Clasificación

Modelos para entender una realidad caótica

Aprendizaje supervisado y no supervisado

Problema de Clasificación

Aprendizaje supervisado

Clasificación

Datos de entrenamiento que contienen tanto las características x_i como las categorías/etiquetas t_i

Numero de categorías finito

Objetivo: Clasificar datos de entrada en una de un numero finito de categorías

K vecinos más cercanos KNN

KNN – Clasificación

Método tanto de predicción como de clasificación

Idea sencilla e intuitiva

Un pronóstico se basa en las k observaciones mas “parecidas”

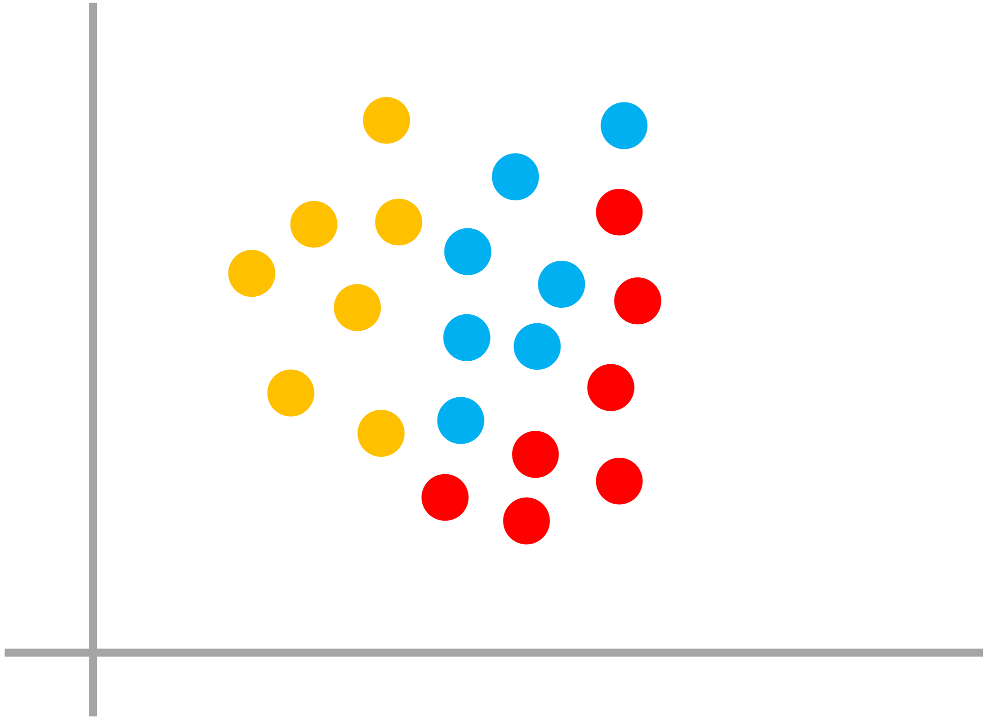
Cada observación en un conjunto de datos es un punto en el espacio n-dimensional. El número de dimensiones es igual al número de variables.

KNN (k-Nearest Neighbors)

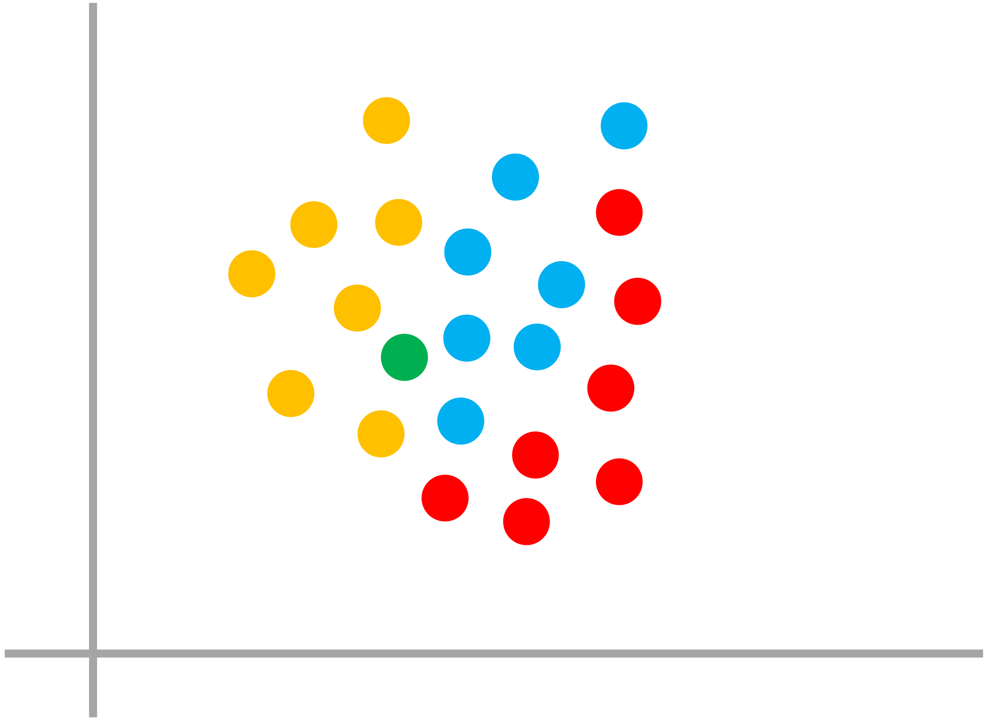
Para predecir la categoría a la que debería pertenecer una nueva observación, se buscan las k observaciones estén más cercanas. Estás serán sus “vecinos”

