir los modelos.

El conjunto de datos MNIST es una colección de dígitos escritos a mano:

- 60,000 imágenes para entrenamiento, 10,000 para pruebas.
- Las imágenes son en blanco y negro, 28x28 píxeles.
- Descargable desde: yann.lecun.com/exdb/mnist

Un problema de clasificación



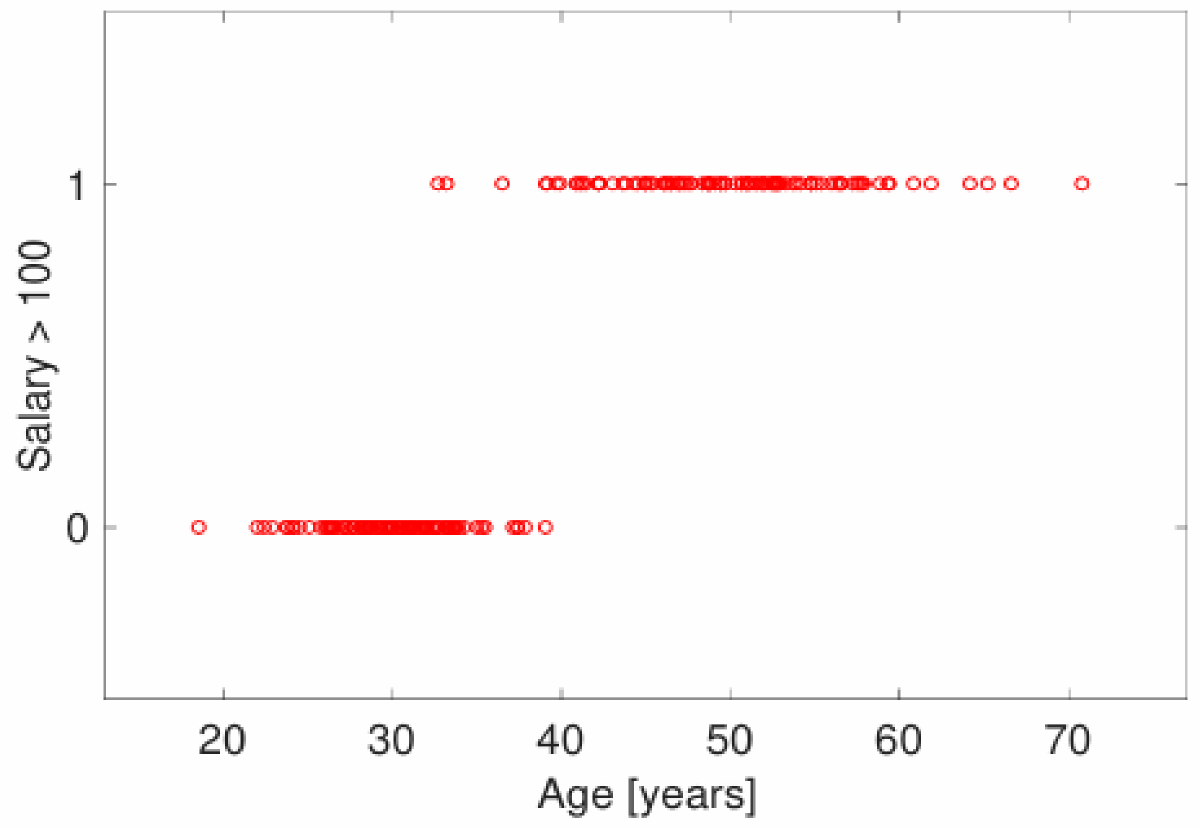
Impago de la tarjeta de crédito

- X1 = Income: ingresos anuales de 10,000 individuos.
- X2 = Balance: saldo (de la deuda) de la tarjeta de crédito.
- Y= Default: impago de la tarjeta de crédito {Yes, No}

El conjunto de datos en el espacio de atributos

Las etiquetas pueden ser representadas por valores numéricos en un eje vertical. Ten cuidado: las nociones habituales de orden y distancia no se aplican a variables categóricas.

Un predictor, dos clases



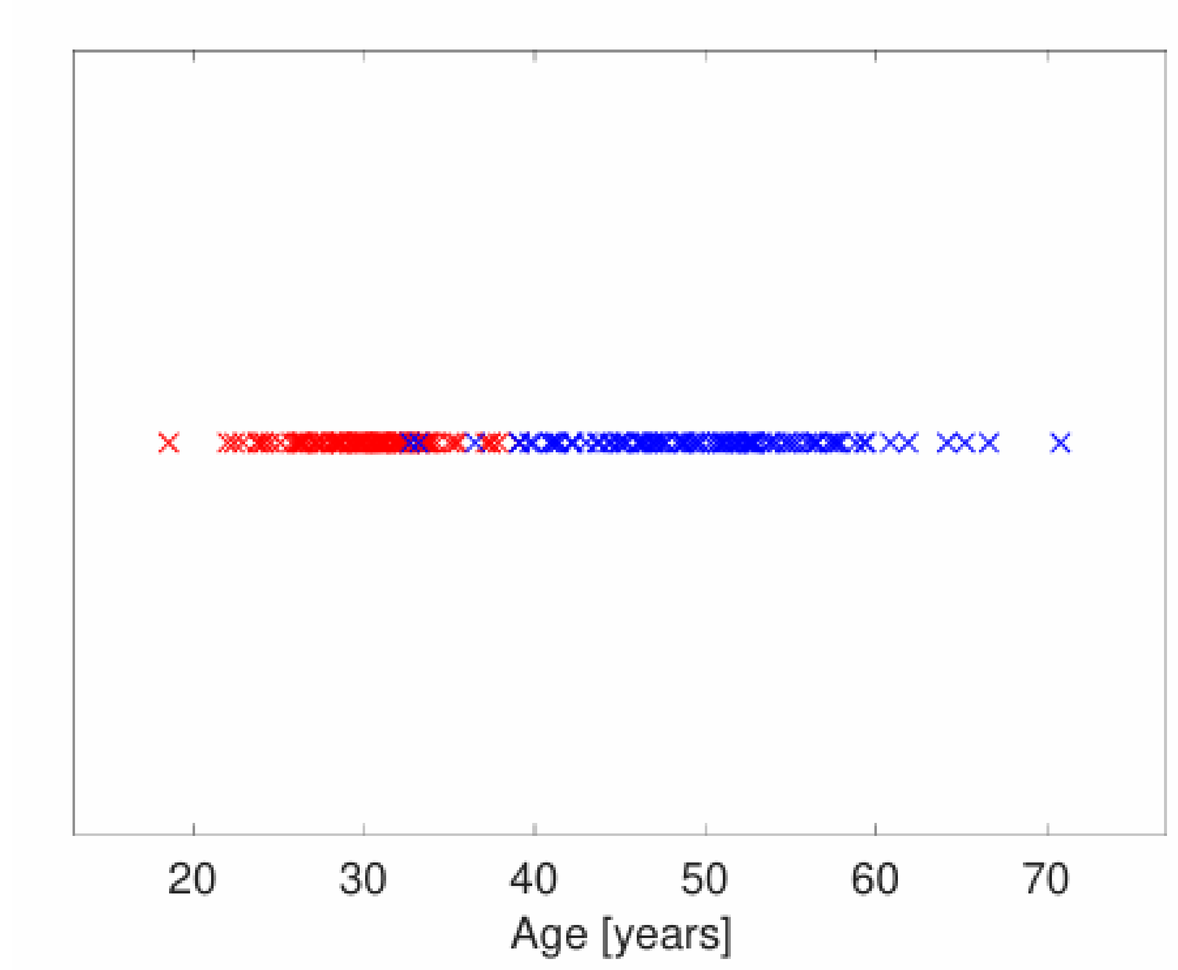
dos predictores, tres clases

No es la mejor manera de representar los datos en el espacio

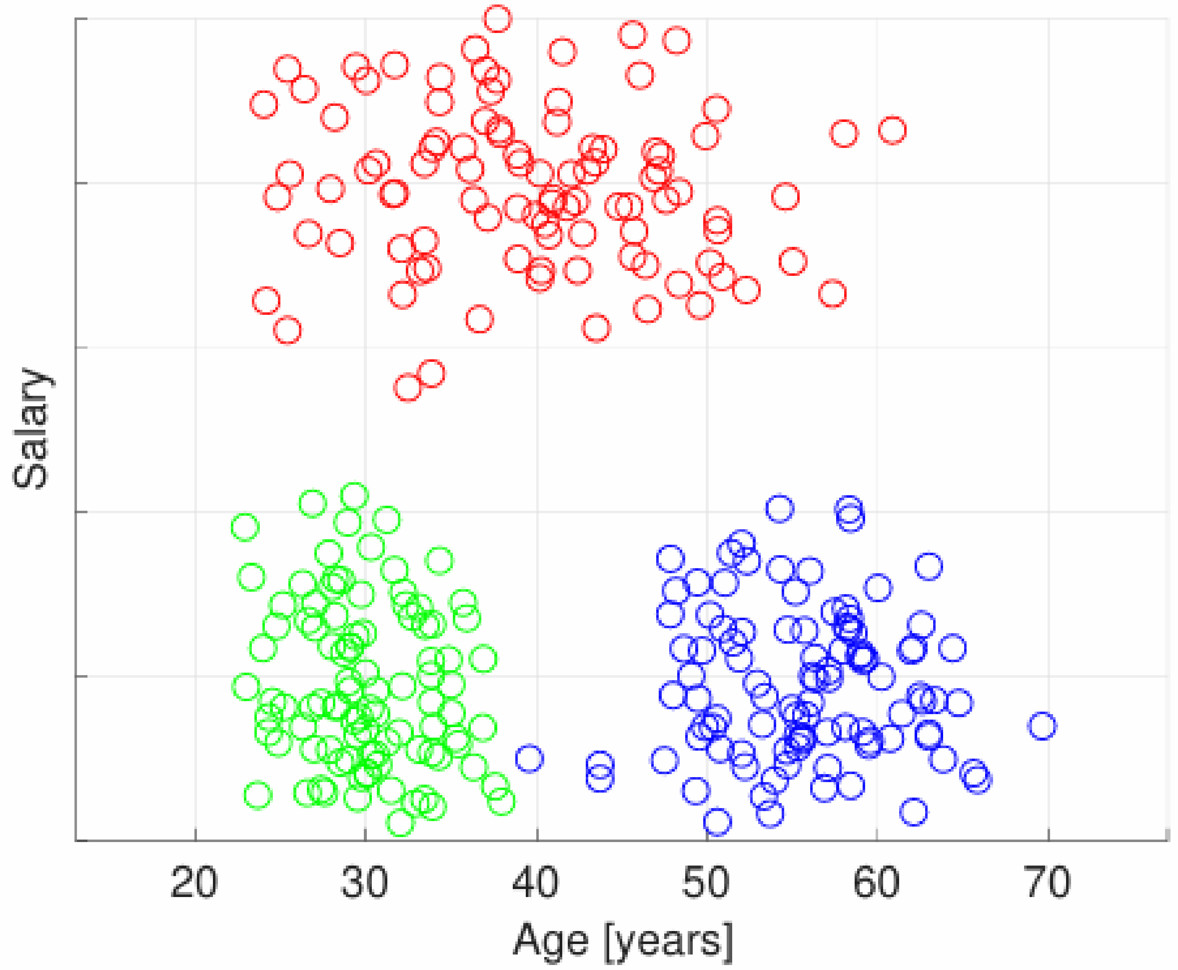
El conjunto de datos en el espacio de predictores

