

Machine Learning

Hernán Aguirre

Universidad San Francisco de Quito

Universidad de Shinshu, Japón

Contenido del Curso

1. Introducción
2. Aprendizaje estadístico
3. Regresión lineal
4. Clasificación
5. Métodos de remuestreo
6. Selección y regularización de modelos lineales
7. Más allá de la linealidad
8. Métodos basados en árboles
9. Máquinas de vectores de soporte
10. Aprendizaje profundo
11. Análisis de supervivencia y datos censurados
12. Aprendizaje sin supervisión
13. Pruebas múltiples

4 Classification

1. Una descripción general de la clasificación
2. ¿Por qué no la regresión lineal?
3. Regresión logística
4. Modelos generativos de clasificación
5. Una comparación de métodos de clasificación
6. Modelos lineales generalizados
7. Lab: Regresión logística, LDA, QDA y KNN
8. Ejercicios

Clasificación

- Respuesta Y es cualitativa
- Predecir una respuesta cualitativa para una observación se denomina *clasificar esa observación*
 - → implica asignar la observación a una categoría o clase
- Como base para hacer la clasificación, a menudo, primero se predice la probabilidad de que la observación pertenezca a cada una de las categorías de una variable cualitativa
- En este sentido también se comportan como métodos de regresión

Clasificadores Ampliamente Utilizados

- Regresión logística,
- Análisis discriminante lineal,
- Análisis discriminante cuadrático,
- Bayes ingenuo
- K-vecinos más cercanos

Métodos de Clasificación más Intensivos en Computación

- Modelos aditivos generalizados
- Arboles,
- Bosques aleatorios
- Boosting (impulso)
- Máquinas de vectores de soporte

Descripción General de la Clasificación

- Los problemas de clasificación ocurren con frecuencia, quizás incluso más que los problemas de regresión
- Una persona llega a la sala de emergencias con *una serie de síntomas que* posiblemente podrían atribuirse a una de tres condiciones médicas. *¿Cuál de las tres condiciones tiene el individuo?*
- Un servicio de banca en línea debe poder determinar si una *transacción* que se realiza en el sitio es *fraudulenta o no*, basándose en la dirección IP del usuario, el historial de transacciones anteriores, etc.
- A partir de los datos de la secuencia de ADN de varios pacientes con y sin una determinada enfermedad, a un biólogo le gustaría *determinar qué mutaciones del ADN son perjudiciales y cuáles no*.

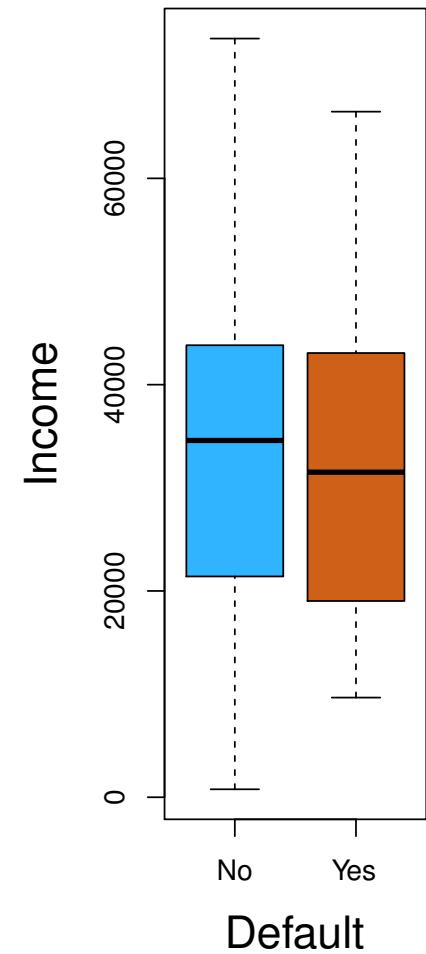
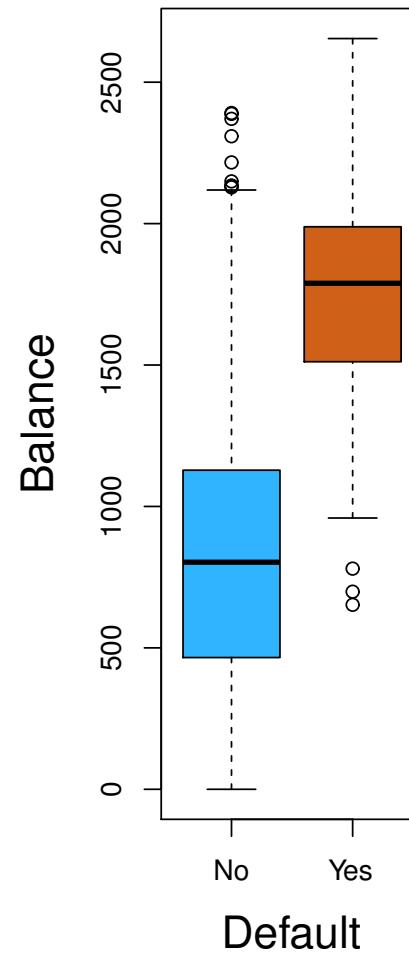
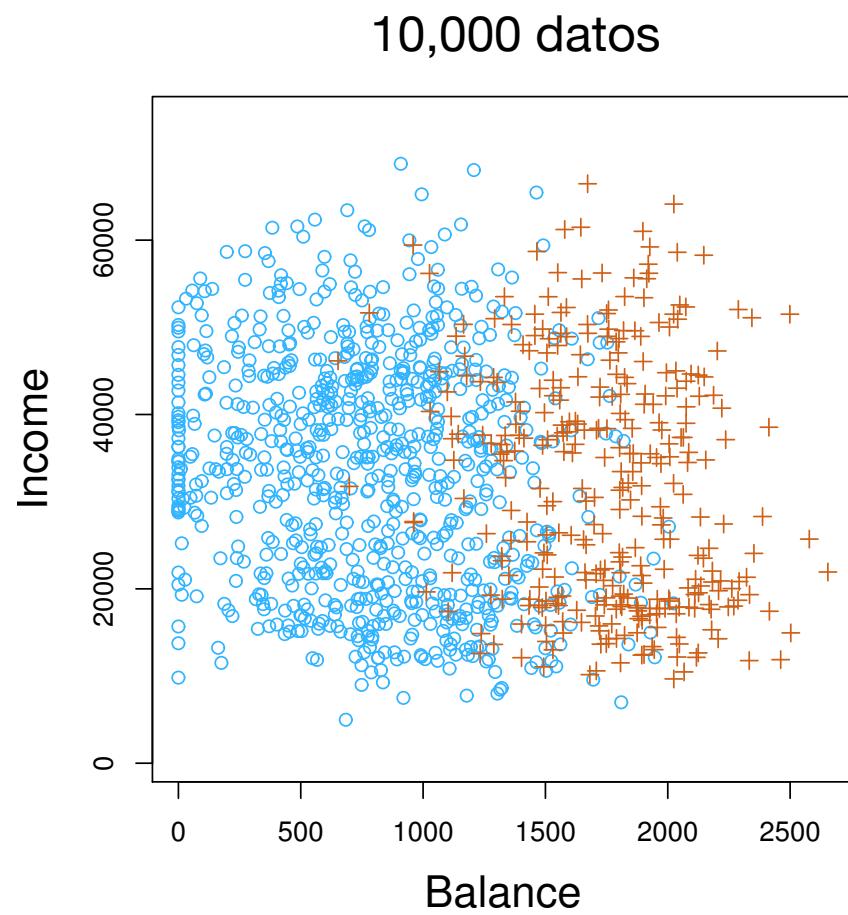
Entorno de Clasificación

Datos $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ $y_i \in \{c_1, \dots, c_k\}$

X	y	Learning	Predicción	Error
x_1	y_1		\hat{y}_1	e_1
x_2	y_2	Clasificador	\hat{y}_2	e_2
\vdots	\vdots		\vdots	\vdots
x_n	y_n		\hat{y}_n	e_n

