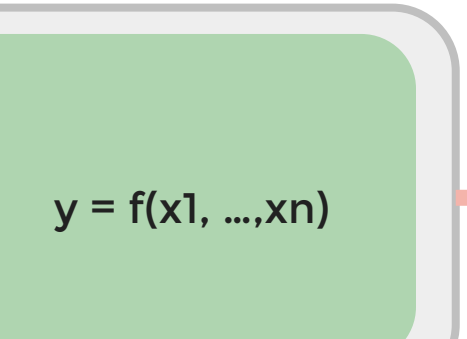
Modelos de clasificación

Fernanda Mora
10 octubre, 2016

Diplomado de Minería de Datos
para soporte a la toma de decisiones

Contenido

- Regresión logística
  - Ejercicio práctico 1
- Análisis discriminante
  - Ejercicio práctico 2



Regresión logística

Preliminares

- Regresión lineal: estimamos la esperanza condicional de una variable continua Y
- Las variables de soporte podían ser categóricas o discretas
- ¿Qué pasa si Y no es continua?
- Caso más común es Y es binaria:
  - ¿El paciente tendrá diabetes?: Y=1 sí, Y=0 no
  - ¿Este cliente pagará el crédito o no?
  - ¿Este producto tendrá éxito en el mercado o no?
- La variable *Y* puede tener muchas categorías

Preliminares

- Al problema de modelar una variable categórica Y en términos de otras variables X_1, X_2, ...X_k se le llama problema de clasificación
- Otra manera de verlo es preguntarnos cómo le hacemos para, dadas las variables de soporte X_1, X_2, ...X_k, asignarle una etiqueta a una variable objetivo *Y*
- Las clases pueden verse como etiquetas
- Los modelos de clasificación pertenecen la subárea de machine learning conocida como aprendizaje supervisado (supervisado significa que tengo un conjunto de datos etiquetados)

Preliminares

- ¿Cómo generar un modelo que prediga esas etiquetas/clases?
- Primer approach: hacer un modelo basado en reglas
  - Si estudio 5 días a la semana, y
  - Si hago más de la mitad de las tareas, y
  - Si apruebo los exámenes parciales, y
    - Entonces aprobaré la materia
- ¿Cómo generar esas reglas?
- ¡Crudo, difícil y no escalable! ¿Qué pasa si agrego una nueva variable? ¿Si los datos cambian?

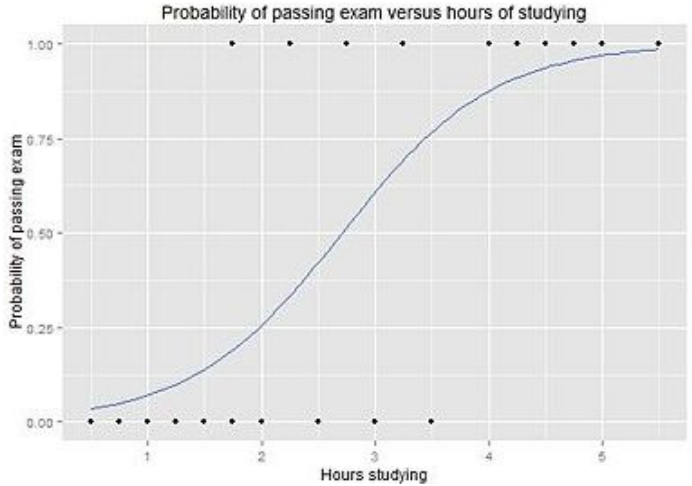
Preliminares

- Mejor approach: modelo que arroje una probabilidad/score:
  - La probabilidad de aprobar es de 60% si estudias 5 días a la semana, si haces más de la mitad de las tarea, si apruebas los exámenes parciales.
- Mejor aún: modelo que arroje $Prob(Y \mid X_1,\ldots,X_k)$
  - Aprobar (sí/no) dado cierto nivel de estudio, número de tareas y número de exámenes aprobados.
- Nos permitirá estimar con cierto grado de confianza
  - $Prob(Y \mid X_1,\ldots,X_k)=0.55$ vs $Prob(Y \mid X_1,\ldots,X_k)=0.98$