Una representación más conveniente es usar **diferentes símbolos para cada etiqueta** en el espacio del **espacio del predictor**

Un predictor, dos clases



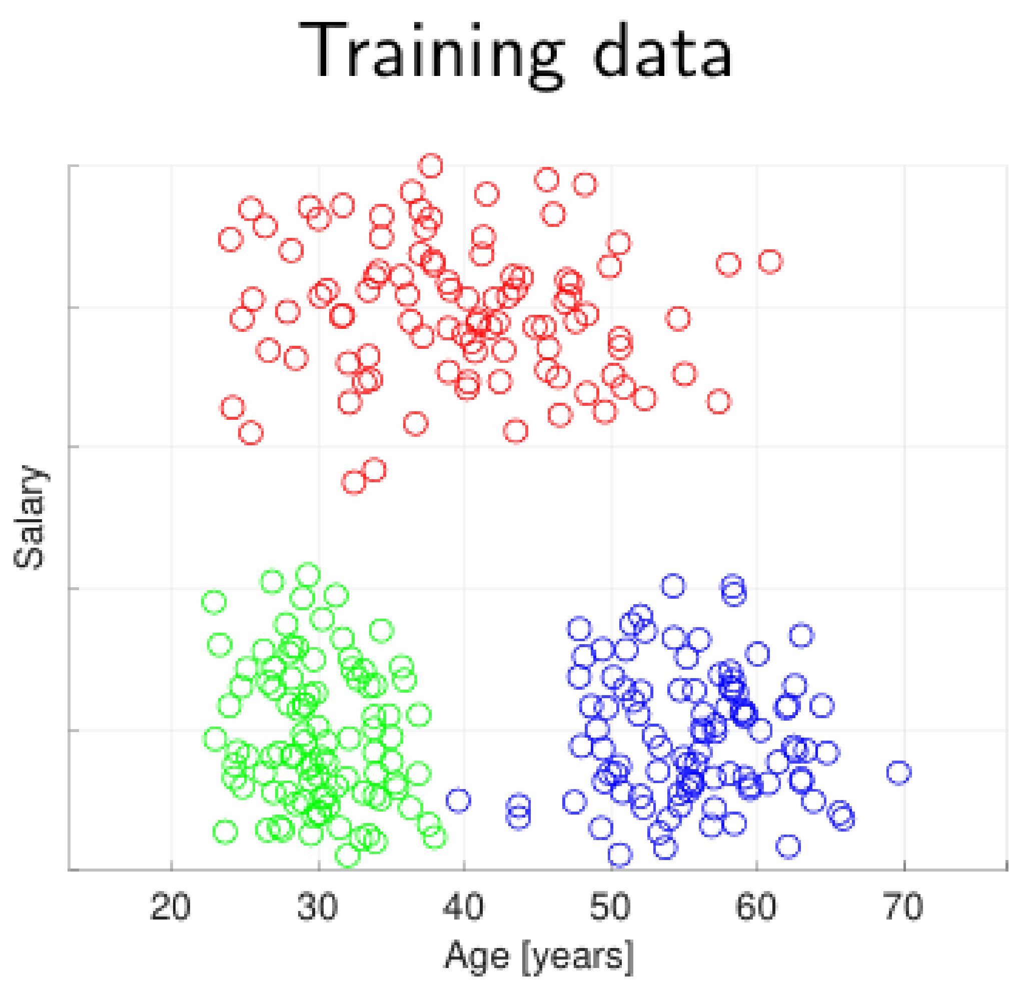
dos predictores, tres clases

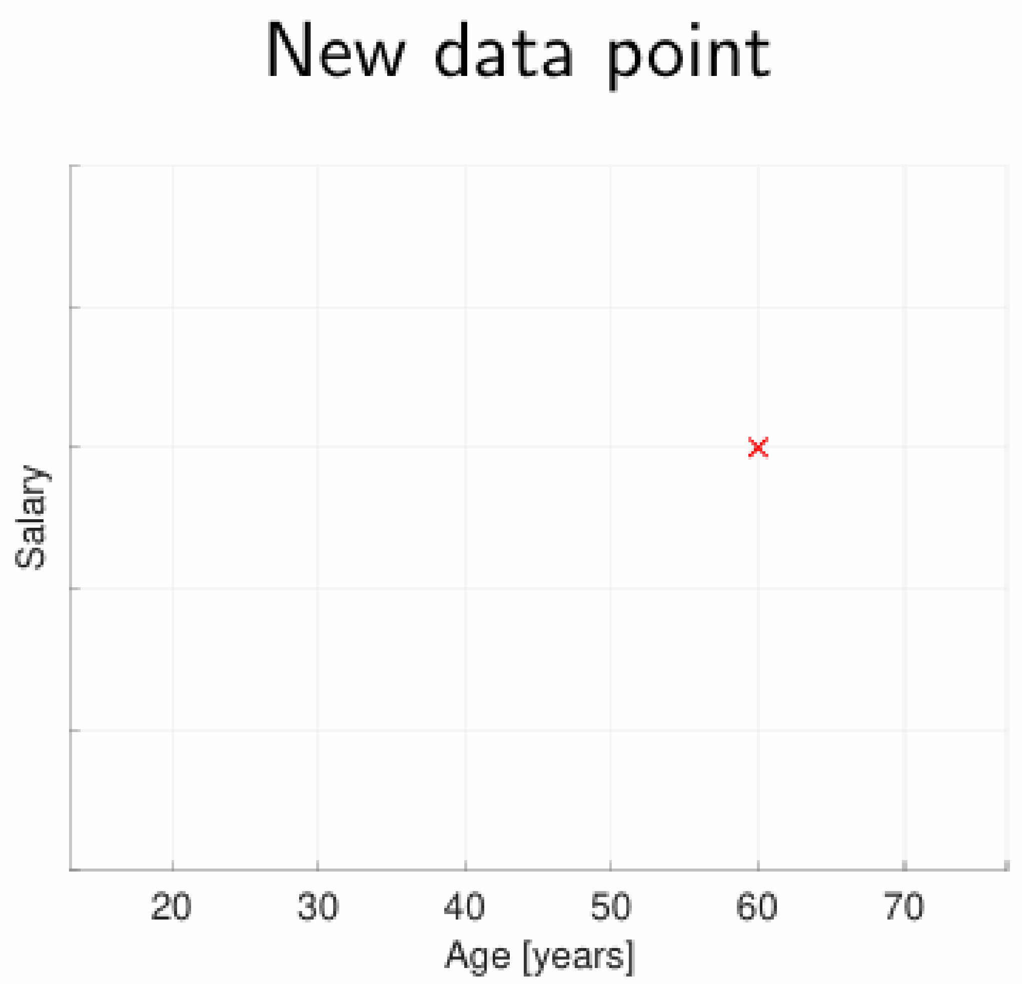


Cómo luce un clasificador

Los **modelos de regresión** pueden representarse como curvas/superficies/hipersuperficies en el espacio de atributos.

Ahora que sabemos cómo representar nuestro conjunto de datos en el espacio del predictor, ¿cómo podemos representar un **modelo de clasificación**?"