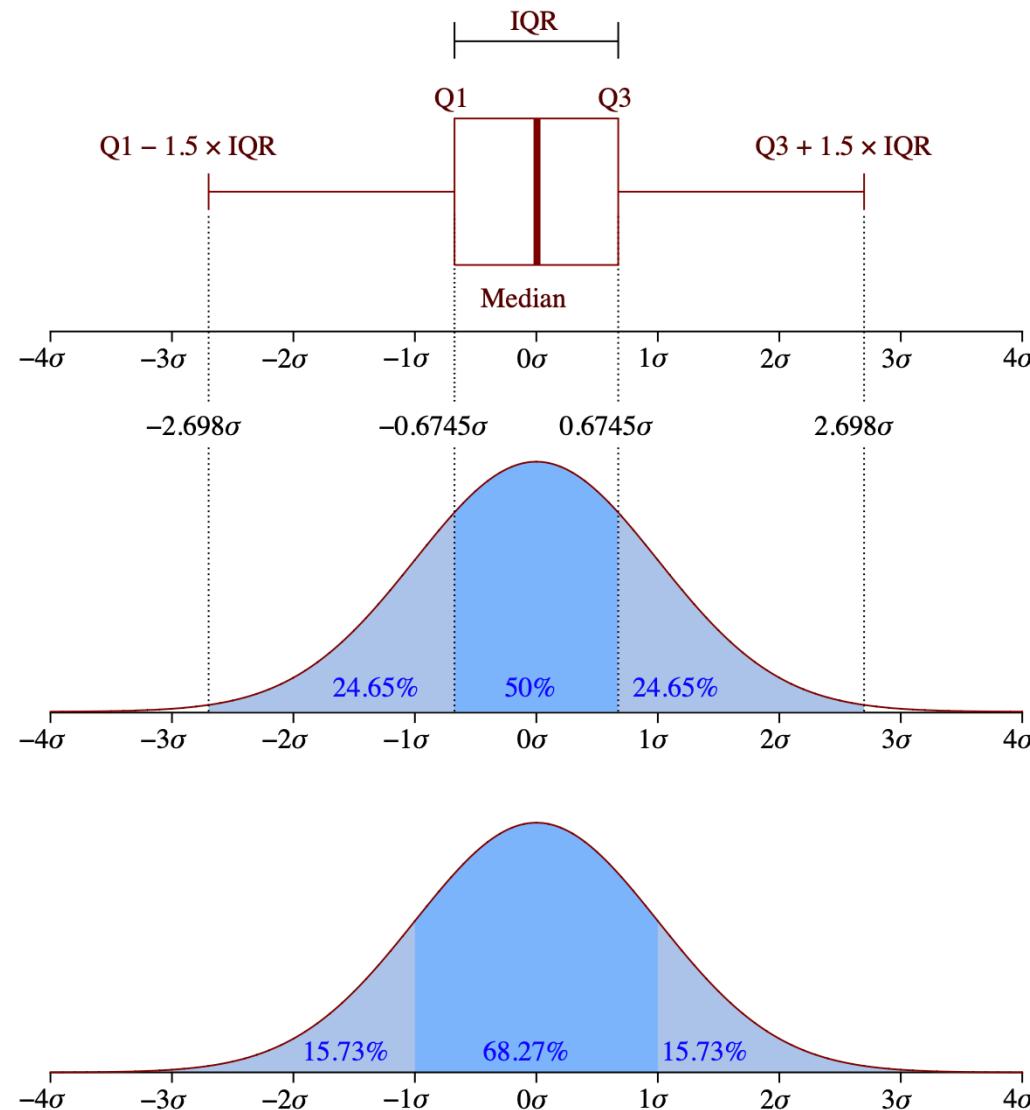
- Queremos que el clasificador funcione bien tanto en los *datos de entrenamiento*, como en las *observaciones de prueba que no se utilizaron para entrenar* el clasificador.

Incumplimiento de Pago



Predecir si un individuo incumplirá el pago de su tarjeta de crédito, sobre la base del ingreso anual y el saldo mensual de la tarjeta de crédito

Box Plot



¿Por qué no la Regresión Lineal?

- Un método de regresión no puede acomodar una respuesta cualitativa con más de dos clases
- Un método de regresión no proporcionará estimaciones significativas de $Pr(Y|X)$, incluso con sólo dos clases
- → es preferible utilizar un método de clasificación que sea realmente adecuado para valores de respuesta cualitativos

Ejemplo

- Predecir la condición médica de un paciente en la sala de emergencias en función de sus síntomas.
- Tres diagnósticos posibles: accidente cerebrovascular, sobredosis de drogas y ataque epiléptico.

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

$$Y = \begin{cases} 1 & \text{if } \texttt{epileptic seizure}; \\ 2 & \text{if } \texttt{stroke}; \\ 3 & \text{if } \texttt{drug overdose}, \end{cases}$$

Observaciones

- Si los valores de la variable de respuesta adoptaran un orden natural, como *leve, moderado y grave*, y pensáramos que la brecha entre leve y moderado era similar a la brecha entre moderado y grave, entonces una codificación 1, 2, 3 sería razonable.
- Desafortunadamente, en general no existe una forma natural de convertir una variable de respuesta cualitativa con más de dos niveles en una respuesta cuantitativa que esté lista para la regresión lineal.

Observaciones (2)

- Para una respuesta cualitativa binaria, la situación es mejor. Por ejemplo: accidente cerebrovascular y sobredosis de drogas. Entonces podríamos utilizar el enfoque de variable ficticia (*dummy*).
- Luego podríamos ajustar una regresión lineal a esta respuesta binaria y predecir
 - sobredosis de drogas si $\hat{Y} > 0.5$
 - accidente cerebrovascular en caso contrario.
- En el caso binario, incluso si invertimos la codificación anterior, la regresión lineal producirá las mismas predicciones finales.

Observaciones (3)

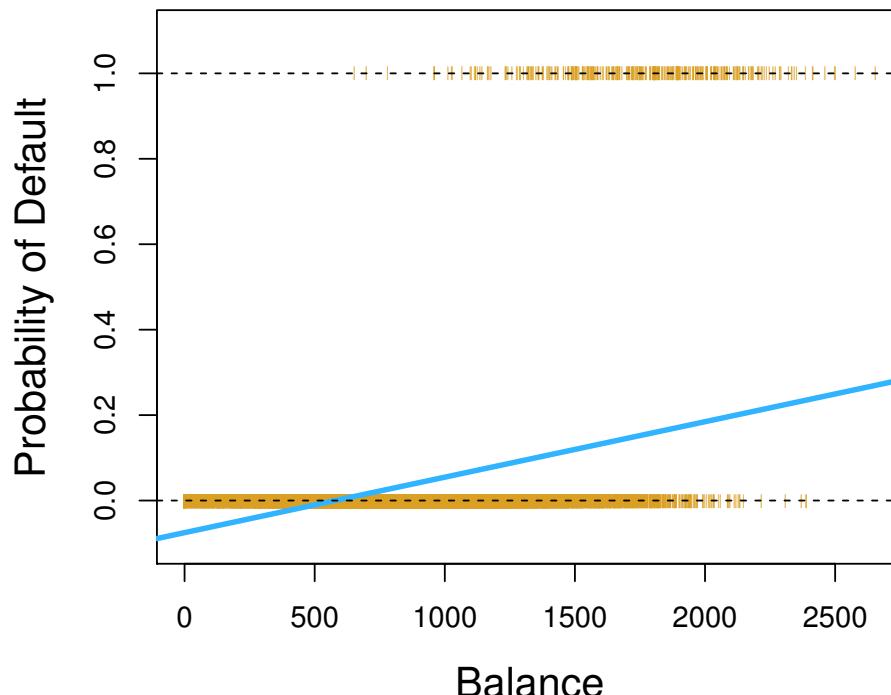
- Para una respuesta binaria con una codificación 0/1 como la anterior, la regresión por mínimos cuadrados no es del todo irrazonable:
 - El $X\hat{\beta}$ obtenido usando la regresión lineal es de hecho una estimación de $Pr(\text{ sobredosis de droga} | X)$ en este caso.
 - Sin embargo, si utilizamos regresión lineal, algunas de nuestras estimaciones podrían estar fuera del intervalo $[0, 1]$, lo que hace que sea difícil interpretarlas como probabilidades.

Observaciones (4)

- Es preferible utilizar un método de clasificación que sea realmente adecuado para valores de respuesta cualitativos.