Ejemplo

- Estudiantes: Aprobado/no aprobado vs Horas de estudio

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1



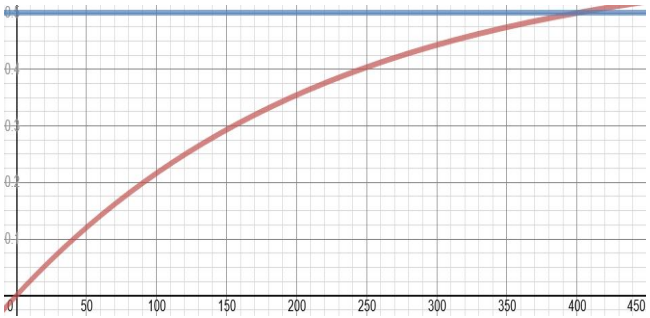
Hours of study	Probability of passing exam
1	0.07
2	0.26
3	0.61
4	0.87
5	0.97

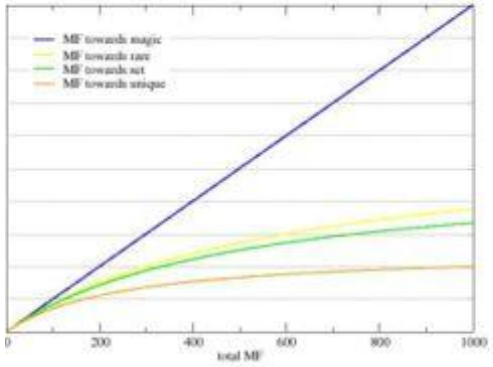
Planteamiento

- Dos clases: 0 y 1 (ausencia o presencia de atributo)
- La variable objetivo *Y* se vuelve una variable indicadora: Y=1 ó Y=0
- La variable objetivo *Y* se puede modelar como una variable aleatoria Bernoulli(p) y entonces Pr(Y=1) = E(Y)
- Pero queremos que *Y* dependa de $X_1,...X_k$
- Si condicionamos al valor de las variables soporte $X_1,...X_k$, entonces Pr(Y=1 | X=x) = E(Y | X=x)
- ¿Para qué nos sirve esto?

Planteamiento

- Queremos modelar Pr(Y=1 | X=x) = E(Y | X=x) = p(x)
- 3 propuestas:
  1. Usar una aproximación lineal a p(x) -> Problema: queremos que p(x) esté entre 0 y 1 y la regresión lineal no nos garantiza eso. Además queremos que p(x) crezca a tasas decrecientes

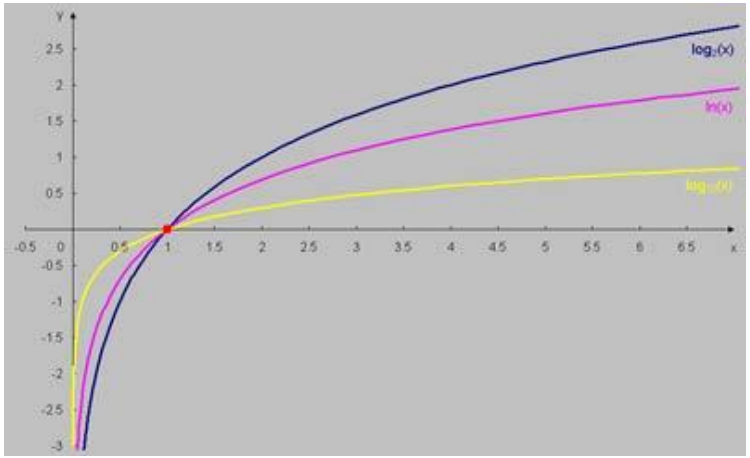




Planteamiento

- Queremos modelar Pr(Y=1 | X=x) = E(Y | X=x) = p(x)
- 3 propuestas:

  2. Usar log(p(x)) y modelarlo como una función lineal. Problema: el soporte (la probabilidad) no está entre [0,1]



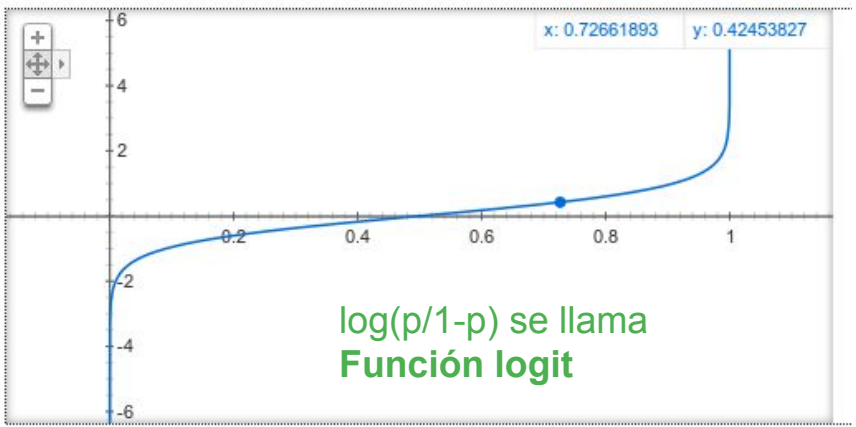
Planteamiento

- Queremos modelar $Pr(Y=1 \mid X=x) = E(Y \mid X=x) = p(x)$
- 3 propuestas:

  3. Usar $\log(p/(1-p))$. Ventajas: se puede modelar como una función lineal (¡sabemos usar regresión!)

El dominio está entre [0,1] -> podemos modelar probabilidad/score

El rango está entre [-infinito, +infinito] -> podemos usar regresión lineal



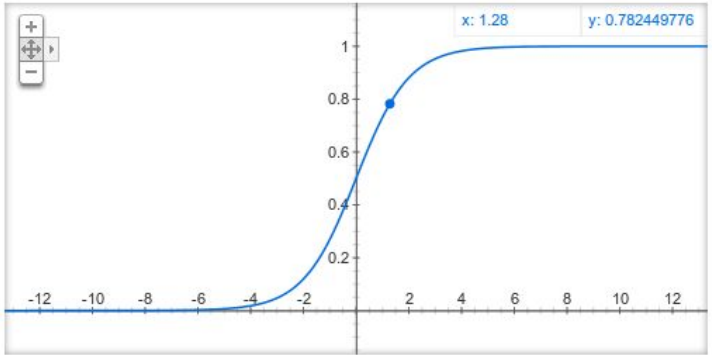
Regresión logística

- El modelo de regresión logística es:
  - $\log(p(x)/(1-p(x)) = \beta_0 + x\beta$
  - Entonces despejando p(x):

1/(1+e^-x)

Betas son cero

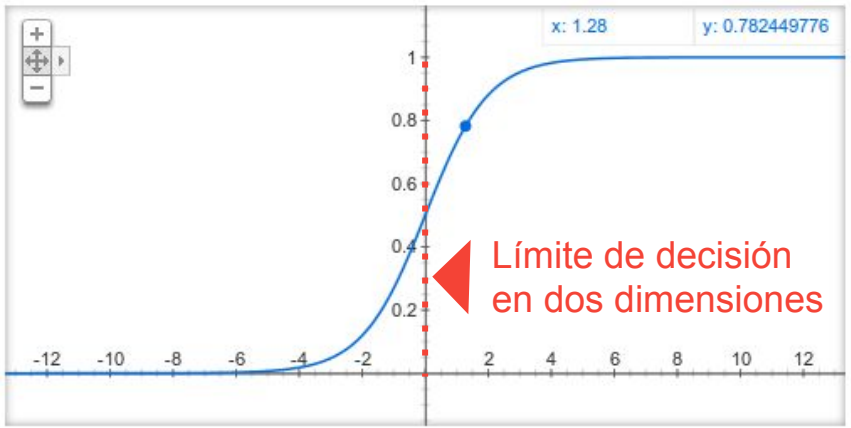
$$p(x) = \frac{1}{1 + e^{-(B_0 + B_1 x)}} = \frac{e^{B_0 + B_1 x}}{1 + e^{B_0 + B_1 x}}.$$



- Probabilidad puede verse como un Score
- Recordar esta función, la veremos después para redes neurales y se llamará perceptrón de una capa