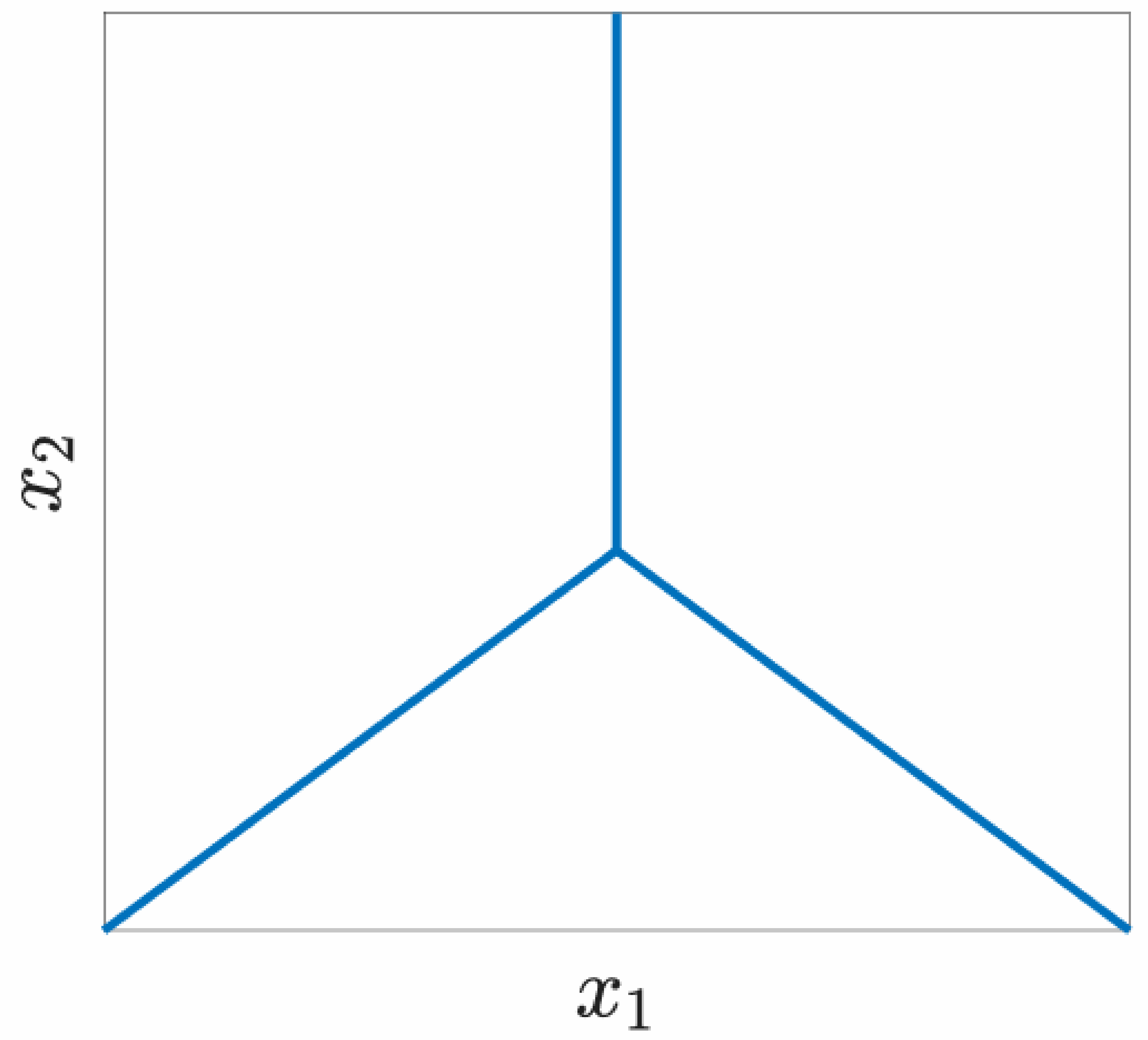




Cómo luce un clasificador

En problemas de clasificación, utilizamos la noción de **regiones de decisión** en el espacio del predictor.

- Una región de decisión está compuesta **por puntos que están asociados a la misma etiqueta.**
- Las regiones pueden definirse identificando sus **límites**.
- Un modelo de clasificación es una **partición del espacio del predictor** en regiones de decisión separadas por límites de decisión.

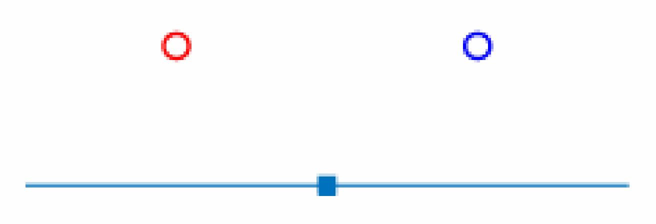
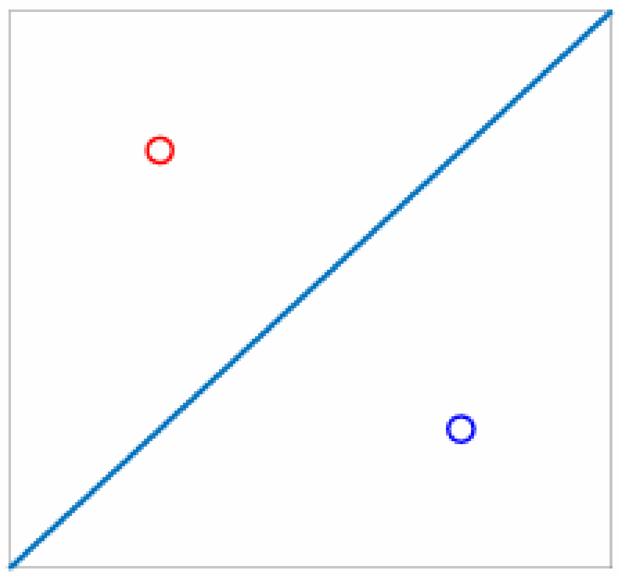
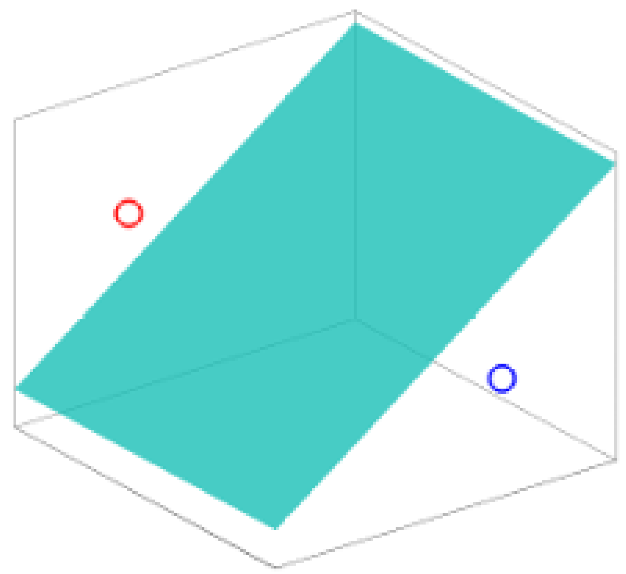


Clasificadores lineales

Consideremos un problema de **clasificación binaria**. El límite más simple es:

- Un punto (conocido como **umbral**) en espacios predictores de 1D.
- Una línea recta en espacios predictores de 2D.
- Una superficie plana en espacios predictores de 3D.

Todos estos límites son lineales

Definición clasificadores lineales

Los límites lineales están definidos por la ecuación líneal $w^T x = 0$: El vector extendido **x** contiene los predictores y **w** es el vector de coeficientes. Una muestra se clasifica identificando el lado del límite en el que se encuentra.

Si conocemos el vector de coeficientes **w** de un límite lineal, clasificar una muestra es simple:

Construye el vector extendido **xi** y calcula $w^T x_i$.

- Si $w^T x_i > 0$ estamos en un lado del límite.
- Si $w^T x_i < 0$ estamos en el otro lado.
- Si $w^T x_i = 0$ ¿dónde estamos?

Nuestro siguiente paso será encontrar el **mejor clasificador lineal** para un conjunto de datos dado. Para responder a esta pregunta, primero necesitamos definir **nuestra métrica de calidad**

Métrica de calidad

La única operación que podemos realizar con variables categóricas es la **comparación**, es decir, podemos evaluar si $y_i = \hat{y}_i$ rdadero o falso.

Al comparar las predicciones y las etiquetas verdaderas, podemos identificar en un conjunto de datos:

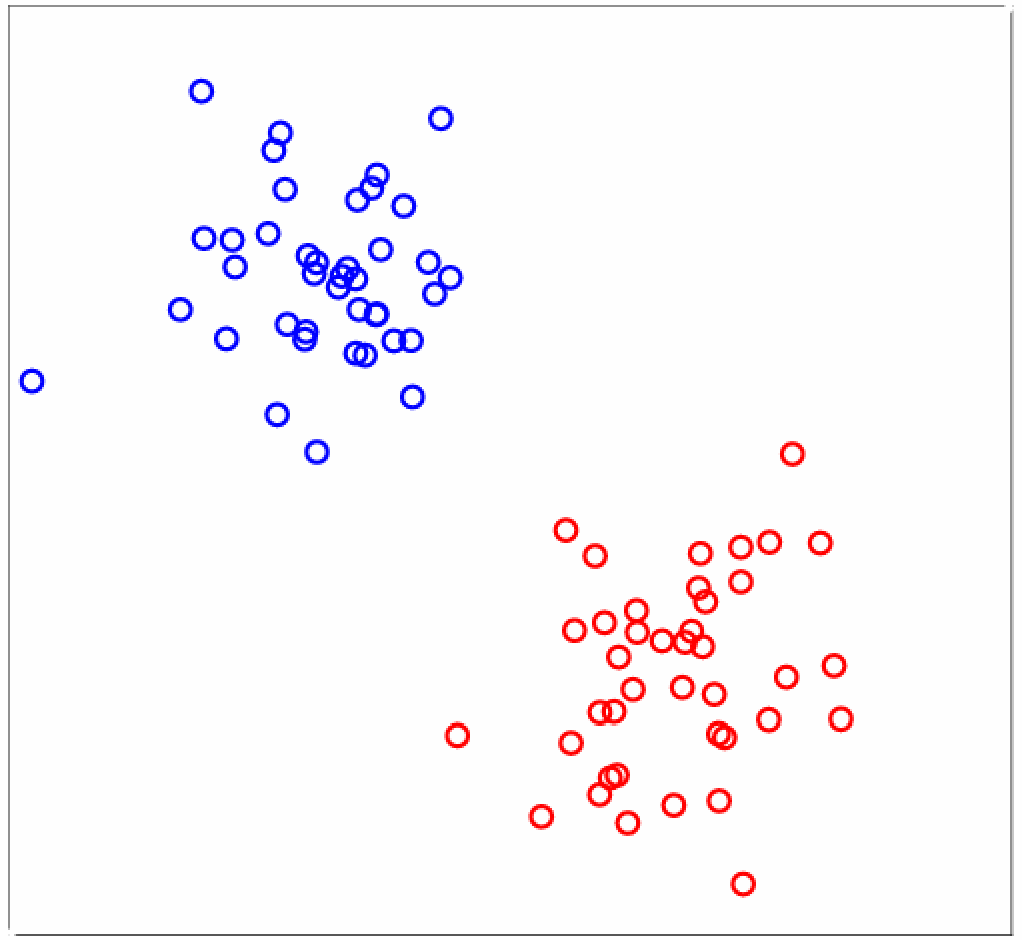
- Las muestras correctamente clasificadas **(predicciones verdaderas)** en cada clase.
- Las muestras clasificadas incorrectamente **(predicciones falsas)** en cada clase.