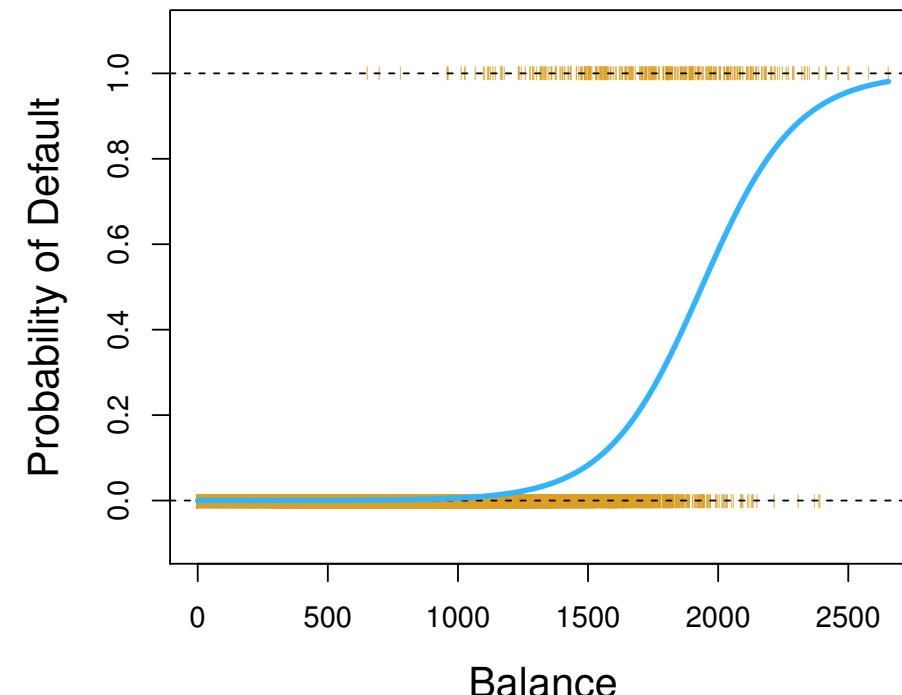
Clasificación Incumplimiento de Pago

Probabilidad estimada de incumplimiento



Regresión Lineal

¡Probabilidades negativas!



Regresión Logística

Probabilidades entre 0 y 1

Las **marcas** indican los valores 0/1 codificados para *Falla de Pago* (No o Sí)

Regresión logística

- Incumplimiento de pago → respuesta es Sí o No
- En lugar de modelar esta respuesta Y directamente, la regresión logística modela la probabilidad de que Y pertenezca a una categoría particular

$$Pr(\text{incumplimiento} = \text{Sí} \mid \text{saldo}) \in [0,1]$$

- Para cualquier valor dado de *saldo*, se puede hacer una predicción del *incumplimiento*
 - *incumplimiento = Sí* si $p(\text{saldo}) > 0.5$
 - *incumplimiento = Sí* si $p(\text{saldo}) > 0.1$

4.3.1

El Modelo Logístico

- ¿Cómo deberíamos modelar la relación entre $p(X) = \Pr(Y = 1 | X)$ y X ?
 - código genérico 0/1 para la respuesta
 - Regresión lineal

$$p(x) = \beta_0 + \beta_1 X$$

- Siempre que se ajusta una línea recta a una respuesta binaria codificada como 0 o 1, en principio siempre podemos predecir $p(X) < 0$ para algunos valores de X y $p(X) > 1$ para otros

Función Logística

- Para evitar este problema, debemos modelar $p(X)$ usando una función que proporcione resultados entre 0 y 1 para todos los valores de X
- Muchas funciones cumplen con esta descripción
- Regresión logística usa la *función logística*

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

- Para *ajustar el modelo*, se usa un método llamado *máxima verosimilitud* (*maximum likelihood*)

Cuotas o Momios (Odds)

$$0 \leq \frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X} < \infty$$

- Por ejemplo, en promedio 1 de cada 5 personas con una cuota de $1/4$ incumplirá, ya que $p(X) = 0.2$ implica una cuota de $0.2/(1 - 0.2) = 1/4$
- *log odds* o *logit*

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

- El modelo de *regresión logística* tiene un *logit* que es lineal en X

Observaciones

- Modelo de regresión lineal
 - β_1 da el cambio promedio en Y asociado con un aumento de una unidad en X
- Modelo de regresión logística
 - Aumentar X en una unidad **cambia *log odds* en β_1**
 - De manera equivalente, **multiplica los *odds* por e^{β_1}**

Observaciones (2)

- La relación entre $p(X)$ y X no es una línea recta, por lo tanto β_1 no corresponde al cambio en $p(X)$ asociado con un aumento de una unidad en X
- La cantidad que $p(X)$ cambia, debido a un cambio en una unidad en X , depende del valor actual de X
- Esto puede observar en la figura anterior

Observaciones (3)

- Independientemente del valor de X ,
 - si β_1 es positivo, un aumento de X estará asociado con un aumento de $p(X)$
 - si β_1 es negativo, un aumento de X estará asociado con una disminución de $p(X)$

4.3.2

Estimación de los Coeficientes de Regresión

- Los coeficientes se estiman a partir de los datos
- Método de máxima verosimilitud (*maximum likelihood*)
- Función de verosimilitud (likelihood function)

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

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

- Las estimaciones β_0 y β_1 se eligen de tal forma que maximicen esta función de verosimilitud

Incumplimiento de Pago

$$p(\text{incumplimiento} = \text{Sí} | \text{saldo}) = \frac{e^{\beta_0 + \beta_1 \text{saldo}}}{1 + e^{\beta_0 + \beta_1 \text{saldo}}}$$

	Coefficient	Std. error	z-statistic	p-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

TABLE 4.1. For the **Default** data, estimated coefficients of the logistic regression model that predicts the probability of **default** using **balance**. A one-unit increase in **balance** is associated with an increase in the log odds of **default** by 0.0055 units.

- Podemos medir la precisión de las estimaciones de los coeficientes calculando sus errores estándar
- La estadística z desempeña el mismo papel que la estadística t en la regresión lineal

Hipótesis nula $H_0 : \beta_1 = 0$

- $z = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$
 - Un valor grande (absoluto) de z indica evidencia en contra de la hipótesis nula
- $p(X) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$
- $p(\text{incumplimiento} = \text{Sí} | \text{saldo}) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$
- La probabilidad de incumplimiento no depende del saldo

4.3.3

Haciendo Predicciones

-

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}}$$

saldo	$p(\text{incumplimiento} = \text{Sí} \text{saldo})$	%
1,000	0.00576	< 1%
2,000	0.586	58%

Predictores Cualitativos

- Se pueden utilizar predictores cualitativos con el modelo de regresión logística utilizando el enfoque de variables *dummy* (ficticias)

$$x_i = \begin{cases} 1, & \text{estudiante = Sí} \\ 0, & \text{estudiante = No} \end{cases}$$

	Coefficient	Std. error	z-statistic	p-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student [Yes]	0.4049	0.1150	3.52	0.0004

TABLE 4.2. For the **Default** data, estimated coefficients of the logistic regression model that predicts the probability of **default** using student status. Student status is encoded as a dummy variable, with a value of 1 for a student and a value of 0 for a non-student, and represented by the variable **student [Yes]** in the table.

Predictores Cualitativos (2)

$$p(\text{incumplimiento} = \text{Sí} | \text{estudiante} = \text{Sí}) = \frac{e^{\beta_0 + \beta_1 \times 1}}{1 + e^{\beta_0 + \beta_1 \times 1}}$$

$$p(\text{incumplimiento} = \text{Sí} | \text{estudiante} = \text{No}) = \frac{e^{\beta_0 + \beta_1 \times 0}}{1 + e^{\beta_0 + \beta_1 \times 0}}$$

4.3.4

Regresión Logística Múltiple

- Predecir una **respuesta binaria** utilizando **múltiples predictores**

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$X = (X_1, \dots, X_p)$$

- Método máxima verosimilitud (maximum likelihood) para estimar $\beta_0, \beta_1, \dots, \beta_p$

Incumplimiento de Pago

$$p(\text{incumplimiento} = \text{Sí} | (\text{saldo}, \text{ingreso}, \text{estudiante})) = \frac{e^{\beta_0 + \beta_1 \text{saldo} + \beta_2 \text{ingreso} + \beta_3 \text{estudiante}}}{1 + e^{\beta_0 + \beta_1 \text{saldo} + \beta_2 \text{ingreso} + \beta_3 \text{estudiante}}}$$

	Coefficient	Std. error	z-statistic	p-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student [Yes]	−0.6468	0.2362	−2.74	0.0062

TABLE 4.3. For the `Default` data, estimated coefficients of the logistic regression model that predicts the probability of `default` using `balance`, `income`, and student status. Student status is encoded as a dummy variable `student [Yes]`, with a value of 1 for a student and a value of 0 for a non-student. In fitting this model, `income` was measured in thousands of dollars.