Luego, se toma como predicción la categoría modal (Aquella que aparezca en mas ocasiones en los “vecinos”

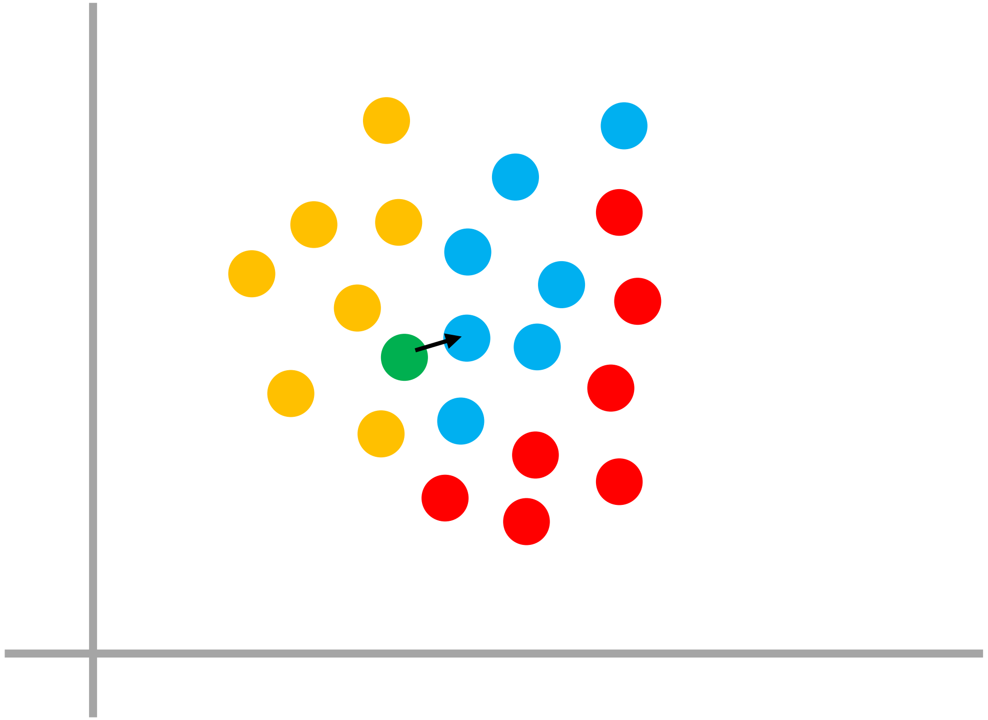
KNN (k-Nearest Neighbors)



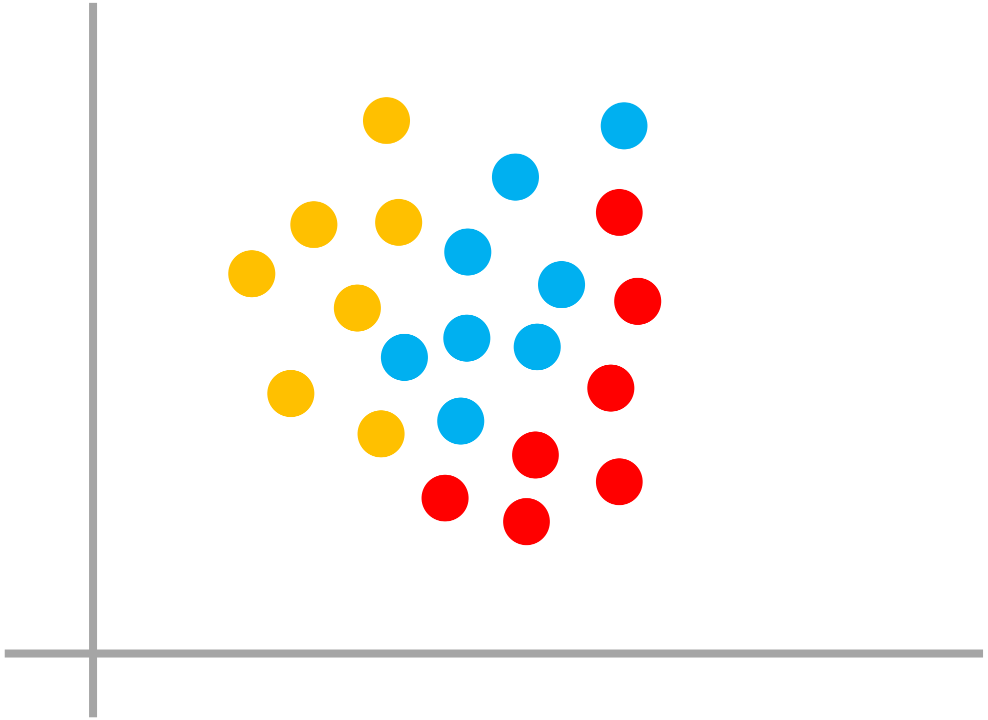
KNN (k-Nearest Neighbors)