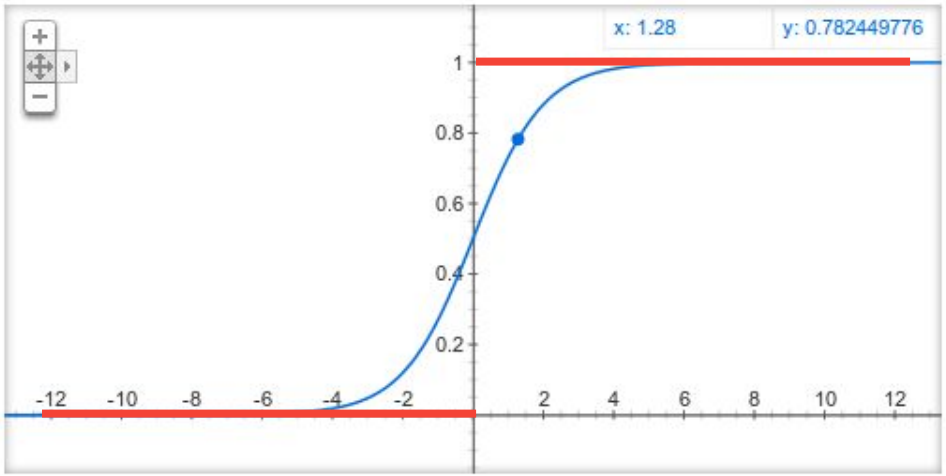
Regresión logística

- Para minimizar la tasa de error de clasificación, podríamos predecir
  - $Y=1$ si $p>=0.5$ y $Y=0$ si $p<0.5$
  - O sea $Y=1$ si $\beta_0 + x\beta > 0$ y $Y=0$ si $\beta_0 + x\beta <= 0$
    - $\beta_0 + x\beta = 0$ define el límite de decisión



Regresión logística

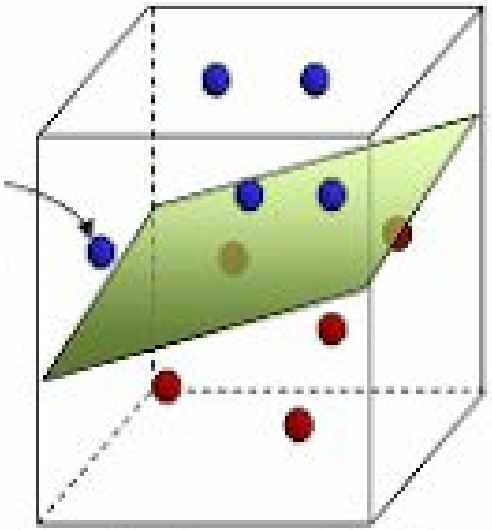
- ¿Por qué este enfoque es mejor que el enfoque tradicional?



- El enfoque tradicional sólo toma dos posibles valores: 0 ó 1. Es más crudo
- Este enfoque puede tomar valor entre 0 y 1. Es más flexible y hace predicciones más detalladas