Observaciones

- Los *valores-p* asociados con el *saldo* y la variable ficticia para la condición de *estudiante* son muy pequeños
 - cada una de estas variables está asociada con la probabilidad de incumplimiento
- El *coeficiente* de la variable ficticia es *negativo*
 - los *estudiantes* tienen menos probabilidades de incumplir que los no estudiantes
- En cambio, el *coeficiente* de la variable ficticia es *positivo* cuando *estudiante* es el único predictor (tabla anterior)

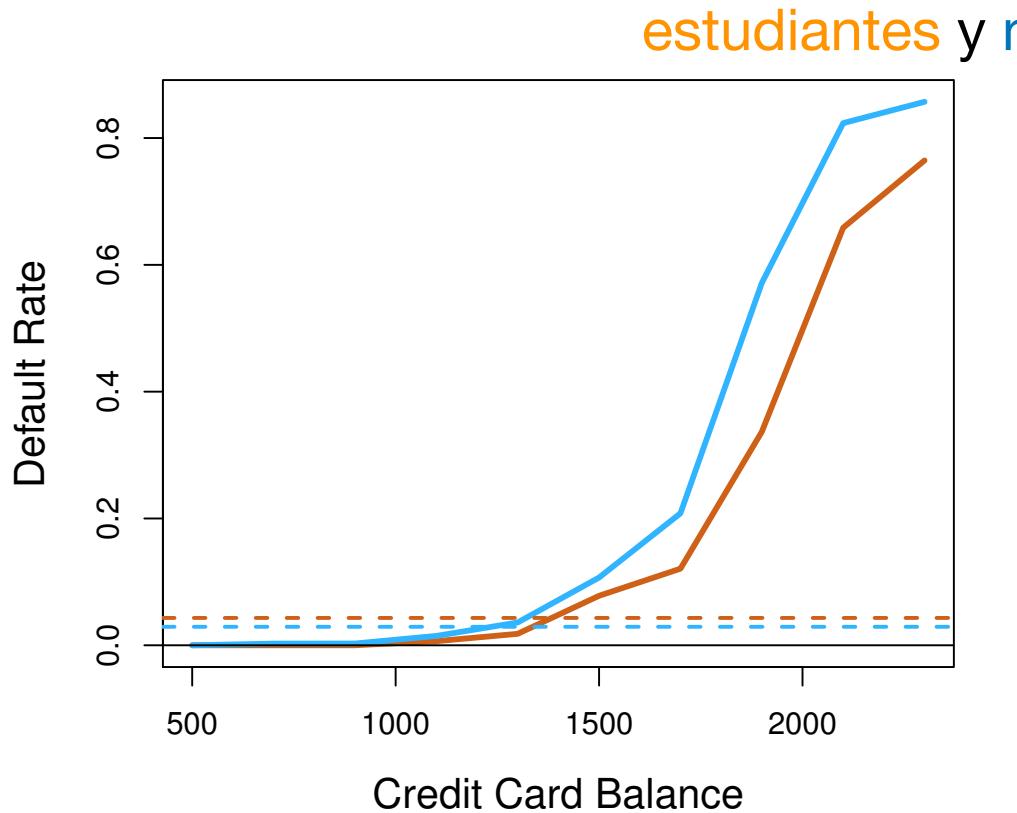
Condición de Estudiante

	Coefficient	Std. error	z-statistic	p-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student [Yes]	0.4049	0.1150	3.52	0.0004

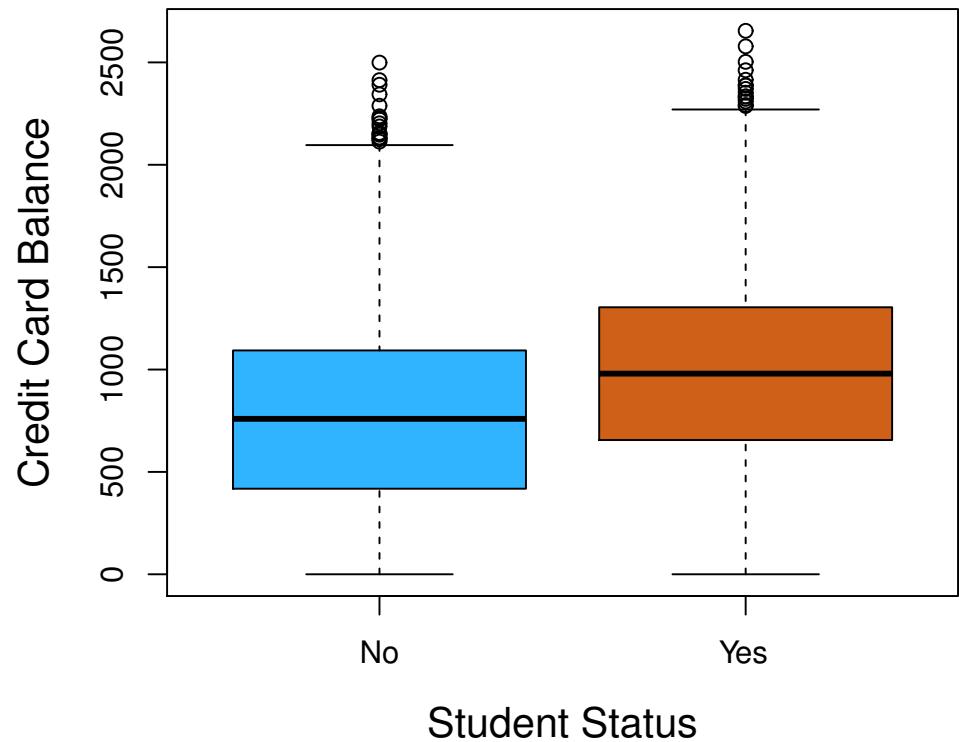
	Coefficient	Std. error	z-statistic	p-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student [Yes]	−0.6468	0.2362	−2.74	0.0062

¿Cómo es posible que la condición de estudiante esté asociada con un aumento en la probabilidad de incumplimiento en un caso y una disminución de dicha probabilidad en el otro?

Condición de Estudiante (2)



Tasa de incumplimiento y saldo



Saldo y estatus de estudiante

- Un estudiante representa más riesgo que un no es estudiante si no hay información disponible sobre el saldo de la tarjeta de crédito.
- Sin embargo, ¡ese estudiante representa menos riesgo que un no estudiante con el mismo saldo de tarjeta de crédito!

4.3.5

Regresión Logística Multinomial

- Clasificar una variable de *respuesta* que tiene **más de dos clases**.
- Por ejemplo, tres categorías de afecciones médicas en la sala de emergencias:
 - accidente cerebrovascular
 - sobredosis de drogas
 - ataque epiléptico
- Es posible extender la regresión logística de dos clases al escenario de $K > 2$ clases para la variable de respuesta
 - Regresión Logística Multinomial

Concepto RLM

- Seleccionar una clase multinomial para que sirva como referencia. Sin pérdida de generalidad, seleccionamos la clase K éSIMA clase como referencia:
- $k = 1, \dots, K - 1$

$$Pr(Y = k | X = x) = \frac{e^{\beta_{k0} + \beta_{k1}x_1 + \dots + \beta_{kp}x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0} + \beta_{l1}x_1 + \dots + \beta_{lp}x_p}}$$

- K

$$Pr(Y = K | X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0} + \beta_{l1}x_1 + \dots + \beta_{lp}x_p}}$$

Concepto RLM

- *log odds*

$$\log\left(\frac{Pr(Y = k | X = x)}{Pr(Y = K | X = x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p$$

- Los *log odds* entre cualquier par de clases son lineales en los predictores

Concepto RLM

- La decisión de cual de las clases se toma como referencia es irrelevante
- Visitas a la sala de emergencias Y
 - accidentes cerebrovasculares, sobredosis de drogas y ataques epilépticos
- Ajustamos dos modelos de regresión logística multinomial
 - *accidente cerebrovascular* → clase de referencia
 - *sobredosis de medicamentos* → clase de referencia
 - Las estimaciones de los coeficientes diferirán entre los dos modelos ajustados debido a la diferente elección de la clase de referencia, pero los valores ajustados (predicciones), los *log odds* entre cualquier par de clases y los demás resultados clave del modelo seguirán siendo los mismos

Interpretación

- La interpretación de los coeficientes en un modelo de regresión logística multinomial está ligada a la elección de la clase de referencia
- Por ejemplo, si el *ataque epiléptico* es la clase de referencia
- → entonces podemos interpretar $\beta_{\text{accidente cerebrovascular} \ 0}$ como el *log odd* de *accidente cerebrovascular* versus *ataque epiléptico*, dado que $x_1 = \dots = x_p = 0$

Interpretación (2)

- Además, un aumento de una unidad en X_j se asocia con un aumento de $\beta_{\text{accidente cerebrovascular } j}$ en log odds de *accidente cerebrovascular* sobre *ataque epiléptico*.
- Dicho de otra manera, si X_j aumenta en una unidad, entonces

$$\log \left(\frac{Pr(Y = \text{accidente cerebrovascular} | X = x)}{Pr(Y = \text{ataque epiléptico} | X = x)} \right)$$

- aumenta en $\beta_{\text{accidente cerebrovascular } j}$

Codificación Softmax

- Codificación alternativa para la Regresión Logística Multinomial
- Equivalente a la codificación que se acaba de describir
 - Los valores ajustados
 - Log odds entre cualquier par de clases
 - Otros resultados clave del modelo