Dos nociones comunes y equivalentes de calidad son la precisión (**accuracy**) ***A*** y la **tasa de error *E*** (o missclassification) E=1−A definidas como:

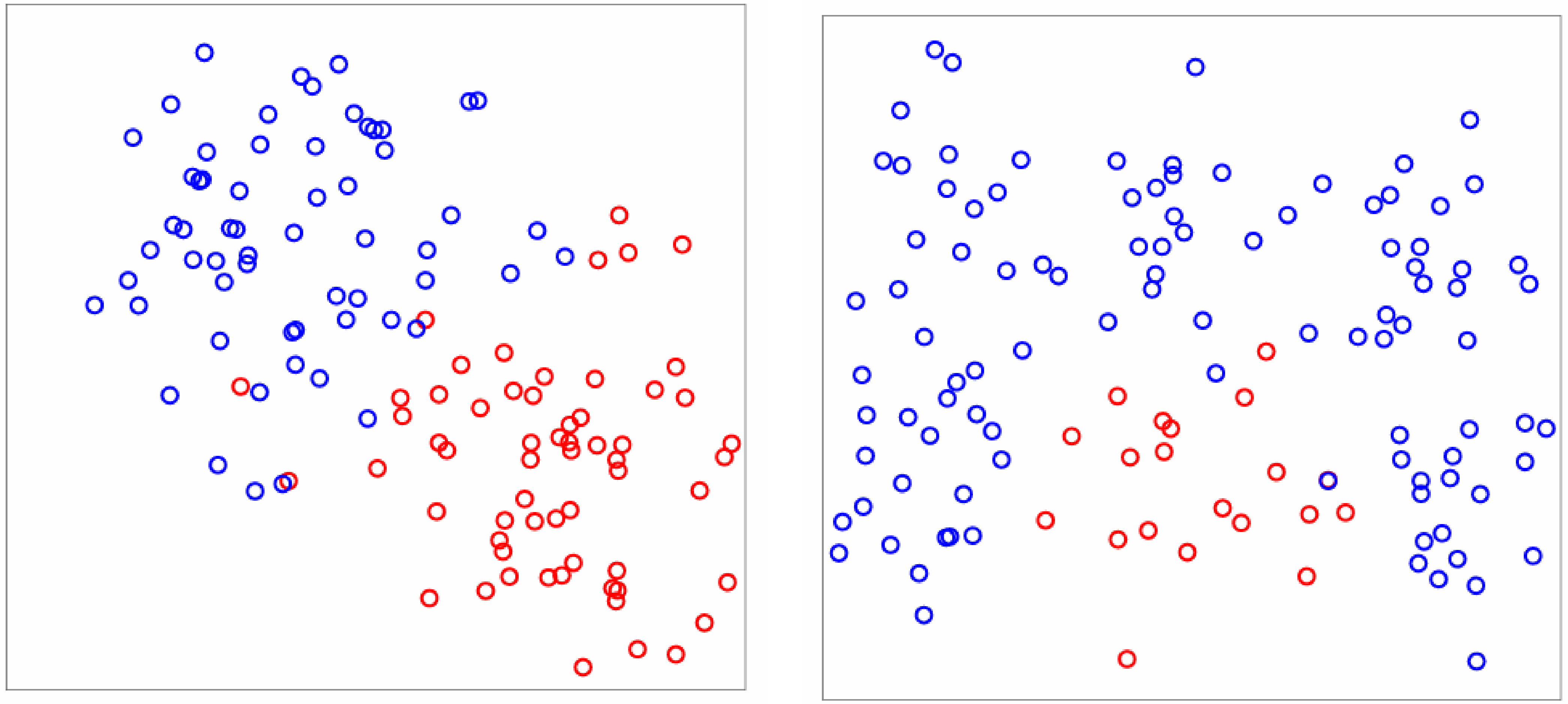
$$A = \frac{\#\text{correctly classif. samples}}{\#\text{samples}} \quad , \quad E = \frac{\#\text{incorrectly classif. samples}}{\#\text{samples}}$$

Utilizando estas nociones de calidad, el mejor clasificador se puede definir como aquel con la mayor precisión (o la tasa de clasificación errónea más baja)

Ejemplos



Dataset linealmente separable



Dataset No linealmente separable

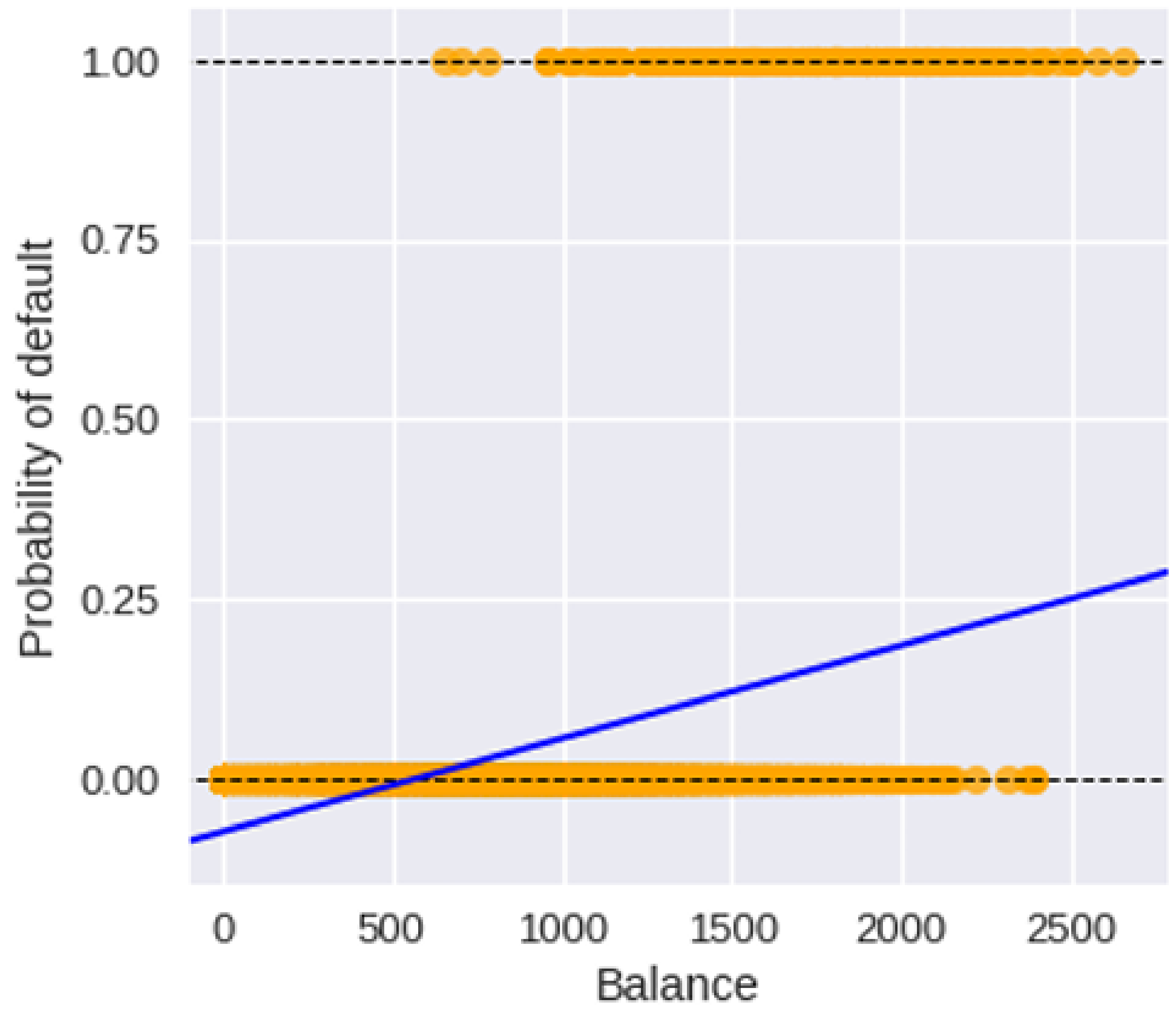
$$A = \frac{\#\text{correctly classif. samples}}{\#\text{samples}} \quad , \quad E = \frac{\#\text{incorrectly classif. samples}}{\#\text{samples}}$$

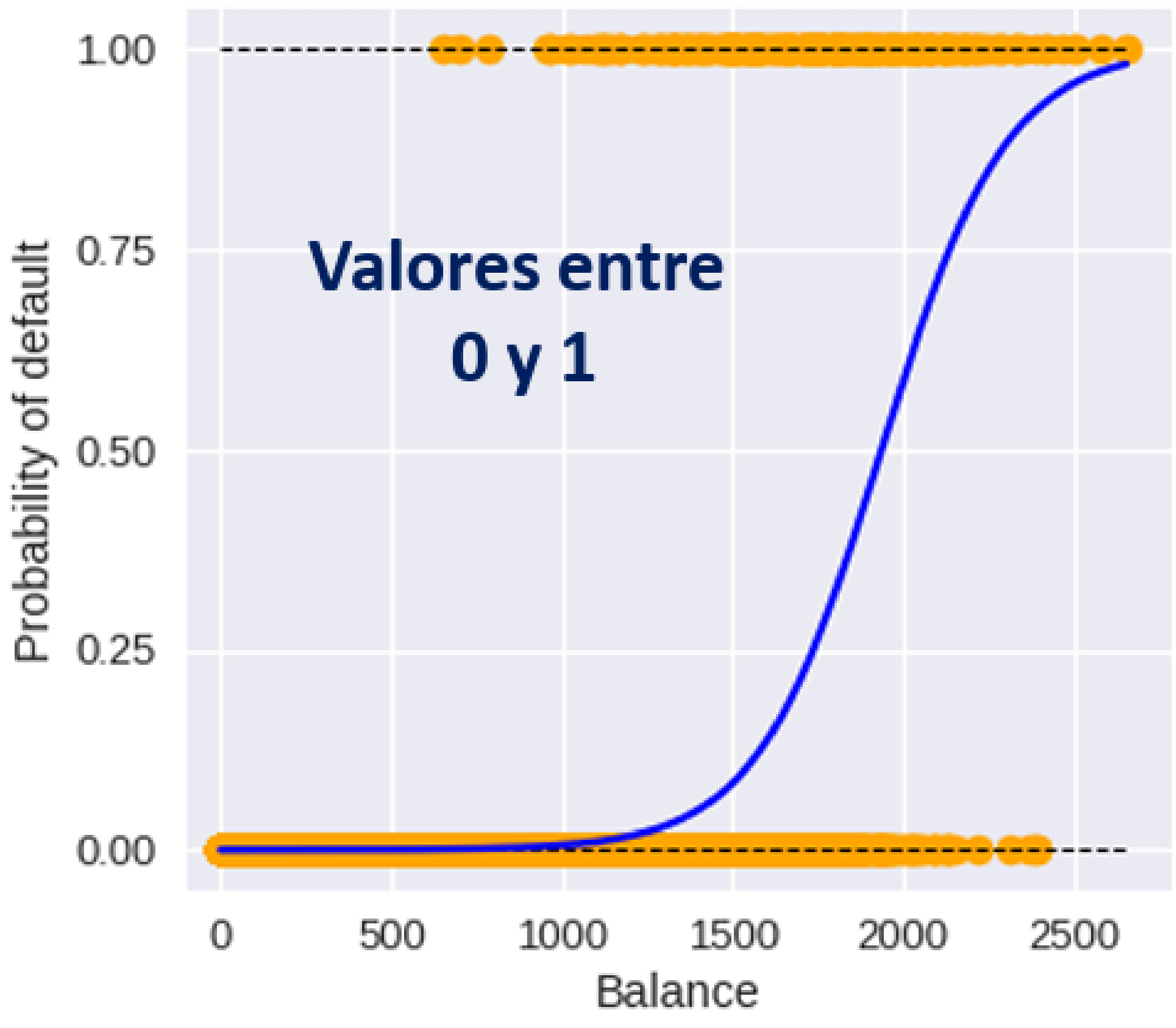
¿Modelo logístico o Regresión logística?