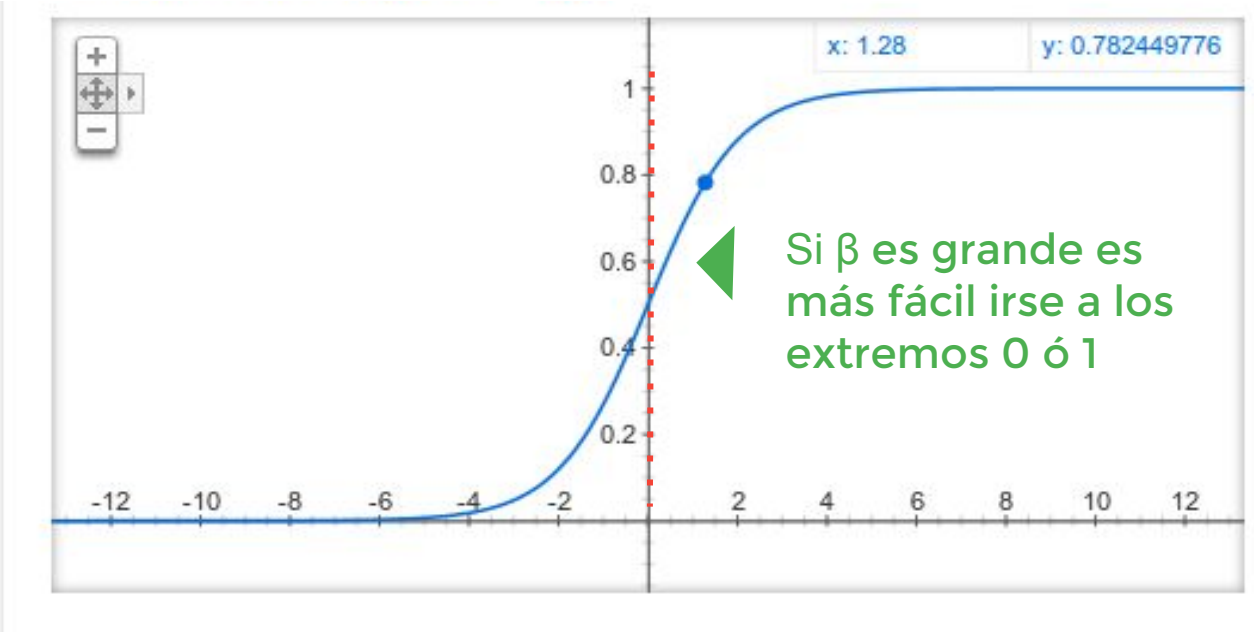
Regresión logística

- Límite de decisión $\beta_0 + x\beta = 0$:
  - Recta en dos dimensiones
  - Plano en tres dimensiones
  - Hiperplano en más dimensiones



Regresión logística

- La probabilidad/score dependerá de la distancia al límite de decisión $\beta_0 + x\beta = 0$



Regresión logística

- La regresión logística es uno de los modelos de clasificación más usados por:
  - Muchas veces funciona sorprendentemente bien como clasificador, pero no siempre
  - Es parte de modelos más generales (modelos lineales generalizados)
  - Se ha usado mucho en la práctica: tradición
- Pero así como existen muchos modelos para el problema de regresión (i.e. explicar una variable cuantitativa en términos de otras) y sólo vimos el lineal, en el problema de clasificación existen muchos otros modelos (veremos otros)

Regresión logística

- Nuevamente queremos estimar los parámetros

$$p(x;b,w) = \frac{e^{\beta_0 + x \cdot \beta}}{1 + e^{\beta_0 + x \cdot \beta}} = \frac{1}{1 + e^{-(\beta_0 + x \cdot \beta)}}$$

¿Cuánto valen las betas?

- No es trivial hacerlo
- Se construye una función de log-verosimilitud y se aproxima numéricamente a la solución que la minimiza
- Se generaliza al caso en múltiples dimensiones (i.e. k variables de soporte): regresión logística multinomial

Regresión logística

- Interpretación:

$$\frac{p(x)}{1-p(x)} = e^{B_0+B_1 x}$$

  - Odds en favor del éxito
  - También se llaman momios
  - Los **momios** se usan para expresar los riesgos relativos (por eso se expresan como un cociente)

- Supongamos que p(50) = ⅔, entonces

$$\frac{p(50)}{1-p(50)} = \frac{\frac{2}{3}}{1-\frac{2}{3}} = 2.$$

  - Un éxito es dos veces más probable que una falla cuando x=2

- Si $\frac{p(x)}{1-p(x)} < 1$, el fracaso (denominador) tiene menor probabilidad que el éxito (numerador)
- Si $\frac{p(x)}{1-p(x)} > 1$ el éxito tiene mayor probabilidad que el fracaso

Regresión logística

- También podemos comparar odds entre individuos:
  - Se compara la situación de la observación i con la de la observación j (que suele ser la de referencia)
  - El cociente entre odds mide cuanto es más probable que se dé el éxito en el individuo i que en el individuo j

$$\text{Cociente entre odds} = \frac{\frac{M_i}{(1-M_i)}}{\frac{M_j}{(1-M_j)}} = \frac{e^{\alpha+\beta_k x_{ki}}}{e^{\alpha+\beta_k x_{kj}}} = e^{\beta_k (x_{ki}-x_{kj})}$$

Regresión logística

- Lo anterior nos sirve para interpretar los coeficientes beta:
  - Factor de cambio en el cociente entre odds cuando el valor de la variable X_k aumenta en una unidad y el resto de variables explicativas se mantienen constantes.

$$Cociente\,entre\,Odds = \frac{\frac{M_{i+1}}{(1-M_{i+1})}}{\frac{M_i}{(1-M_i)}} = \frac{e^{\alpha+\beta_k(x_{ki}+1)}}{e^{\alpha+\beta_k x_{ki}}} = e^{\beta_k(x_{ki}+1-x_{kj})} = e^{\beta_k}$$

  - β_k se interpreta como el número de veces que incrementa el logaritmo del éxito frente al fracaso cuando la variable X_k incrementa en una unidad
  - Qué tanto impacto tiene en *Y*, un cambio en X_k

Regresión logística

Recordemos α y β de regresión lineal, definían tipos de error:

	Null hypothesis (H_0) is valid: Innocent	**Null hypothesis (H_0) is invalid: Guilty**
Reject H_0 **I think he is guilty!**	Type I error False positive Convicted! α Inocente en la cárcel	Correct outcome True positive Convicted!
Don't reject H_0 **I think he is innocent!**	Correct outcome True negative Freed!	Type II error False negative Freed! β Delincuente en las calles

Regresión logística

- Tenemos una tabla muy parecida para problemas de clasificación
- ¿Qué pasa si mi modelo dice que es de tipo 1 pero realmente no lo era?