→ seguirán siendo los mismos
- Se usa ampliamente en algunas áreas, por lo que vale la pena tenerla en cuenta

Codificación Softmax (2)

- En la codificación softmax, en lugar de seleccionar una clase de referencia, tratamos todas las K clases simétricamente y asumimos que para $k = 1, \dots, K$,

$$Pr(Y = k | X = x) = \frac{e^{\beta_{k0} + \beta_{k1}x_1 + \dots + \beta_{kp}x_p}}{1 + \sum_{l=1}^K e^{\beta_{l0} + \beta_{l1}x_1 + \dots + \beta_{lp}x_p}}$$

- Por lo tanto, en lugar de estimar coeficientes para $K - 1$ clases, en realidad **estimamos coeficientes para todas las K clases**
- El *log odds* entre las clases k y k' es

$$\log \left(\frac{Pr(Y = k | X = x)}{Pr(Y = k' | X = x)} \right) = (\beta_{k0} - \beta_{k'0}) + (\beta_{k1} - \beta_{k'1})x_1 + \dots + (\beta_{kp} - \beta_{k'p})x_p$$

Modelos Generativos de Clasificación

- La **regresión logística** implica modelar directamente $Pr(Y = k | X = x)$, la **distribución condicional de la respuesta Y dados los predictores X** , usando la función logística.
- Un **enfoque alternativo** y menos directo para estimar estas probabilidades
 - Modelamos la **distribución de los predictores X** por separado en cada una de las clases de respuesta; es decir, **para cada valor de Y** .
 - Usamos el *teorema de Bayes* para convertirlos en estimaciones de $Pr(Y = k | X = x)$. Si la distribución de X dentro de cada clase es normal → el modelo es muy similar en forma a la regresión logística

¿Por qué necesitamos otro método cuando tenemos Regresión Logística?

- Cuando existe una **separación sustancial entre las dos clases**, las estimaciones de los parámetros para el modelo de regresión logística son sorprendentemente inestables. Los métodos que consideramos en esta sección no sufren este problema.
- Si la **distribución de los predictores X es** aproximadamente **normal** en cada una de las clases y el tamaño de la muestra es pequeño, entonces los enfoques de esta sección pueden ser más precisos que la regresión logística.
- Los métodos de esta sección **se pueden extender** al caso de **más de dos clases de respuesta**. (O podemos utilizar la regresión logística multinomial)

Concepto MGC

- Clasificar una observación en una de K clases, $K \geq 2$
- Sea π_k la **probabilidad previa** de que una observación elegida al azar provenga de la k -ésima clase
- Sea $f_k(X) \equiv \Pr(X | Y = k)$ la función de densidad de X para una observación que proviene de la k -ésima clase
 - $f_k(x)$ es relativamente grande si existe una alta probabilidad de que una observación de la k -ésima clase tenga $X \approx x$
 - $f_k(x)$ es pequeña si es muy improbable que una observación de la k -ésima clase tenga $X \approx x$

Concepto MGC (2)

- El *teorema de Bayes* establece que

$$Pr(Y = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_k(l)}$$

- $p_k(x) = Pr(Y = k | X = x)$
 - *probabilidad posterior* de que una observación posterior $X = x$ pertenezca a la k -ésima clase
- En lugar de calcular directamente $p_k(x)$, podemos simplemente sustituir las estimaciones de π_k y $f_k(x)$ en la ecuación anterior

Concepto MGC (3)

- Estimar π_k es fácil si tenemos una muestra aleatoria de la población
 - calculamos la fracción de las observaciones de entrenamiento que pertenecen a la k -ésima clase
- Sin embargo, estimar la función de densidad $f_k(x)$ es más desafiante. Requiere suposiciones simplicadoras
 - si podemos encontrar una manera de estimar $f_k(x)$, entonces podemos sustituirla en la ecuación para aproximar el *clasificador de Bayes*

Aproximando el Clasificador de Bayes

$$Pr(Y = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_k(l)}$$

- Tres clasificadores que utilizan diferentes estimaciones de $f_k(x)$ para aproximar el clasificador de Bayes
 - Análisis discriminante lineal
 - Análisis discriminante cuadrático
 - Bayes ingenuo

Análisis Discriminante Lineal

- Suponemos que $f_k(x)$ es normal o gaussiana

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right),$$

$$p_k = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_k(l)}$$

- Por ahora, supongamos además que $\sigma_1^2 = \dots = \sigma_K^2$
 - Hay un término de varianza σ_2 compartido entre todas las K clases

Clasificador de Bayes

- El *clasificador de Bayes* implica asignar una observación $X = x$ a la clase para la cual $p_k(x)$ es mayor

- $$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

- En la práctica, para aplicar el *clasificador de Bayes* todavía tenemos que **estimar los parámetros**

- $\mu_1, \dots, \mu_K \quad \pi_1, \dots, \pi_K \quad \sigma^2$

Estimaciones de LDA

- El método de *análisis discriminante lineal (LDA)* se aproxima al *clasificador de Bayes* usando las estimaciones

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

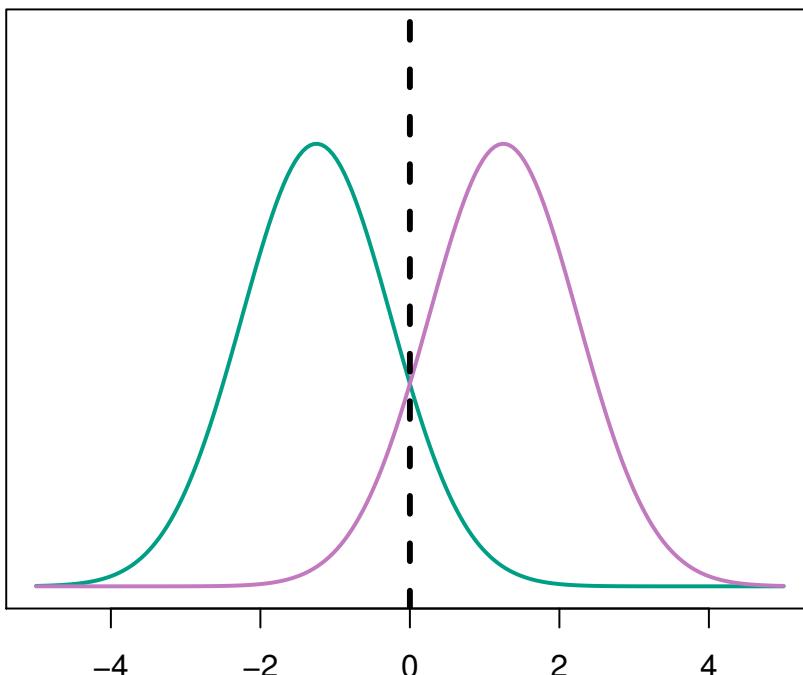
$$\hat{\sigma}^2 = \frac{1}{n - K} \sum_{k=1}^K \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

$$\hat{\pi}_k = n_k/n.$$

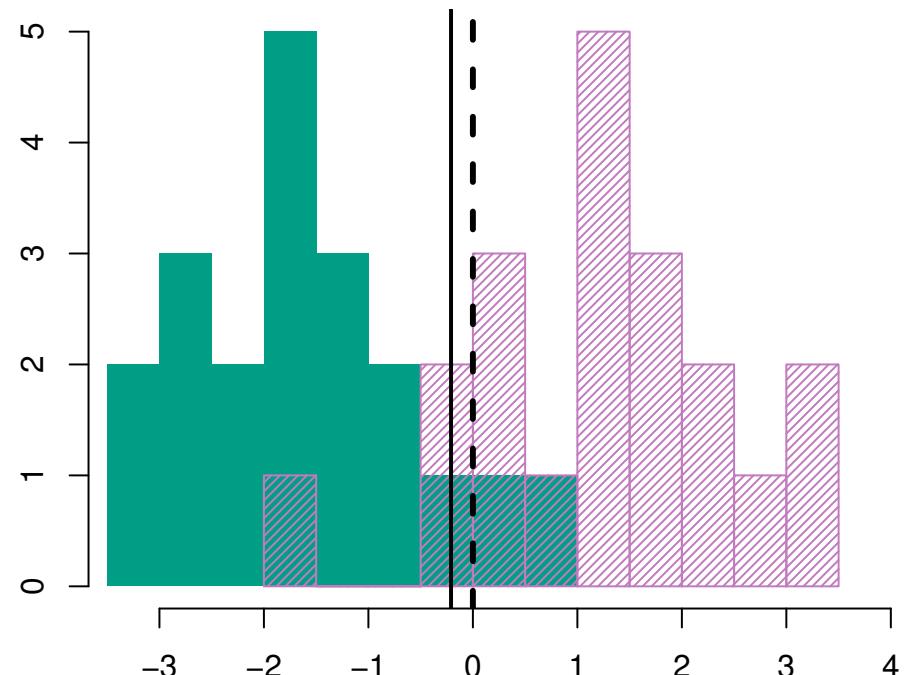
- número de observaciones de entrenamiento
 - n : total, n_k : en la k -ésima clase