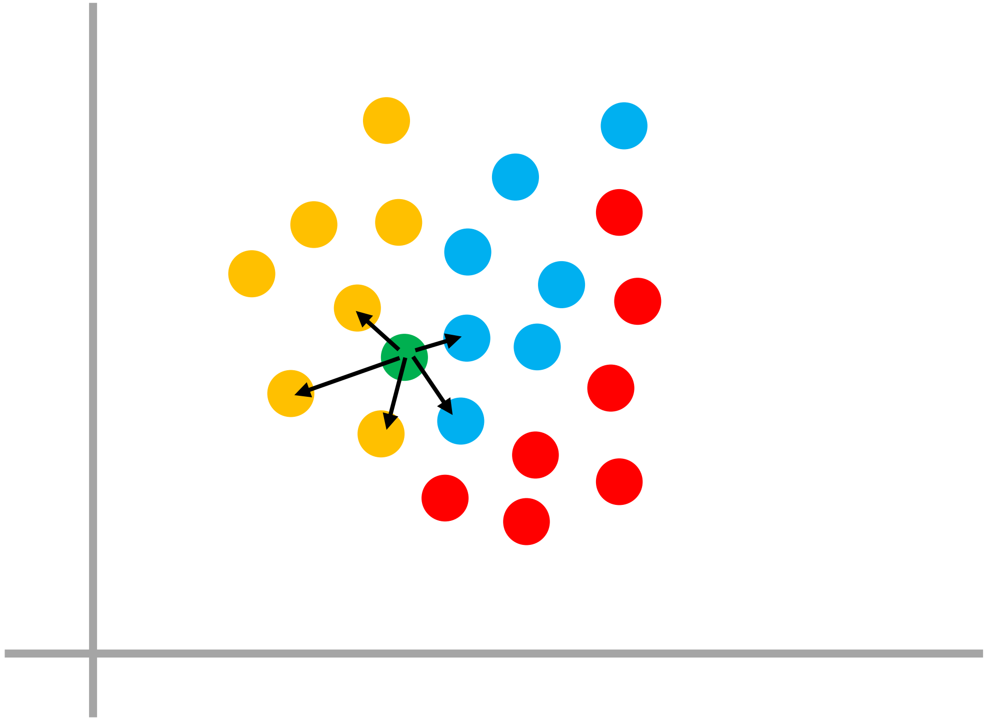
KNN (k-Nearest Neighbors)



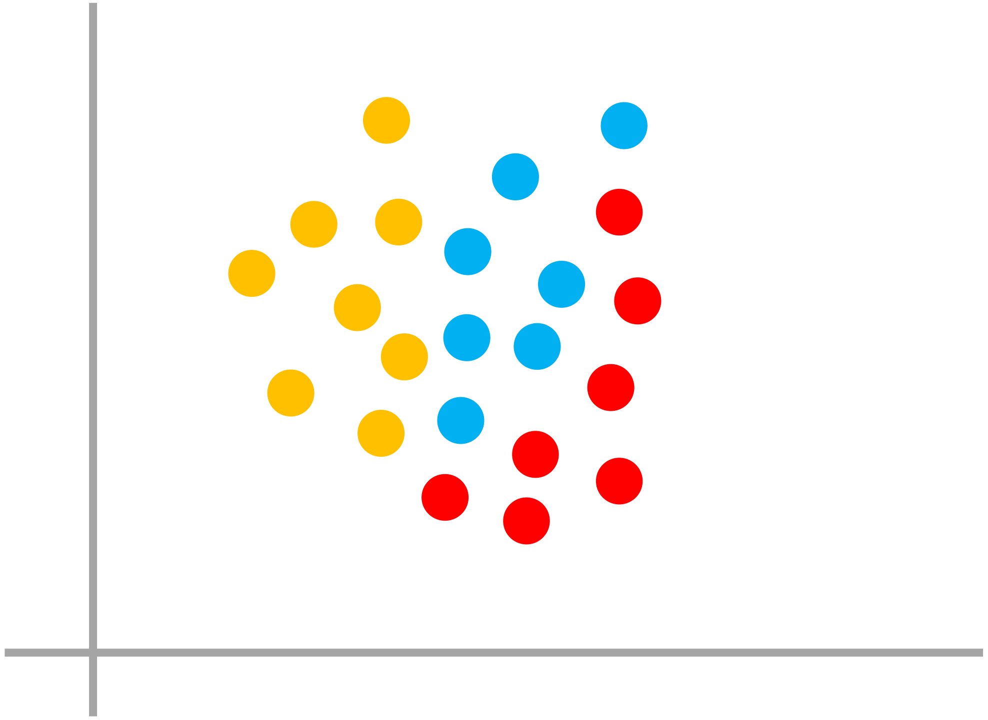
KNN (k-Nearest Neighbors)



KNN (k-Nearest Neighbors)



KNN (k-Nearest Neighbors)



KNN (k-Nearest Neighbors)

Para predecir la categoría a la que debería pertenecer una nueva observación, calcular las distancias euclidianas entre las variables numéricas y quedarse con las k observaciones que presenten menor distancia euclidiana a esta nueva observación. Estás serán sus “vecinos”

Luego, se toma como predicción la categoría modal (Aquella que aparezca en mas ocasiones en los “vecinos”

Implementación

1. Normalizar los datos (Ponerlos en una escala comparable en cada una de las variables)

2. Para predecir la categoría a la que debería pertenecer una nueva observación, calcular las distancias euclidianas entre las variables numéricas y quedarse con las k observaciones que presenten menor distancia euclidiana a esta nueva observación. Estás serán sus “vecinos”

3. Tomar como predicción la categoría modal (Aquella que aparezca en mas ocasiones en los “vecinos”

Ejercicio 1

1. Crear una función que (1) normalice o (2) estandarice los datos a través de:

$$z_i = \frac{x_i - \min(X)}{\max(X) - \min(X)}$$

$$z_i = \frac{x_i - \bar{X}}{sd(X)}$$

Donde X es la variable a ser normalizada y x_i es la observación a ser normalizada