Realidad

Predicción

	Realmente es 0 (inocente)	Realmente es 1 (culpable)
Predigo 1 (es culpable)	Queremos tener pocas observaciones aquí Type I error False positive Convicted! Inocente en la cárcel	Correct outcome True positive Convicted! Delincuente en la cárcel
Predigo 0 (es inocente)	Correct outcome True negative Freed! Inocente en las calles	Queremos tener pocas observaciones aquí Type II error False negative Freed! Delincuente en las calles

Regresión logística

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	**Specificity** = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

Ejercicio práctico 1

Análisis discriminante

Motivación

- Vimos cómo clasificar una variable discreta *Y* usando un conjunto de variables X_1, ...,X_k
- Es decir, vimos un modelo para predecir clases/categorías a partir de variables de soporte, i.e. un modelo de clasificación
- Asumimos que X_1, ...,X_k nos ayudan a explicar Y
- Demos un paso atrás: ¿realmente X_1, ...,X_k nos ayudan a explicar Y?
- Es decir, ¿realmente el nivel educativo, el ingreso, el sexo, la edad, el estado civil me ayudan a predecir si un acreditado pagará?

Motivación

- ¿Qué tan diferentes son estas variables dentro de los grupos: sí pagará y no pagará?
- ¡Esperamos que sean diferentes!
- Ejemplo:
  - No pagará: ingreso bajo, joven, sin educación, casado
  - Sí pagará: ingreso alto, adulto maduro, profesionista, soltero
- ¿Será así? ¿qué tan diferentes?
- ¿Cómo saber?

Análisis discriminante

- Si son muy diferentes entre los grupos definidos por *Y* decimos que discriminan
- Análisis discriminante: describir (si existen) diferencias significativas entre *g* grupos de individuos sobre los que se observan *p* variables
- Una vez encontradas (si las hay), explicarlas
- Construir un modelo que me permita clasificar nuevos individuos
- ¡Análisis discriminante -igual que regresión logística- también es un modelo de clasificación!

Análisis discriminante

- Si son muy diferentes entre los grupos definidos por Y decimos que discriminan
- Análisis discriminante: describir (si existen) diferencias significativas entre *g* grupos de individuos sobre los que se observan *p* variables
- Una vez encontradas (si las hay), explicarlas
- Construir un modelo que me permita clasificar nuevos individuos
- ¡Análisis discriminante -igual que regresión logística- también es un modelo de clasificación!

Análisis discriminante

- Modelo debido al estadístico y biólogo Fisher (1890 - 1962)
- AD: Identificar las características que diferencian a dos o más grupos y a crear una función capaz de distinguir con la mayor precisión posible a los miembros de uno u otro grupo
- Análisis discriminante supone que las variables de soporte X_1, ...,X_k son normales, cuantitativas y continuas
  - Si no son normales a veces es posible transformar las variables para que lo sean
  - Si nos es posible hacerlas normales mejor usar regresión logística
- Sin embargo, cuando se cuenta con pocos individuos regresión logística predice pobremente

Análisis discriminante

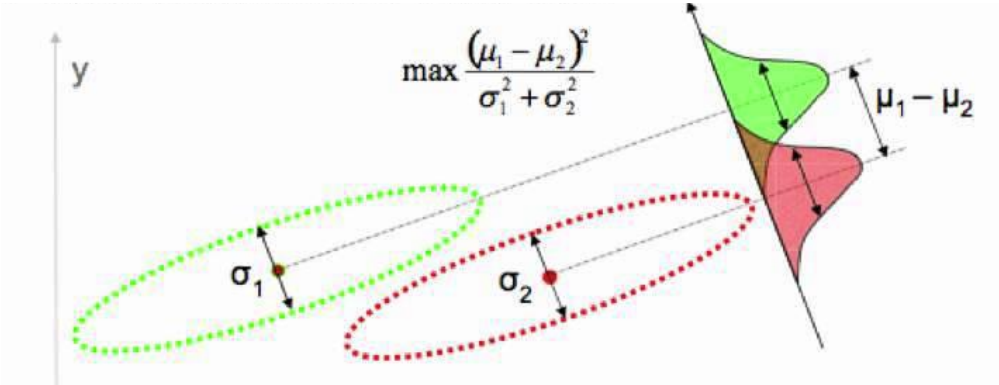
- El objetivo es llegar a una función discriminante
- Hay varios tipos de análisis de discriminante que dependen de la función discriminante
- ¿Uso regresión logística o análisis discriminante?
  - Si tengo pocos datos, estoy en problemas. Intentar análisis discriminante (regresión logística es inestable)
  - Si las variables de soporte son cuantitativas y normales y usar análisis discriminante
  - Si las variables de soporte no son normales ó no son todas cuantitativas entonces usar regresión logística