Ejemplo LDA

Tasa de error de Bayes 10,6%



Tasa de error de prueba LDA 11,1%

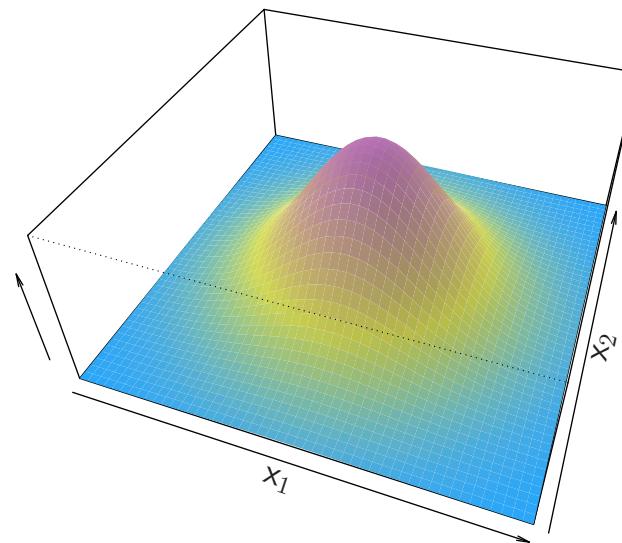


- Dos funciones de densidad normal unidimensionales.
- La línea vertical discontinua : límite de decisión de Bayes.

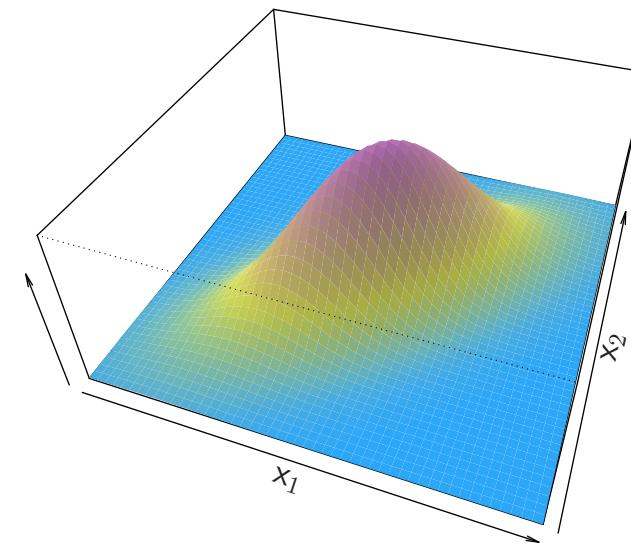
- 20 observaciones de cada una de las dos clases y se muestran como histogramas.
- La línea vertical continua: límite de decisión LDA estimado a partir de los datos de entrenamiento.

Análisis Discriminante Lineal

- Suponemos que $X = (X_1, X_2, \dots, X_p)$ se extrae de una distribución gaussiana multivariada, con un vector medio específico de clase y una matriz de covarianza común.



Los dos predictores no están correlacionados



Las dos predictores tienen una correlación de 0.7

Dos funciones de densidad gaussianas multivariadas, con $p = 2$

$p > 1$

Análisis Discriminante Lineal (2)

- Distribución Gausiana Multivariada

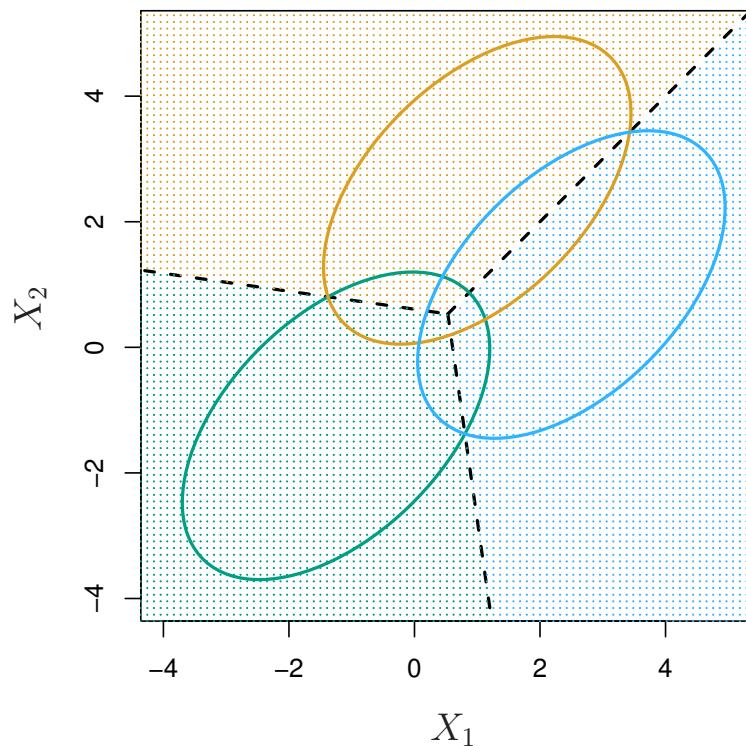
$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right).$$

$$p_k = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_k(l)}$$

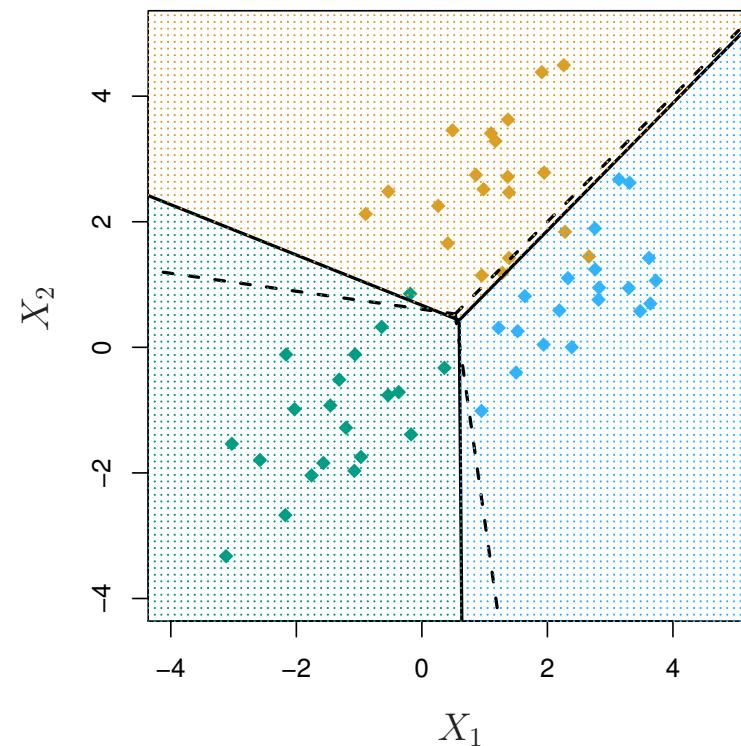
- Clasificador de Bayes

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

Un Ejemplo con Tres Clases



Elipses contienen el 95 % de la probabilidad para cada una de las tres clases. Líneas discontinuas: límites de decisión de Bayes.