Ejercicio 2

2. Aplicar la función de normalización a todas las variables del conjunto de datos, almacenar estas nuevas bases como datos_norm y datos_est

Implementación en R (base iris)

```
set.seed(1234)

split <- sample.split(iris$labels, SplitRatio = 0.6)

training_set <- subset(iris, split == TRUE)

test_set <- subset(iris, split == FALSE)

knn.3 <-  knn(train=training_set[, -5],
          test=test_set[, -5],
          cl=training_set$Species,
          k=3,
          prob = TRUE)
```

Ejercicio 3

Implementar el algoritmo knn para 5, 10, 15 vecinos más cercanos

Validación

Validación cruzada (precisión)

Calcular el número de observaciones que fueron bien clasificadas y dividirlas por el total de datos de prueba

Estaremos calculando la probabilidad de clasificar bien una observación en la muestra de prueba

Ejercicio 4

Calcular la precisión del modelo

¿Cómo seleccionar el número de vecinos que deberían ser escogidos?

Para mejorar el rendimiento del modelo se puede encontrar el valor de “k” que nos provee la precisión máxima del modelo

Para seleccionar este valor:

1. Se ajusta el modelo para los diferentes valores de k y se calcula su precisión
2. Se selecciona el valor de “k” que presente la mayor precisión

Ejercicio 5

Calcular la precisión del modelo desde 1 hasta 40 vecinos y seleccionar el valor de “k” que maximiza la precisión

Hacer un gráfico que permita observar este resultado

Ejercicio 6

Hacer un clasificador para el grupo al que pertenece un país de acuerdo a sus características para la base de datos UN de la librería carData

Regresión Logística

Datos

El conjunto de datos contiene la información de 1456 empleados de una empresa

Variables:

```
> names(datos)
 [1] "satisfaction_level"    "last_evaluation"       "number_project"
 [4] "average_montly_hours"  "time_spend_company"    "Work_accident"
 [7] "left"                  "promotion_last_5years" "sales"
[10] "salary"
```

Objetivo:

Saber si un empleado va a abandonar la compañía (left) de acuerdo con algunas variables de las presentadas

Clasificación: Modelos con respuesta binaria

¿Qué pasa si la variable respuesta es una variable binaria?

En este caso, la variable "*y*" puede tomar los valores cero o uno.

Supongamos que:

$$y_i = \begin{cases} 1 \; con \; probabilidad \; p_i \\ 0 \; con \; probabilidad \; 1 - p_i \end{cases}$$

$$E(y_i) = p_i$$

Si x toma valores en un rango diferente a cero o uno este modelo no es adecuado

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

Para nuestro caso

$$y_i = \begin{cases} El\ empleado\ si\ abandona\ va\ con\ probabilidad\ p_i \\ El\ empleado\ no\ abandona\ con\ probabilidad\ 1 - p_i \end{cases}$$