Análisis discriminante

- Busco encontrar una combinación lineal de las variables de soporte que me ayude a separar las clases
- ¡Igual que regresión logística será un clasificador lineal!
  - En regresión logística teníamos: Y=1 si $\beta_0 + x\beta > 0$ y Y=0 si $\beta_0 + x\beta <= 0$
- ¿Cómo separo en análisis discriminante?

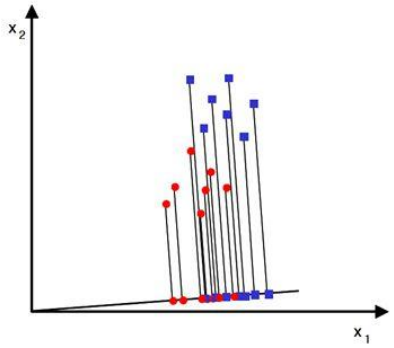
Análisis discriminante

- Buscaremos un nuevo eje de coordenadas tal que al recolocar los puntos:
  - Haya una máxima separación entre las medias de los grupos
  - La varianza sea mínima en cada grupo

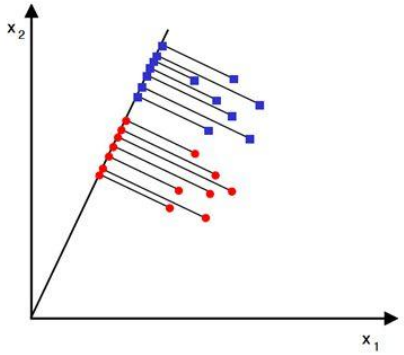


Análisis discriminante

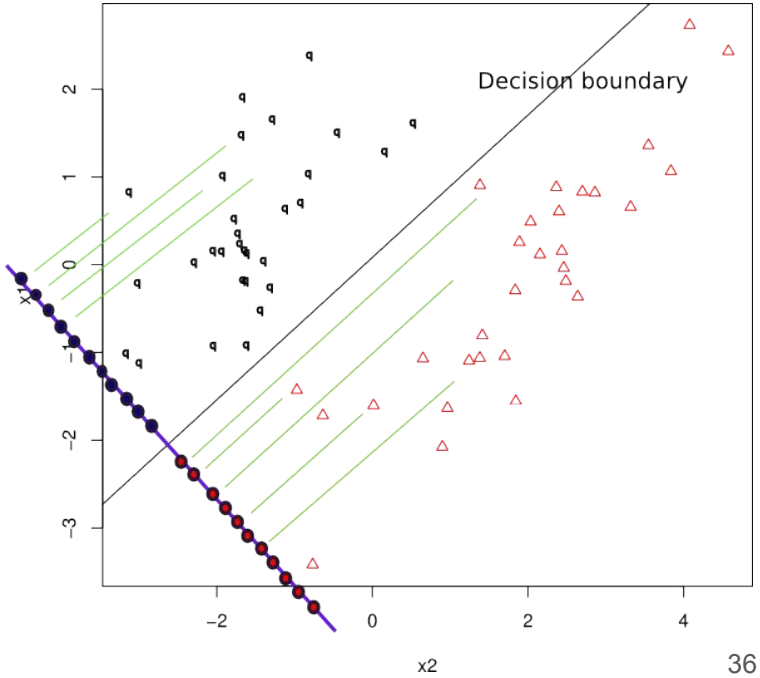
- ¿Por qué queremos un nuevo eje de coordenadas?



A projection with non-ideal separation



A projection with ideal separation



Ejercicio práctico 2

Gracias