20 observaciones de cada clase. Líneas continuas : Los límites de decisión LDA

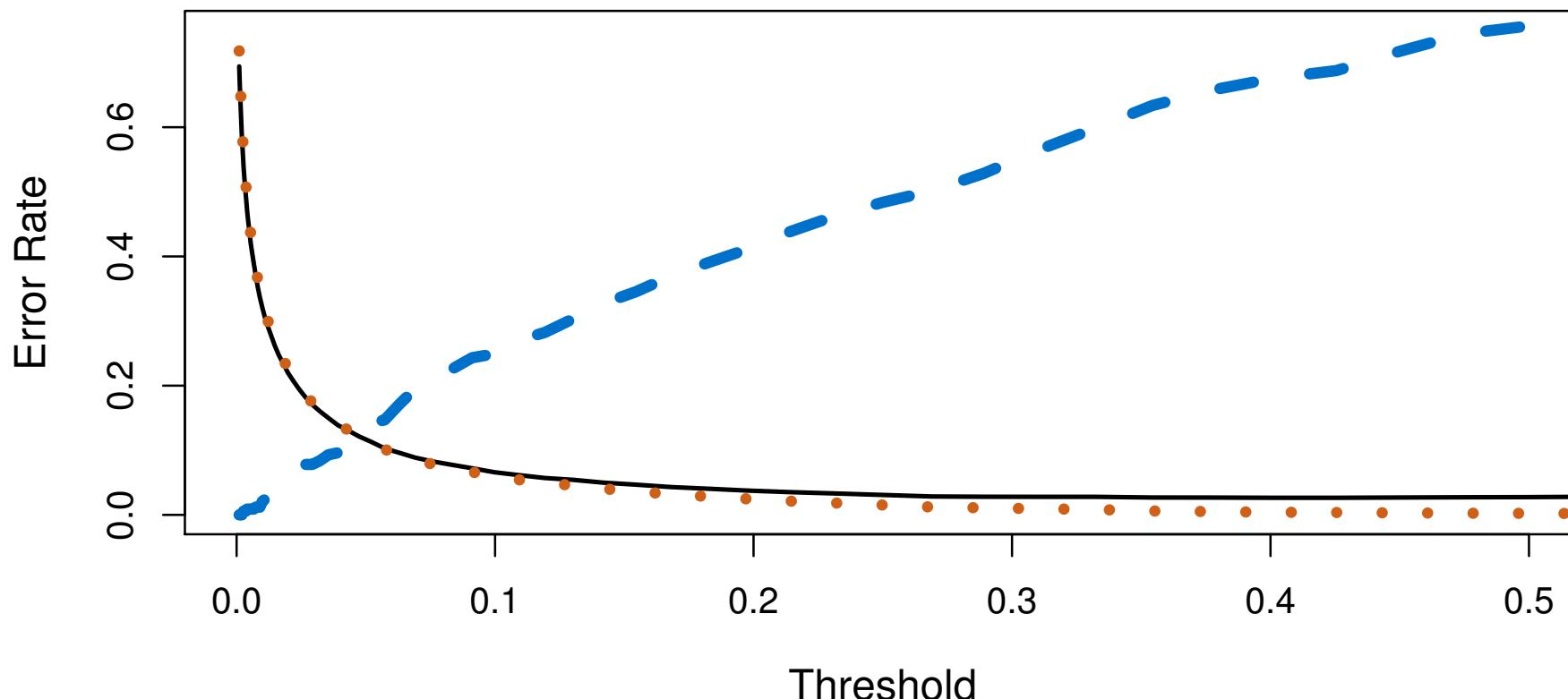
Observaciones de cada clase se extraen de una distribución gaussiana multivariada con $p = 2$, con un vector medio específico de la clase y una matriz de covarianza común

Incumplimiento de Pago

		<i>True default status</i>		
		No	Yes	Total
<i>Predicted default status</i>	No	9644	252	9896
	Yes	23	81	104
Total		9667	333	10000

TABLE 4.4. A confusion matrix compares the LDA predictions to the true default statuses for the 10,000 training observations in the **Default** data set. Elements on the diagonal of the matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were misclassified. LDA made incorrect predictions for 23 individuals who did not default and for 252 individuals who did default.

Incumplimiento de Pago

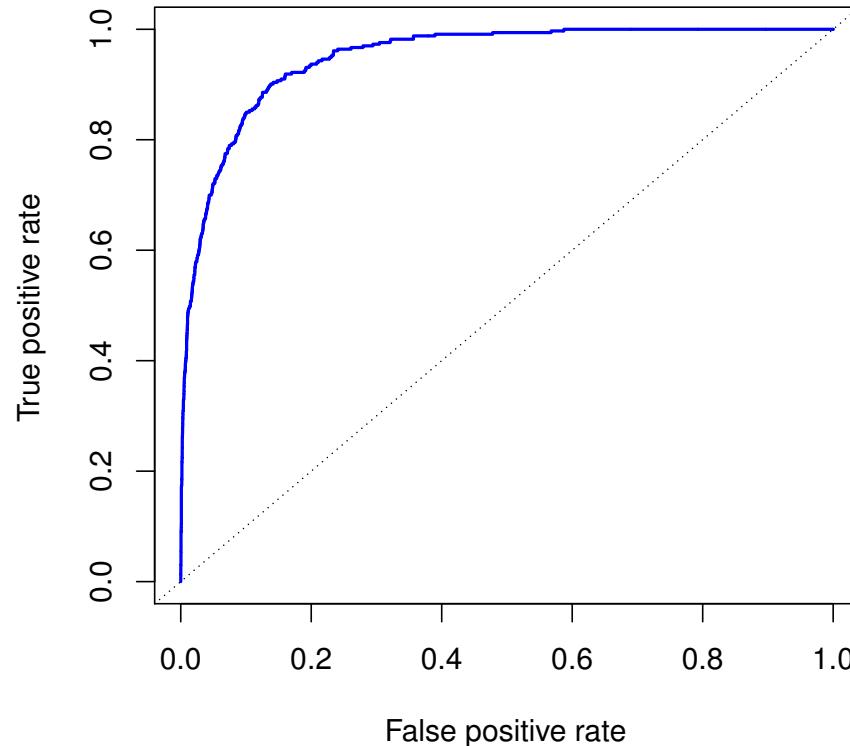


Tasas de error como una función del valor umbral para la probabilidad posterior que se utiliza para realizar la asignación de clase.

- **Tasa de error general.**
- **Fracción de clientes morosos que están clasificados incorrectamente.**
- **Fracción de errores entre los clientes no morosos.**

ROC

La *tasa de verdaderos positivos* (true positives) es la *sensibilidad*: la fracción de morosos que se identifican correctamente



La *tasa de falsos positivos* (false positives) es de *especificidad-1*: la fracción de no morosos que clasificamos incorrectamente como morosos

- La curva ROC traza dos tipos de error a medida que variamos el valor umbral para la probabilidad posterior de incumplimiento. Umbrales no se puestran.
- La curva ROC ideal se aproxima a la esquina superior izquierda, lo que indica una *alta tasa de verdaderos positivos* y una *baja tasa de falsos positivos*.
- La línea de puntos representa el clasificador “sin información”; esto es lo que esperaríamos si el estatus de estudiante y el saldo de la tarjeta de crédito no estuvieran asociados con la probabilidad de incumplimiento.

Análisis Discriminante Cuadrático (QDA)

- Observaciones dentro de cada clase se extraen de una distribución gaussiana multivariada con un vector medio específico de la clase y
- *Cada clase tiene su matriz de covarianza.* Se supone que una observación de la k -ésima clase es de la forma $X \sim N(\mu_k, \Sigma_k)$, donde Σ_k es una matriz de covarianza para la k -ésima clase
- Introducir estimaciones de los parámetros en el teorema de Bayes para realizar la predicción

Diferencia entre QDA y LDA

Característica	LDA	QDA
Observaciones dentro de cada clase se extraen de una distribución gaussiana multivariada con un vector medio específico de la clase	✓	✓
Matriz de covarianza común a todas las K clases	✓	
Cada clase tiene su matriz de covarianza		✓
Introducir estimaciones de los parámetros en el teorema de Bayes para realizar la predicción	✓	✓

QDA

$$p_k = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_k(l)}$$

- Clasificador de Bayes

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) - \frac{1}{2} \log |\Sigma_k| + \log \pi_k \\ &= -\frac{1}{2}x^T \Sigma_k^{-1}x + x^T \Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T \Sigma_k^{-1}\mu_k - \frac{1}{2} \log |\Sigma_k| + \log \pi_k\end{aligned}$$

- $\delta_k(x)$ es una función cuadrática
- QDA provee estimaciones para Σ_k , μ_k y π_k . Asigna una observación $X = x$ a la clase para la cual δ_k es mayor

¿QDA o LDA?