$$E(y_i) = p_i$$

Función Logística

$$p_i = \frac{e^{z_i}}{(1 + e^{z_i})}$$

Se desea estimar la probabilidad de que un empleado abandone la compañía

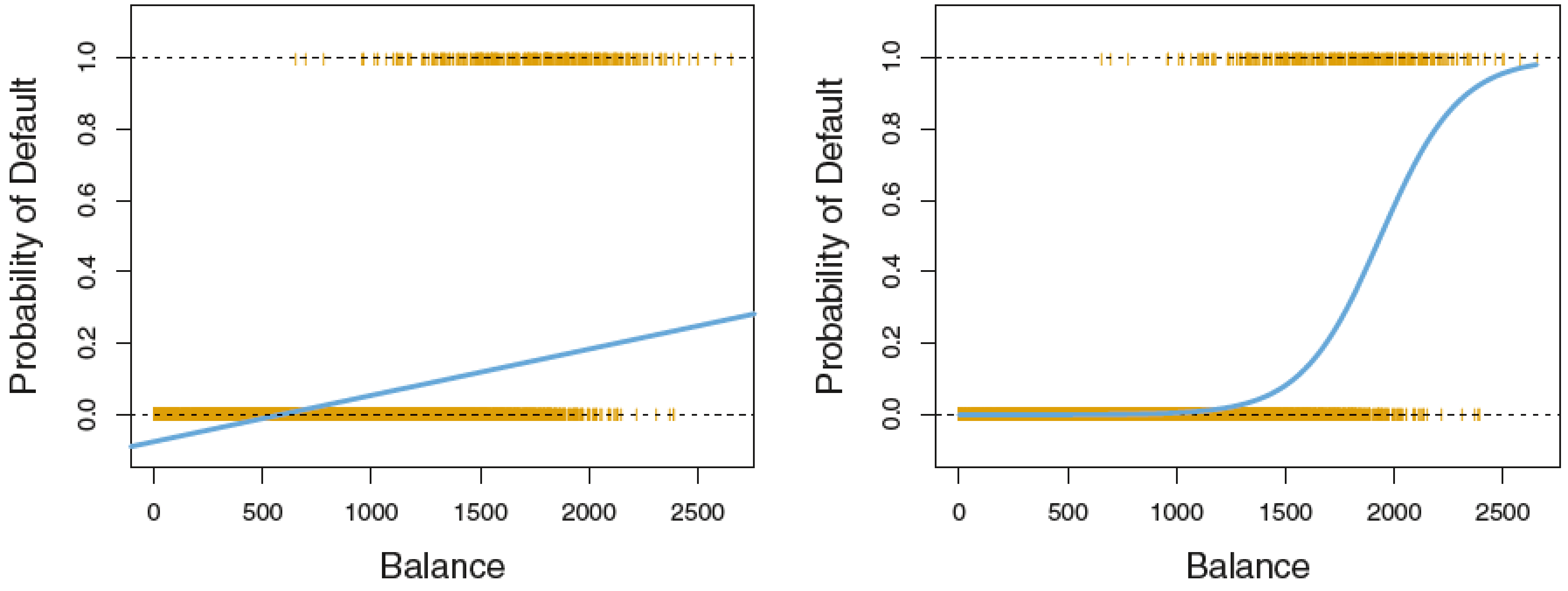
$$\beta_0 + \beta_1 x_i = z_i$$

Ejercicio 6

$$p_i = \frac{e^{z_i}}{(1 + e^{z_i})}$$

Hacer una grafica en R que permita comparar los valores de la función logística para para valores de z entre -10 y 10 (avanzando de 0.2 unidades)

Ejercicio 1



Probability of Default
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
Balance
Probability of Default
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
Balance

Función logística - Implementación

$$p_i = \frac{e^{z_i}}{(1 + e^{z_i})}$$

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i = z_i$$

Regresión Logística

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i = z_i$$

$$p_i = \frac{e^{z_i}}{(1+e^{z_i})}$$

Razón de chances (odss):

$$\frac{p_i}{1-p_i}$$

Por ejemplo, si la probabilidad de que ocurra un evento es del 80% entonces se espera que haya 4 veces más probabilidad de que ocurra que de que NO ocurra (se modela el logaritmo de la razón de chances)

$$\frac{0.8}{0.2} = 4$$

Regresión Logística

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i = z_i \qquad\qquad p_i = \frac{e^{z_i}}{(1+e^{z_i})}$$

Los odds y el logaritmo de odds cumplen que:

Si p(verdadero) = p(falso), entonces odds(verdadero) = 1
Si p(verdadero) < p(falso), entonces odds(verdadero) < 1
Si p(verdadero) > p(falso), entonces odds(verdadero) > 1
A diferencia de la probabilidad que no puede exceder el 1, los odds no tienen límite superior.
Si odds(verdadero) = 1, entonces logit(p) = 0
Si odds(verdadero) < 1, entonces logit(p) < 0
Si odds(verdadero) > 1, entonces logit(p) > 0
La transformación logit no existe para p = 0

Implementación del modelo

Muestras de test y muestras de entrenamiento

Las muestras de test son las muestras con las que vamos a verificar si el modelo tiene o no un buen desempeño

Las muestras de entrenamiento son las muestras con las que vamos a ajustar el modelo

Se suele tomar un 70% de los datos para entrenar el modelo y un 30% para probar su desempeño

Ajuste del modelo

Ajustar el modelo con los datos de entrenamiento, con la opción de enlace logístico en R

Establecer el punto de corte de la probabilidad (a partir de que punto se dice que es uno o cero)

Una vez estos datos estén ajustados, predecir la probabilidad en los datos de prueba

Validación del modelo

Matriz de confusión

Tabla de contingencia definida como:

		Datos de Prueba		
		Positivos	**Negativos**	
Predicción	**Positivos**	Verdaderos Positivos	Falsos Positivos	P'
	Negativos	Falsos Negativos	Verdaderos Negativos	N'
		P	N	TOTAL

Sensibilidad

Razón de verdaderos positivos, clasificar correctamente a un individuo que abandona la empresa:

$$VPR = \frac{VP}{P} = \frac{VP}{VP + FN}$$

		Datos de Prueba		
		Positivos	**Negativos**	
Predicciones	**Positivos**	VP	FP	P'
	Negativos	FN	VN	N'
		P	N	TOTAL

Especificidad

$$SPS = \frac{VN}{N} = \frac{VN}{FP + VN} = 1 - FPR$$

		Datos de Prueba		
		Positivos	Negativos	
Predicciones	Positivos	VP	FP	P'
	Negativos	FN	VN	N'
		P	N	TOTAL

Probabilidad de clasificar correctamente a un individuo que no abandona la empresa

Ratio

Razón de falsos positivos:

$$FPR = \frac{FP}{N} = \frac{FP}{FP + VN}$$

		Datos de Prueba		
		Positivos	Negativos	
Predicciones	Positivos	VP	FP	P'
	Negativos	FN	VN	N'
		P	N	TOTAL

Probabilidad de clasificar incorrectamente a un individuo que no abandona la empresa

Accuracy - Exactitud

$$ACC = \frac{VP + VN}{P + N}$$

		Datos de Prueba		
		Positivos	Negativos	
Predicciones	Positivos	VP	FP	P'
	Negativos	FN	VN	N'
		P	N	TOTAL

Valor predictivo positivo

$$PPV = \frac{VP}{VP + FP}$$

		Datos de Prueba		
		Positivos	Negativos	
Predicciones	Positivos	VP	FP	P'
	Negativos	FN	VN	N'
		P	N	TOTAL

Valor predictivo negativo

$$NPV = \frac{VN}{VN + FN}$$

		Datos de Prueba		
		Positivos	Negativos	
Predicciones	Positivos	VP	FP	P'
	Negativos	FN	VN	N'
		P	N	TOTAL

Curva ROC

Gráfica de la especificidad vs la sensibilidad a medida que varía el umbral de discriminación