https://mlu-explain.github.io/logistic-regression/

Regresión Lineal vs Modelo logístico

$$p(x) = P_r(Y = 1|X)$$





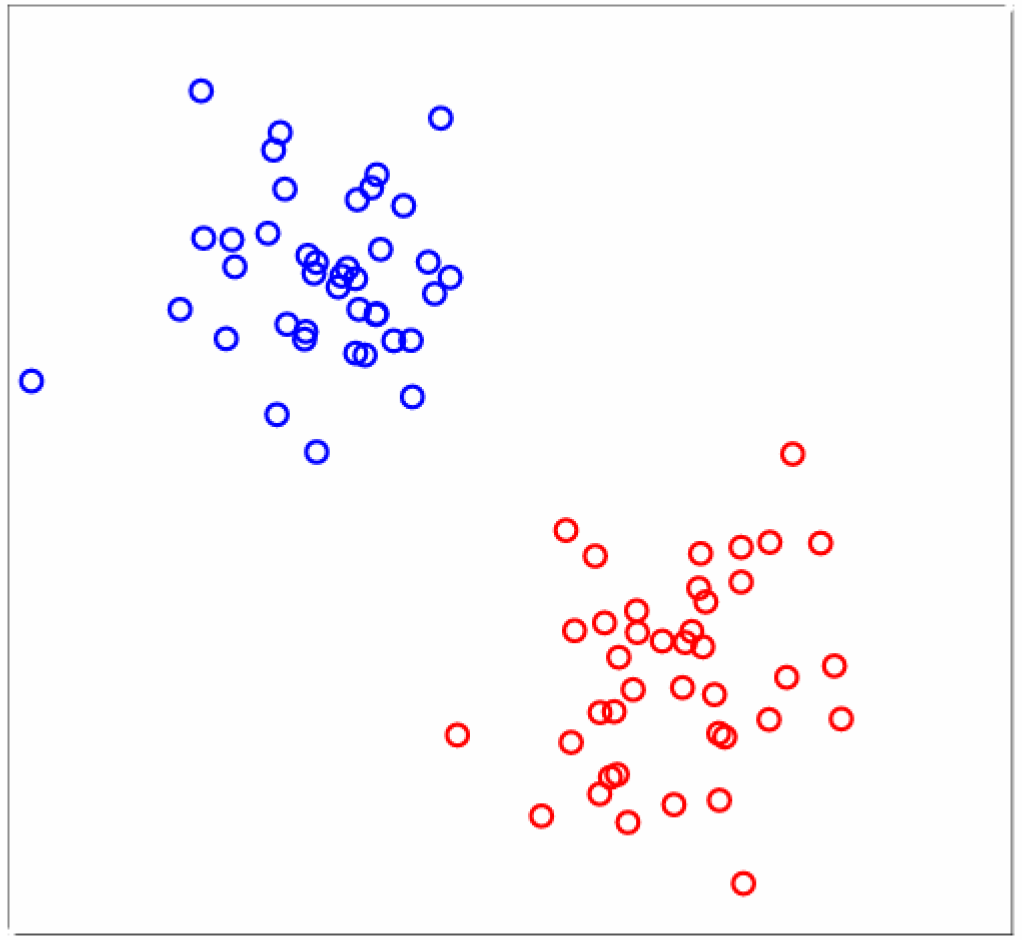
$$\boxed{p(X) = \beta_0 + \beta_1 X}$$

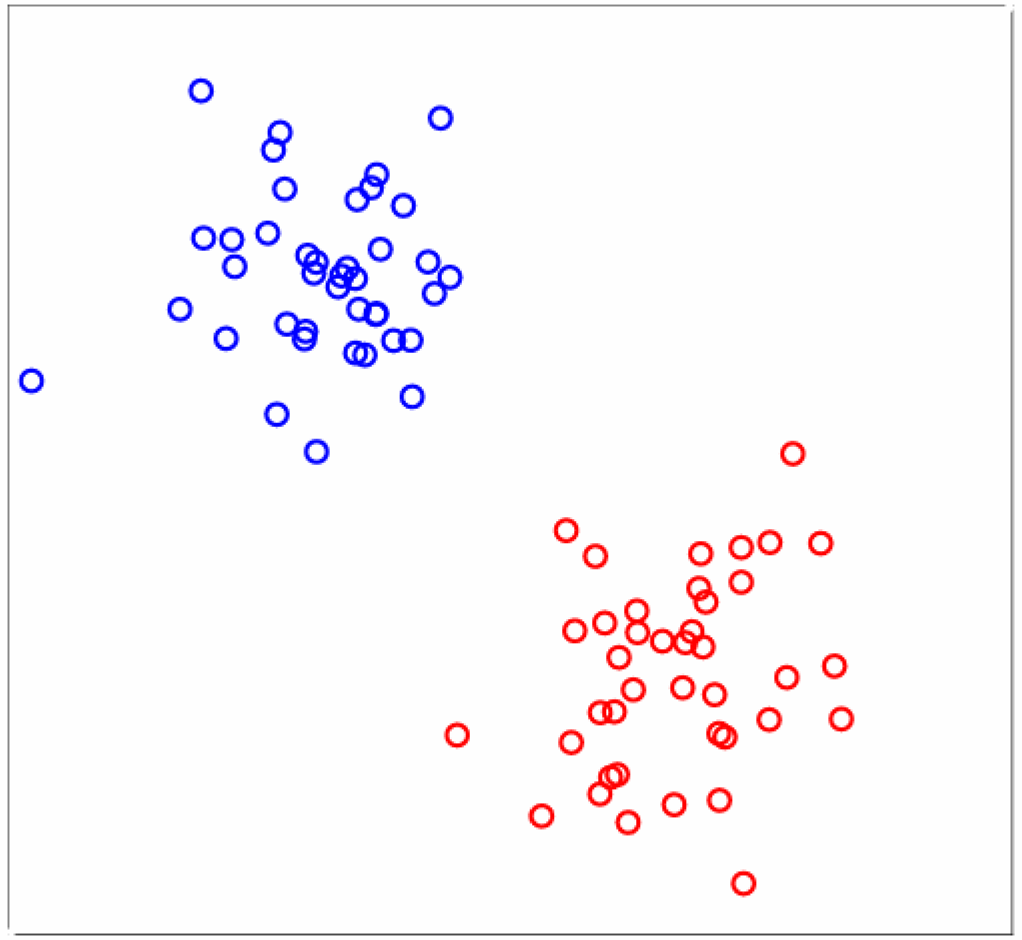
$$\boxed{p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}}$$

Las mejores soluciones lineales, pero arriesgadas

Dibuja dos límites lineales que logren una precisión A=1. ¿Cuál elegirías? ¿Por qué?

Si prefieres uno sobre el otro, podrías estar evaluando inadvertidamente su capacidad de generalización y modelando la distribución de las muestras