A

VP=63	FP=28	91
FN=37	VN=72	109
100	100	200

VPR = 0.63

FPR = 0.28

ACC = 0.68

B

VP=77	FP=77	154
FN=23	VN=23	46
100	100	200

VPR = 0.77

FPR = 0.77

ACC = 0.50

C

VP=24	FP=88	112
FN=76	VN=12	88
100	100	200

VPR = 0.24

FPR = 0.88

ACC = 0.18

C′

VP=76	FP=12	88
FN=24	VN=88	112
100	100	200

VPR = 0.76

FPR = 0.12

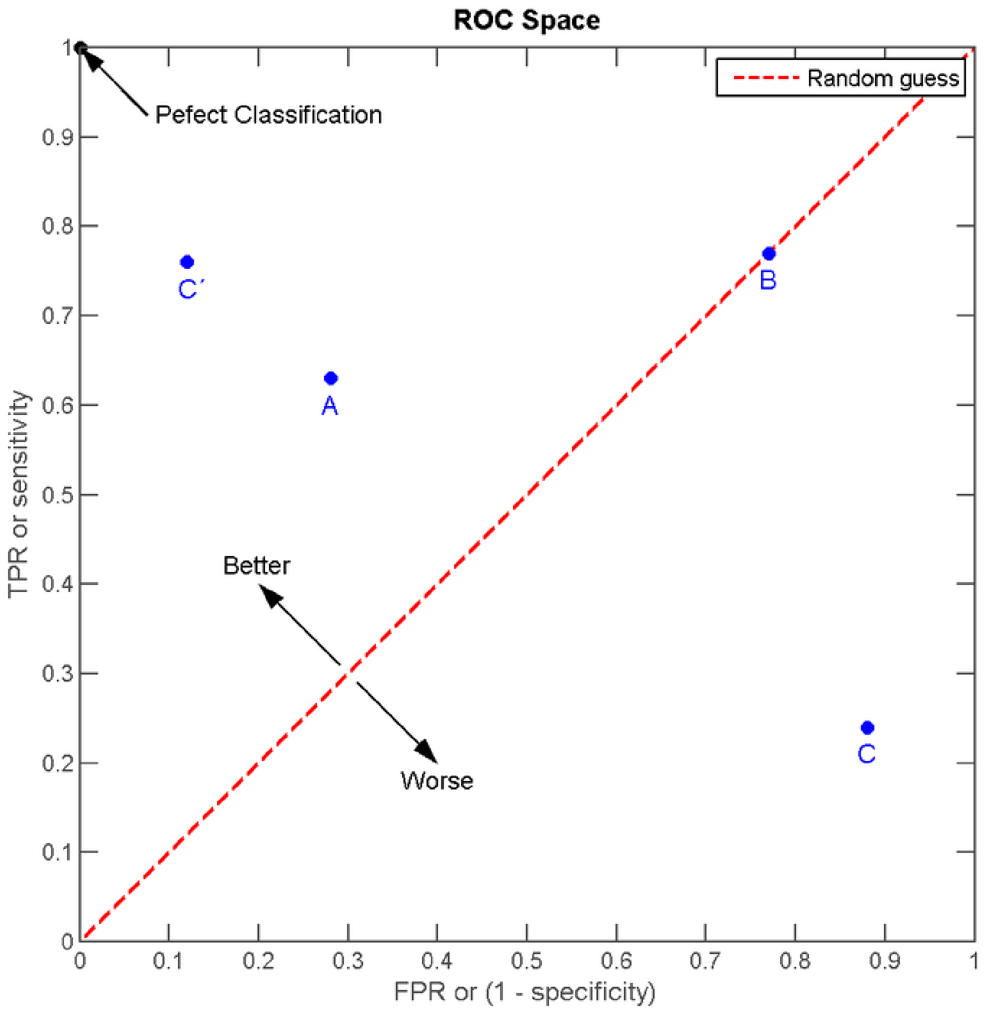
ACC = 0.82

Curva ROC

Dibujar la curva ROC consiste en poner juntos todos los puntos correspondientes a todos los umbrales o puntos de corte, de tal modo que ese conjunto de puntos se parecerá más o menos a una curva en el espacio cuadrado entre (0,0) y (1,1)

Area bajo la curva: Este índice se puede interpretar como la probabilidad de que un clasificador ordenará una instancia positiva elegida aleatoriamente más alta que una negativa. Cuanto mayor sea este valor, mejor es nuestra regresión

Curva ROC



ROC Space
Random guess
Pefect Classification
C´
B
A
Better
Worse
C
TPR or sensitivity
FPR or (1 - specificity)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Curva ROC

A modo de guía para interpretar las curvas ROC se han establecido los siguientes intervalos para los valores de AUC:

[0.5]: Es como lanzar una moneda.

[0.5, 0.6): Test malo.

[0.6, 0.75): Test regular.

[0.75, 0.9): Test bueno.

[0.9, 0.97): Test muy bueno.

[0.97, 1): Test excelente.

Regresión Logística

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i = z_i$$

$$p_i = \frac{e^{z_i}}{(1+e^{z_i})}$$

Razón de chances: Por ejemplo, si la probabilidad de que ocurra un evento es del 80% entonces se espera que haya 4 veces más probabilidad de que ocurra que de que NO ocurra

$$\frac{0.8}{0.2} = 4$$

Regresion Probit

$$\hat{y}_i = \Phi(\beta_0 + \beta_1 x_i)^{-1}$$

Discriminante lineal LDA

Teorema de Bayes – Clasificación

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A|B_j) = \frac{P(A \cap B)}{\sum_j P(B_j|A)}$$

Objetivo LDA

Hay dos o más grupos conocidos a priori, el objetivo es calcular:

$$P(Y = k \mid X = x)$$

Probabilidades a priori

La probabilidad a priori (conocimiento previo) de que que una observación pertenezca a cada una de las categorías

$$P(Y \ = \ k) = \pi_k$$

$$f_k(x) = P(X = x|Y = k)$$

Cuánto mayor sea el valor de $f_k(x)$, mayor será la probabilidad de que la observación pertenezca a la observación k

Probabilidad a posteriori – Clasificación Bayesiana

La probabilidad de que una observación pertenezca a k siendo x el valor del predictor

$$P(Y = k|X = x) = \frac{\pi_k P(X = x|Y = k)}{\sum_i \pi_i P(X = x|Y = i)}$$

$$= \frac{\pi_k f_k(x)}{\sum_i \pi_i f_i(x)}$$

La clasificación con menor error se consigue asignando la observación que maximice la probabilidad a posteriori

Estimadores

$$\widehat{\pi}_i = \frac{n_k}{N}$$

Ho: Las observaciones se distribuyen de forma normal

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2\sigma_k^2}(x-\mu)^2}$$

Ejercicio 1

Generar una muestra aleatoria para 2 clases distribuidas de forma normal.

Graficar un histograma donde se puedan ver las dos variables

Características:

Muestra 1: n = 280, media = 10, desviación = 2

Muestra 2: n = 320, media = 17, desviación = 2

Probabilidad a posteriori

$$P(Y=k|X=x)=\frac{\frac{1}{\sqrt{2\pi\sigma_k}}e^{-\frac{1}{2\sigma_k^2}(x-\mu)^2}\pi_k}{\sum_i \pi_j \frac{1}{\sqrt{2\pi\sigma_k}}e^{-\frac{1}{2\sigma_k^2}(x_i-\mu_i)^2}\pi_k}$$

¿Qué pasa si todas las varianzas son iguales?

Método

$$\delta_k = \ln(Y = k|X = x)$$

$$\delta_k = -\frac{1}{2\sigma^2}\left(x^2 - 2x\mu_k + \mu_k^2\right) + \ln(\pi_k) - \ln(cte)$$

Luego,

$$\delta_k \propto \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \ln(\pi_k)$$

Estimadores

$$\widehat{\pi_i} = \frac{n_k}{N}$$

$$\widehat{\mu_k} = \frac{1}{n_k} \sum_{i \in y_k} x_i$$

$$\hat{\sigma} = \frac{1}{n-k} \sum_{j}^{k} \sum_{i \in y_k}^{n_k} (x_i - \widehat{\mu_k})^2$$

Ejemplo – 2 clases

X se asigna a la clase 1 con probabilidad $P(Y = 1|X = x)$

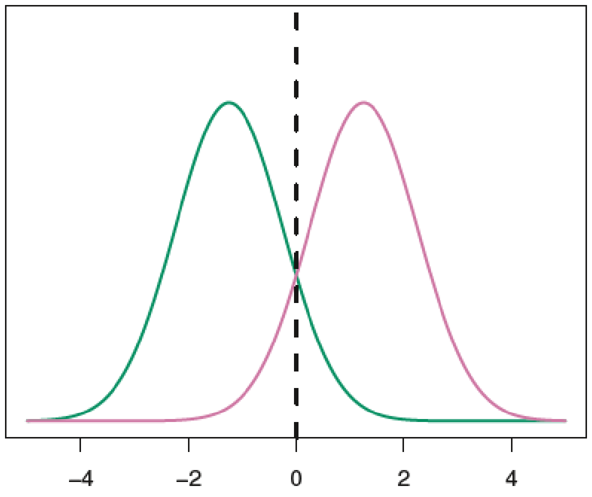
X se asigna a la clase 2 con probabilidad $P(Y = 2|X = x)$

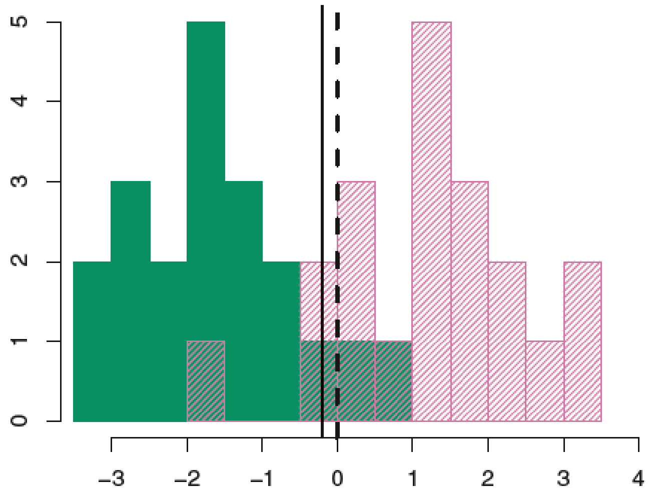
$$\frac{P(Y = 1|X = x)}{P(Y = 2|X = x)} > 1$$

Límite de decisión

Asignar a la observación 1 si:

$$x > \frac{\mu_1 + \mu_2}{2}$$





Ejercicio 2

Graficar la línea discriminante para las dos clases

```
geom_vline(xintercept =  valor_línea , linetype = "longdash")
```