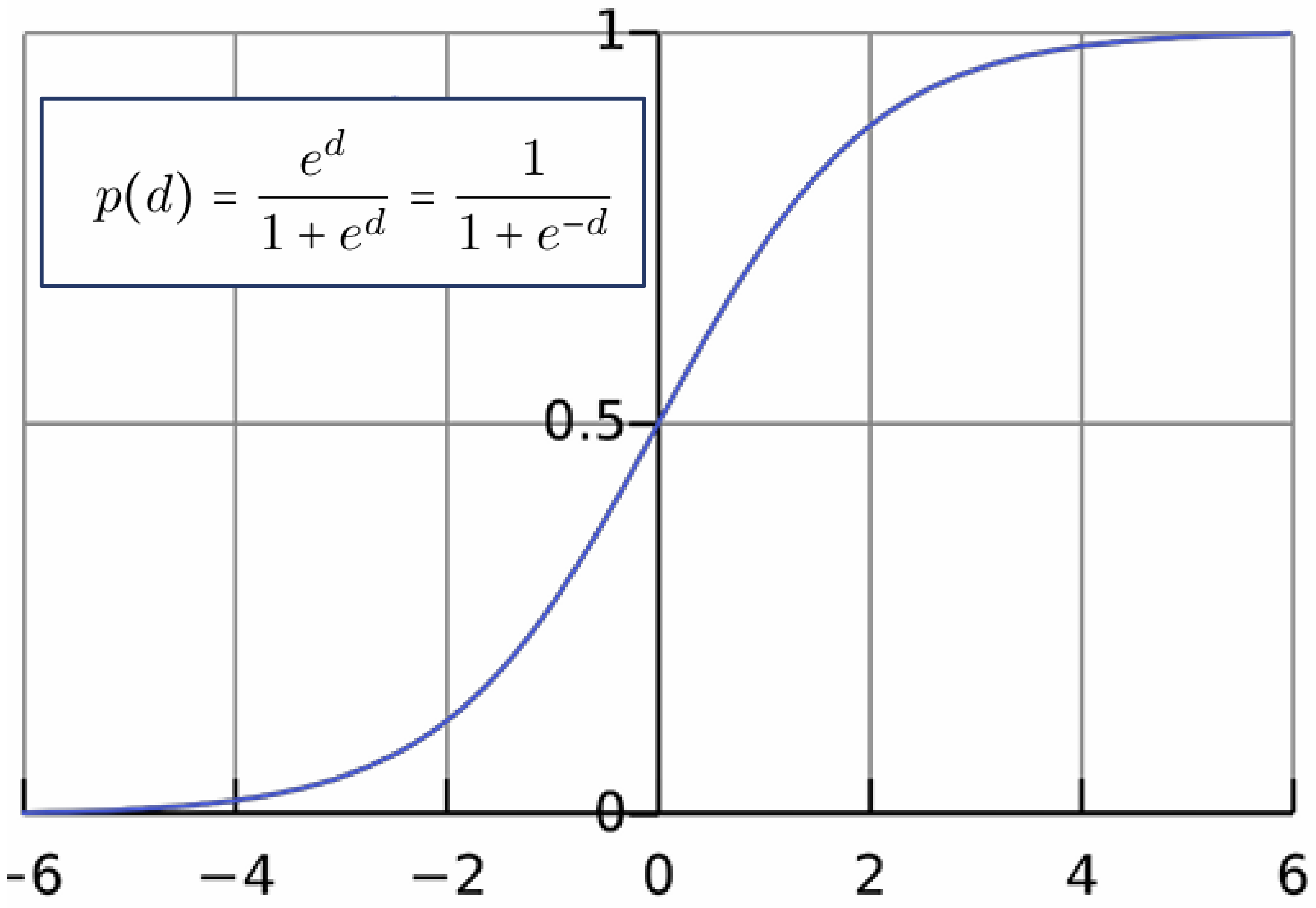


Cuanto más lejos estemos del límite, mayor será nuestra certeza de que estamos clasificando las muestras correctamente

Modelo logístico

La función logística p(d) esta definida por :

$$p(d) = \frac{e^d}{1+e^d} = \frac{1}{1+e^{-d}}$$



- $p(0) = 0.5.$
- As $d \to \infty$, $p(d) \to 1.$
- As $d \to -\infty$, $p(d) \to 0.$

Para modelar fenómenos binarios, nos interesa la probabilidad que el evento suceda o no, la cual tiene valores continuos entre 0 y 1.

Modelo logístico

Dado un límite lineal **w** y un vector predictor **xi**, la cantidad $\boldsymbol{w}^T\boldsymbol{x}_i$ se puede interpretar como la **distancia** desde la muestra hasta el límite.

Si establecemos $d = \boldsymbol{w}^T\boldsymbol{x}_i$ en la función logística, obtenemos

$$p(\boldsymbol{w}^T\boldsymbol{x}_i) = \frac{e^{\boldsymbol{w}^T\boldsymbol{x}_i}}{1 + e^{\boldsymbol{w}^T\boldsymbol{x}_i}}$$

Para un **w** fijo, simplemente lo denotaremos como **p(xi)** para simplificar la notación:

- Cuando $\boldsymbol{w}^T\boldsymbol{x} \to \infty$, tiende a infinito, la función logística $p(\boldsymbol{x}_i) \to 1$
- Cuando $\boldsymbol{w}^T\boldsymbol{x} \to -\infty$, tiende a menos infinito, la función logística $p(\boldsymbol{x}_i) \to 0$

- Utilizaremos la función logística para cuantificar la noción de certeza en los clasificadores.

Modelo logístico

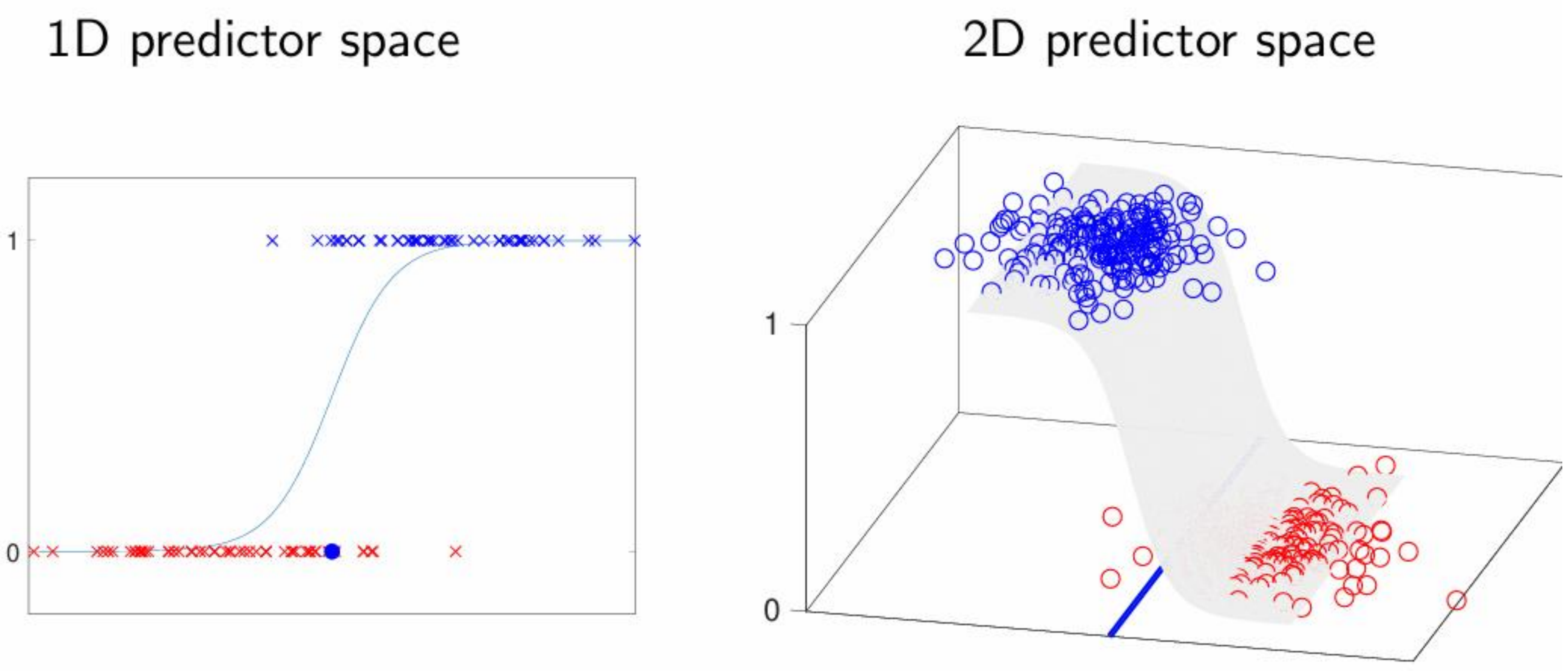
Considera un clasificador lineal **w** que etiqueta las muestras de manera que $\boldsymbol{w}^T\boldsymbol{x}_i > 0$ como azul y muestras donde $\boldsymbol{w}^T\boldsymbol{x}_i < 0$ como rojo.

Observa que:

Si $w^T xi = 0$ (xi está en el límite), $p(xi) = 0.5$.
Si $w^T xi > 0$ (xi está en la región azul), $p(xi)$ tiende a 1 a medida que nos alejamos del límite.
Si $w^T xi < 0$ (xi está en la región roja), $p(xi)$ tiende a 0 a medida que nos alejamos del límite.

p(xi) es la certeza del clasificador de que yi es azul.
1−p(xi) es la certeza del clasificador de que yiyi es rojo.

Visualización modelo logístico



La verisimilitud

Podemos obtener la certeza del clasificador de que xi pertenece a azul o rojo. ¿Podemos calcular la certeza para un conjunto de **datos etiquetado** (xi,yi)?

La respuesta es sí, **multiplicando** las certezas individuales:

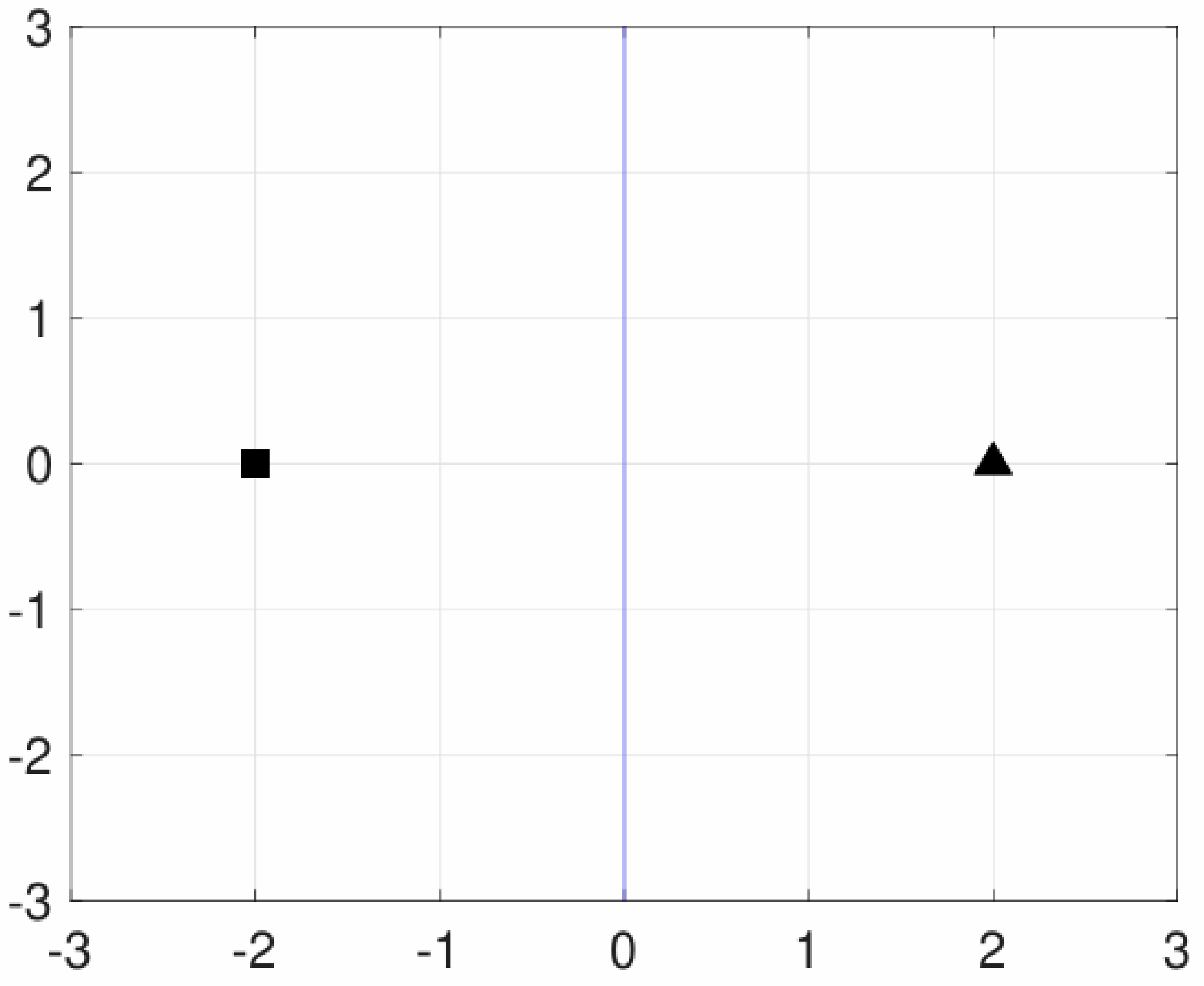
$$L = \prod_{y_i = \text{o}} (1 - p(\boldsymbol{x}_i)) \prod_{y_i = \text{o}} p(\boldsymbol{x}_i)$$

L es conocido como la **función de verosimilitud** (likelihood) y define una **métrica de calidad**. Tomando logaritmos, obtenemos la L es conocido como la función de verosimilitud y define una métrica de calidad. Tomando logaritmos, obtenemos la log-verosimilitud (log-likelihood):

$$l = \sum_{y_i = \text{o}} \log\left[1 - p(\boldsymbol{x}_i)\right] + \sum_{y_i = \text{o}} \log\left[p(\boldsymbol{x}_i)\right]$$

El clasificador lineal que maximiza **L** o ***l*** es conocido como el clasificador de **Regresión Logística**. Se puede encontrar usando el descenso del gradiente.

Ejemplo



- Definimos $d_i = \boldsymbol{w}^T \boldsymbol{x}_i$
- Nosotros podemos escribir la función logística como:

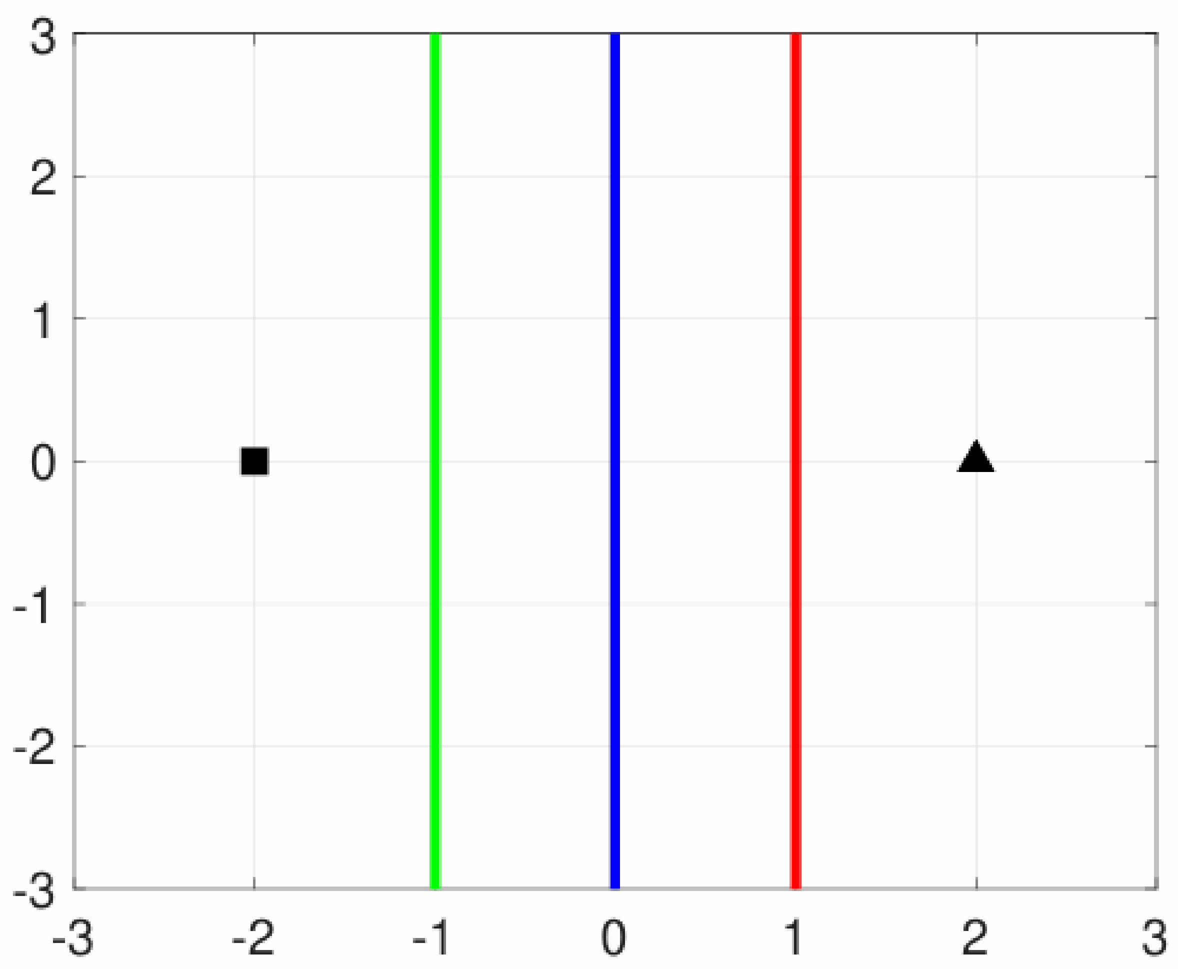
$$p(d_i) = \frac{e^{d_i}}{1 + e^{d_i}}$$

- Por lo tanto, $p(0) = 0.5$, $p(1) \approx 0.73$, $p(2) \approx 0.88$, $p(-1) \approx 0.27$ and $p(-2) \approx 0.12$

Supón que este clasificador lineal etiqueta las muestras en el semiplano derecho como $\triangle$ y las muestras en el semiplano izquierdo como $\square$

$$p(\triangle) \approx 0.88,\ 1 - p(\square) \approx 0.88 \text{ and } L = p(\triangle)\,(1 - p(\square)) \approx 0.77.$$