Algoritmo de clasificación LDA

1. Ajustar $\hat{\delta}_k(x)$ para k = 1, 2,..., c

$$\hat{\delta}_k(x) > \frac{x\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\mu_k^2}{2\hat{\sigma}^2} + \ln(\hat{\pi}_k)$$

2. Buscar el valor de k dónde $\hat{\delta}_k(x)$ sea máximo
3. Asignar la observación a dicha categoría

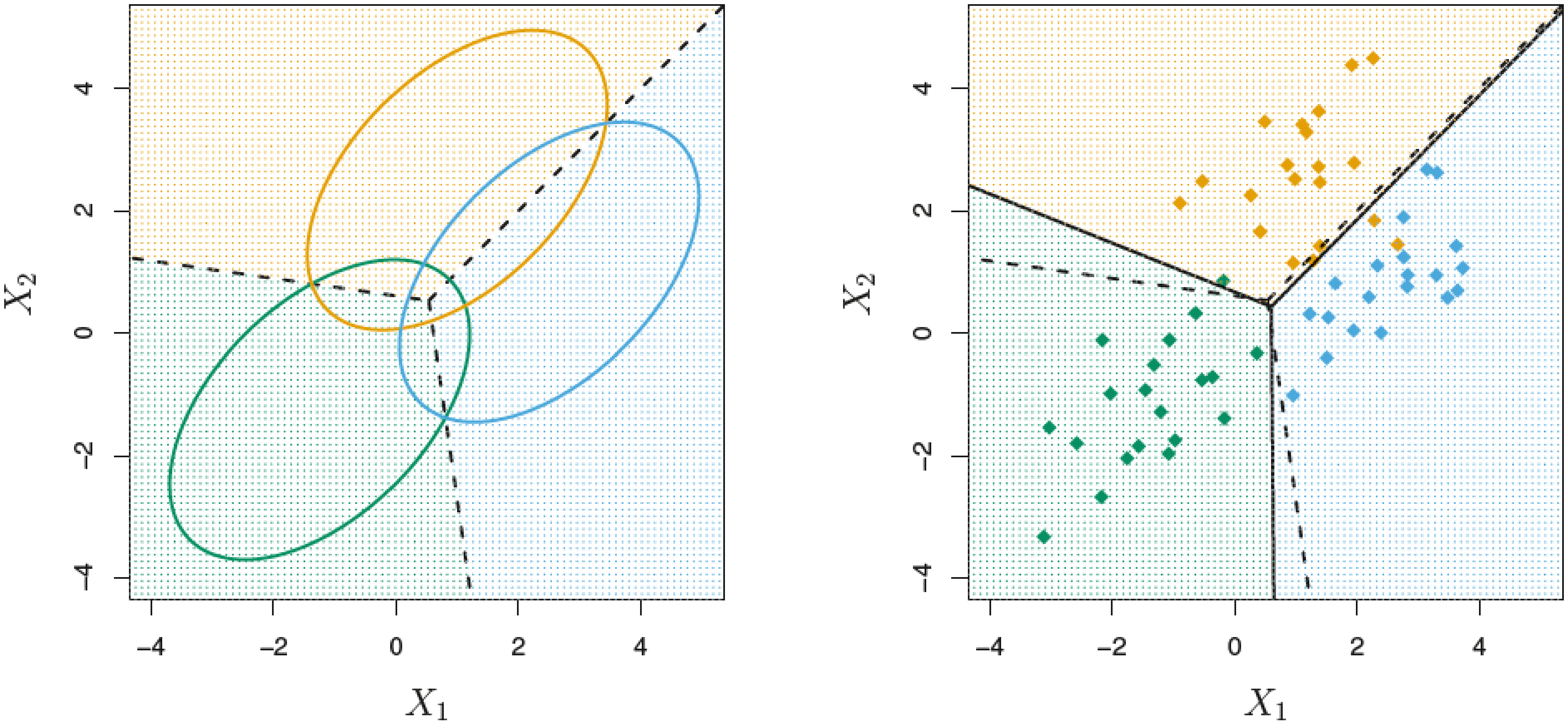
Nota

Se denomina análisis discriminante lineal porque $\hat{\delta}_k(x)$ es lineal para el valor de x

$$\frac{x\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\mu_k^2}{2\hat{\sigma}^2} + \ln(\hat{\pi}_k)$$

$$a = \frac{x\hat{\mu}_k}{\hat{\sigma}^2}, \qquad b = \frac{\mu_k^2}{2\hat{\sigma}^2} + \ln(\hat{\pi}_k)$$

Nota



Caso multivariado

1. Ajustar $\hat{\delta}_k(x)$ para k = 1, 2,…, c

$$\hat{\delta}_k(x) = \mu_k' \Sigma^{-1} X - \frac{1}{2} \mu_k' \Sigma^{-1} \mu_k + \ln(\pi_k)$$

2. Buscar el valor de k dónde $\hat{\delta}_k(x)$ sea máximo
3. Asignar la observación a dicha categoría

Nota

Se supone que todas las clases tienen homogeneidad de varianzas, la estimación corresponde a la matriz de varianzas y covarianzas

Análisis discriminante cuadrático QDA

QDA

Funciona de la misma forma que el LDA, la única diferencia es que en el QDA no hay homogeneidad de las varianzas, luego la función discriminante toma la forma:

$$\hat{\delta}_k(x) = \mu_k' \Sigma_{\mathrm{k}}^{-1} X - \frac{1}{2} \mu_k' \Sigma_{\mathrm{k}}^{-1} \mu_k + \ln(\pi_k)$$