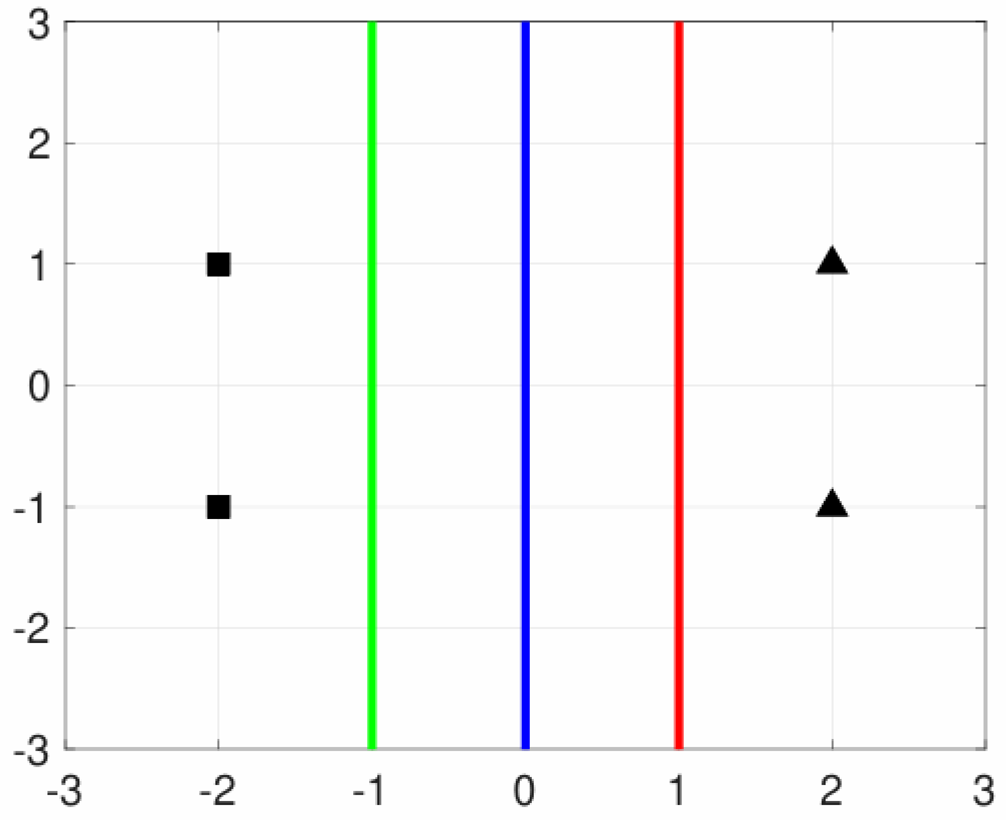
Ejemplo

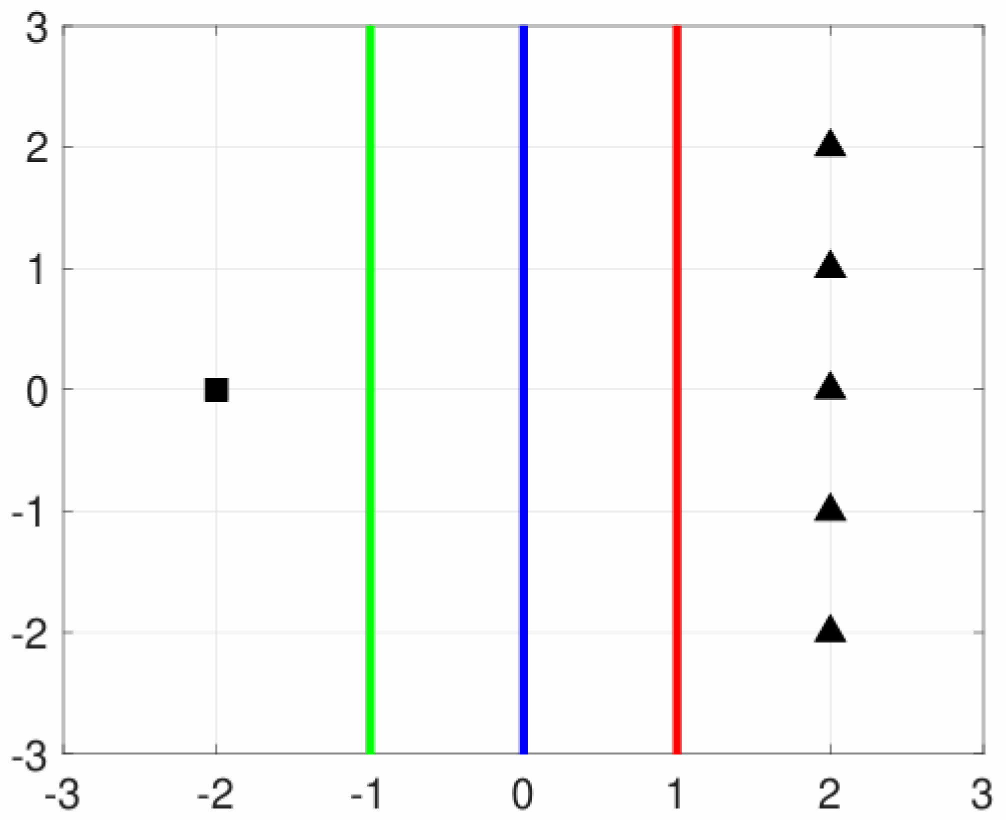


Y las verosimilitudes de cada clasificador

- $L = p(\triangle)\,(1 - p(\square)) \approx 0.70$
- $L = p(\triangle)\,(1 - p(\square)) \approx 0.77$
- $L = p(\triangle)\,(1 - p(\square)) \approx 0.70$

Ejemplo





Y las verosimilitudes de cada clasificador, cuales son?

Laboratorio sencillo

Construir el mejor clasificador para el ítem ¿Esperas que tu hijo ingrese en una carrera de ciencias, para un país determinado ejemplo Mexico? Para ello

- Usa el dataSet de pisa2015
- Usa la variable dependiente la PA032Q03TA

```
data = pd.read_parquet("pisa2015.parquet")
data_Mex = data.loc[data.CNT == 'Mexico']
data_Mex["ciencia_expectativas"]=data_Mex.PA032Q03TA
```

- Separa los datos para entrenamiento y validación
- Implementa el modelo logístico para responder a la pregunta, puede usar LogisticRegression de sklearn o tu mismo crear las funciones para tal efecto

```
logreg_model = LogisticRegression(random_state=0, max_iter=500)
logreg_model.fit(X_train, y_train)
```

- Cual es el log-likelihood

Usa las características que consideres necesarias

Nombre variable	Descripción
DISCLISCI	Clima disciplinario en clases de ciencias (WLE)
TEACHSUP	Apoyo del profesor en clases de ciencias elegidas por los estudiantes (WLE)
IBTEACH	Prácticas de enseñanza y aprendizaje de ciencias basadas en la investigación (WLE)
TDTEACH	Instrucción en ciencias dirigida por el profesor (WLE)
ENVAWARE	Conciencia ambiental (WLE)
JOYSCIE	Disfrute de la ciencia (WLE)
INTBRSCI	Interés en temas amplios de ciencia (WLE)
INSTSCIE	Motivación instrumental (WLE)
SCIEEFF	Autoeficacia en ciencias (WLE)
EPIST	Creencias epistemológicas (WLE)
SCIEACT	Índice de actividades científicas (WLE)
BSMJ	Expectativa de estatus ocupacional del estudiante (SEI)
MISCED	Educación de la madre (ISCED)
FISCED	Educación del padre (ISCED)
OUTHOURS	Tiempo de estudio fuera de la escuela por semana (Suma)
TMINS	Tiempo de aprendizaje (minutos por semana) - Total
BELONG	Bienestar subjetivo: Sentimiento de pertenencia a la escuela (WLE)
ANXTEST	Personalidad: Ansiedad ante los exámenes (WLE)
MOTIVAT	Actitudes, preferencias y creencias auto relacionadas del estudiante: Motivación para el logro (WLE)
COOPERATE	Disposiciones para la colaboración y el trabajo en equipo: Disfrute de la cooperación (WLE)
PERFEED	Retroalimentación percibida (WLE)
unfairteacher	Justicia del profesor (Suma)

HEDRES	Recursos educativos en el hogar (WLE)
HOMEPOS	Posesiones en el hogar (WLE)
ICTRES	Recursos de TIC (WLE)
WEALTH	Riqueza familiar (WLE)
ESCS	Índice de estatus económico, social y cultural (WLE)
PV1MATH	Puntuación de matemáticas de los estudiantes en PISA 2015
PV1READ	Puntuación de lectura de los estudiantes en PISA 2015

Modelos lineales generalizados

Recordando

El modelo se estructura para describir patrones específicos de interacciones y relaciones.

Los parámetros que componen este modelo cuantifican la intensidad de dichas relaciones. La esencia de trabajar con modelos radica en la estimación precisa de estos parámetros.

Para lograrlo, se utilizan herramientas fundamentales de inferencia, tales como la estimación puntual, las pruebas de hipótesis y la construcción de intervalos de confianza.

- Objetivo: ¿Qué busca el modelo?
- Estructura: Variables, fórmula, ecuación
- Supuestos: Premisas del modelo
- Estimación de Parámetros: Interpretación y significado
- Ajuste del Modelo: Evaluación y estadísticas
- Selección: Variables a incluir

Recordando

En el contexto de la regresión lineal simple, considera lo siguiente:

Objetivo: Estimar el valor esperado de una variable continua **Y** como función lineal de un predictor continuo **X**.

Estructura del Modelo: La relación lineal se describe mediante la ecuación $\boxed{Y_i = \beta_0 + \beta_1 x_i + e_i}$
.
Supuestos: La variable Y sigue una distribución normal. Los errores, representados por **ei** también son normalmente distribuidos, independientes entre sí, y con una varianza constante **σ2**. Además, se considera que **X** es una variable no aleatoria con varianza constante.

Estimación e Interpretación de Parámetros: El coeficiente **β0^** es una estimación de la ordenada al origen, mientras que β1^β1^ estima la pendiente. Es fundamental entender su significado: ¿Qué representan en el contexto de tus datos?

Ajuste del Modelo: Utiliza métricas como R cuadrado , análisis de residuos se evaluar cuán bien se ajusta el modelo a los datos.

Selección de Variables: Entre numerosos posibles predictores, es crucial determinar cuáles son relevantes para el modelo.

Modelos lineales generalizados

El Modelado Lineal Generalizado es una extensión de los modelos lineales tradicionales que permite modelar respuestas que no siguen necesariamente una distribución normal (como respuestas binarias o conteos).

1. **Captura Patrones**: El modelo refleja las relaciones y tendencias principales de los datos.

2. **Determina Relaciones Importantes**: Ayuda a identificar qué variables influyen significativamente en la respuesta. (puedes hacer "inferencia",)

3. **Mide la Fuerza de Efectos**: Los coeficientes indican cuánto impacto tiene cada variable en la respuesta. **(importancia estadística)**

4. **Predicciones Claras**: Las predicciones del modelo ofrecen una visión "limpia" de los datos, eliminando el ruido aleatorio. (**"suavizan" los datos**)

GLM van más allá de la Regresión Lineal Simple

Variable Respuesta: puede tener una distribución diferente a la normal — cualquier distribución dentro de una clase de distribuciones conocida como “familia exponencial de distribuciones”, (Normal, Binomial, Poisson, etc.)..

Función de enlace: La relación entre la respuesta (Y) y las variables explicativas no necesita ser simple ("identidad").

Por ejemplo, en lugar de $Y = \alpha + \beta x$ podemos permitir transformaciones de Y : $g(Y) = \alpha + \beta x$

Componente lineal: Como en la regresión lineal $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$

Ejemplo

Los recuentos corresponden a células T/mm en muestras de sangre de 20 pacientes en remisión de la enfermedad de Hodgkin y 20 pacientes en remisión de malignidades diseminadas..

Hodgkin's		Non-Hodgkin's	
396	568	375	375
1212	171	752	208
554	1104	151	116
257	435	736	192
295	397	315	1252
288	1004	657	700
431	795	440	771
1621	1378	688	426
902	958	410	979
1283	2415	377	503

¿Existe una "diferencia" en los recuentos celulares entre las dos enfermedades?

Ejemplo

Qué se Entiende por "Diferencia"?

- Promedio
- Variabilidad
- Forma general de la distribución

Enfoque Ingenuo:

Suponer una distribución normal y realizar una "prueba t" (es decir, calcular la diferencia entre las medias y dividir por el error estándar de la diferencia).

Enfoque más Sofisticado:

Suponer una distribución de Poisson y calcular la diferencia entre el logaritmo de las medias (es decir, la proporción de las medias).

Componentes fundamentales modelo lineal generalizado

1. Random Component: identifica la respuesta de la variable Y. Distribución de probabilidad de la variable de respuesta

Binaria / discontinua	Conteo
Distribución binomial	Distribución Poisson

2. Systematic Component: especifique cuál es el componente explicativo o las variables predictoras son (por ejemplo, X1, X2, etc.). Estas variables entran de manera lineal

$$\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

3. Link: Especifica la relación entre la media o valor esperado del componente aleatorio (es decir, E(Y)) y el componente sistemático $\Longrightarrow g(\mu)$

$$E(Y) = \alpha + \beta x$$

$$\log(E(Y)) = \log(\mu) = \alpha + \beta x$$

Enlace logarítmico

$$\log(\mu/(1-\mu)) = \alpha + \beta x$$

Enlace logit 0 ≤ μ ≤ 1

Formula General

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Distribution	"Natural Parameter"	"Canonical Link"
Normal	μ	Identity
Poisson	$\log(\mu)$	log
Binomial	$\log(\mu/(1-\mu))$	logit

Model	**Random**	**Link**	**Systematic**
Linear Regression	Normal	Identity	Continuous
ANOVA	Normal	Identity	Categorical
ANCOVA	Normal	Identity	Mixed
Logistic Regression	Binomial	Logit	Mixed
Loglinear	Poisson	Log	Categorical
Poisson Regression	Poisson	Log	Mixed
Multinomial response	Multinomial	Generalized Logit	Mixed