- ¿Por qué importa si suponemos o no que las clases K comparten una matriz de covarianza común?
- En otras palabras,
 - ¿por qué preferiríamos LDA a QDA, o viceversa?
- La respuesta está en el equilibrio entre sesgo y varianza

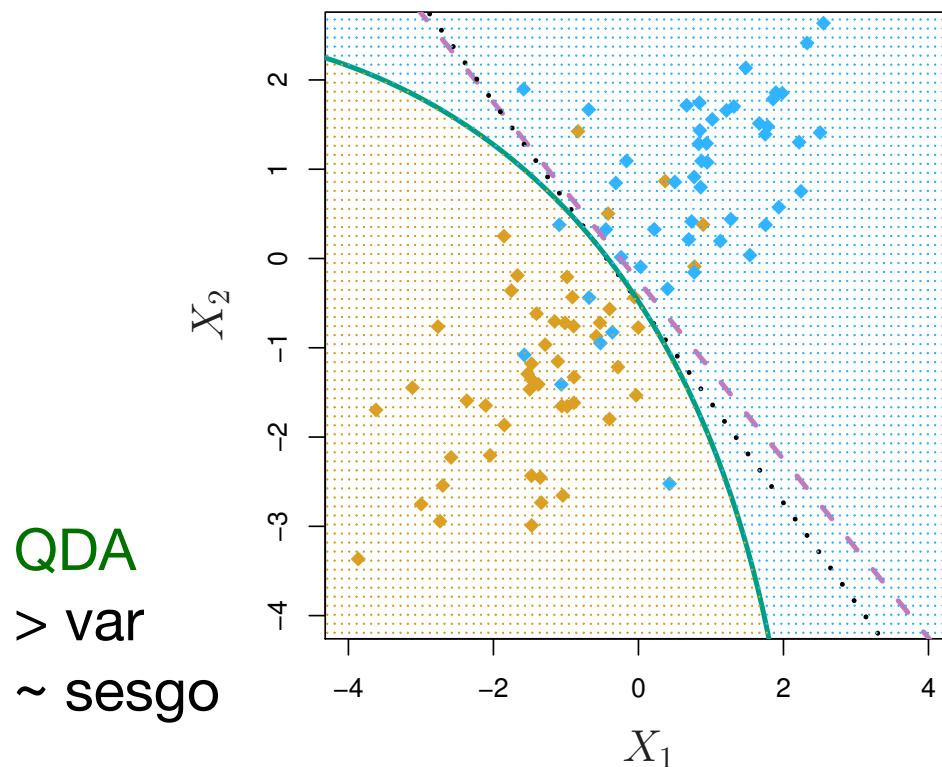
Número de Parámetros

- Cuando hay p predictores, estimar una matriz de covarianza requiere estimar $p(p + 1)/2$ parámetros (LDA)
- QDA estima una matriz de covarianza separada para cada clase, para un total de parámetros $Kp(p + 1)/2$
- LDA es un clasificador mucho **menos flexible** que QDA
 - LDA tiene una **varianza menor**. Potencialmente, mejor predicción.
 - Pero, si el supuesto de LDA de que las clases K comparten una matriz de covarianza común está mal → LDA puede sufrir un **alto sesgo**

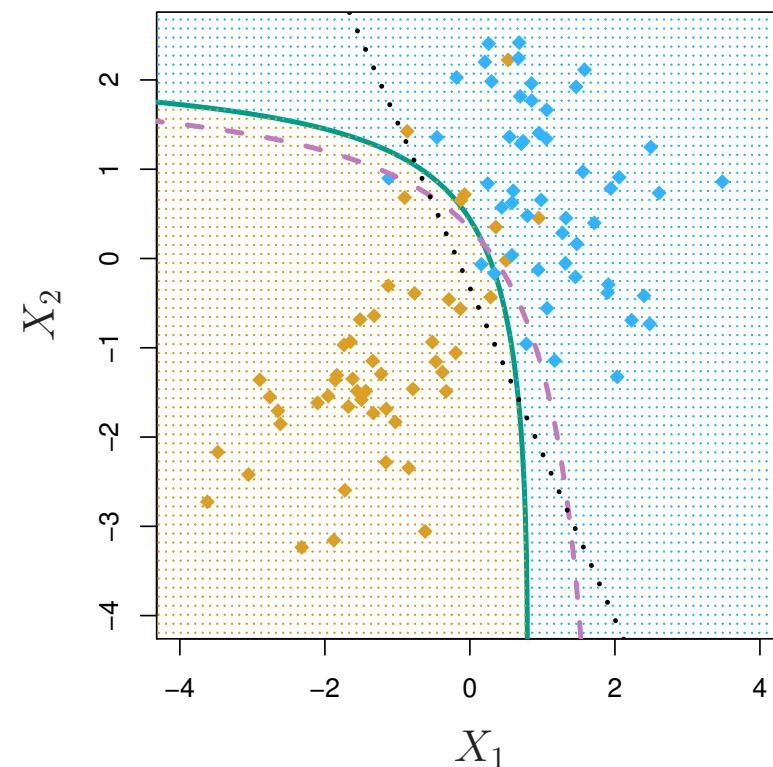
Cuando usar LDA or QDA

- Se recomienda LDA si hay relativamente pocas observaciones de entrenamiento
 - por lo que reducir la varianza es crucial
- Se recomienda QDA si
 - el conjunto de entrenamiento es muy grande, de modo que la varianza del clasificador no es una preocupación importante
 - o si la suposición de una matriz de covarianza común para las clases K es claramente insostenible.

LDA and QDA



$\Sigma_1 = \Sigma_2 = 0.7$. El **límite de decisión de Bayes** es **lineal**
→ **LDA** lo aproxima con mayor precisión que **QDA**.



$\Sigma_1 = 0.7$, $\Sigma_2 = -0.7$. El **límite de decisión de Bayes** es **no lineal**
→ **QDA** lo aproxima con mayor precisión que **LDA**.

Bayes Ingenuo

- Teorema de Bayes

$$p(x) = \Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_k(l)}$$

- Desarrollar LDA, QDA
 - Estimaciones de probabilidades previas π_1, \dots, π_K , y funciones de densidades de probabilidad $f_1(x), \dots, f_K(x)$
- Motivar el popular Clasificador de Bayes Ingenuo

Suposición Principal

- El ingenuo clasificador de Bayes adopta un rumbo diferente para estimar $f_1(x), \dots, f_K(x)$.
- En lugar de suponer que $f_1(x), \dots, f_K(x)$ pertenecen a una familia particular de distribuciones (por ejemplo, normal multivariada), hacemos una única suposición:
 - *Dentro de la k -ésima clase, los p predictores son independientes.*

Suposición Principal (2)

- Expresado matemáticamente, este supuesto significa que para $k = 1, \dots, K$,

$$f_k(x) = f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)$$

- donde f_{jk} es la función de densidad del j -ésimo predictor entre observaciones de la k -ésima clase.

¿Por qué esta suposición es tan poderosa?

- Estimar una función de densidad p -dimensional es un desafío
 - considerar la *distribución marginal* de cada predictor: la distribución de cada predictor por sí solo
 - la *distribución conjunta* de los predictores: la asociación entre los diferentes predictores
- Distribución normal multivariada: la asociación entre los diferentes predictores se resume en los elementos fuera de la diagonal de la matriz de covarianza
 - Esta asociación puede ser muy difícil de caracterizar y extremadamente difícil de estimar
- Al suponer que las p covariables son independientes dentro de cada clase, eliminamos completamente la necesidad de preocuparnos por la asociación entre los p predictores, ¡porque simplemente hemos asumido que no hay asociación entre los predictores!

¿Las p covariables son realmente independientes dentro de cada clase?

- **No!**, en la mayoría de casos
 - Este supuesto de modelado se hace por conveniencia
- **Sin embargo**, estimar una distribución conjunta (LDA, QDA) requiere una cantidad grande de datos → **el ingenuo Bayes es una buena opción** en una amplia gama de entornos
 - Si n no es lo **suficientemente** grande en relación con p para estimar la distribución conjunta de los predictores dentro de cada clase
- **El ingenuo supuesto de Bayes** introduce cierto sesgo, pero reduce la varianza → un clasificador que funciona bien en la práctica como resultado del equilibrio entre sesgo y varianza

Teorema de Bayes y Bayes Ingenuo

$$Pr(Y = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_k(l)}$$

$$Pr(Y = k | X = x) = \frac{\pi_k f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)}{\sum_{l=1}^K \pi_l f_{l1}(x_1) \times f_{l2}(x_2) \times \cdots \times f_{lp}(x_p)}$$

Estimar f_{kj}

- Para estimar la función de densidad unidimensional f_{kj} utilizando datos de entrenamiento x_{1j}, \dots, x_{nj} , tenemos algunas opciones
 - X_j es cuantitativo
 - X_j es cualitativo

X_j es Cuantitativo

- Podemos asumir que $X_j | Y = k \sim N(\mu_{jk}, \sigma_{jk}^2)$.
- Dentro de cada clase, el j -ésimo predictor se extrae de una distribución normal (univariada).
- Parecido a QDA, pero hay una diferencia clave:
 - aquí asumimos que los predictores son independientes
 - equivale a QDA con el supuesto adicional de que la matriz de covarianza específica de la clase es diagonal

X_j es Cuantitativo: Opción 2

- Utilizar una *estimación no paramétrica* para f_{kj}
 - *Histograma* para las observaciones del j -ésimo predictor dentro de cada clase
 - ▶ estimar $f_{kj}(x_j)$ como la fracción de las observaciones de entrenamiento en la k -ésima clase que pertenecen al mismo contenedor de histograma que x_j
 - Alternativamente, podemos usar un *estimador de densidad kernel* (una versión suavizada de un histograma)

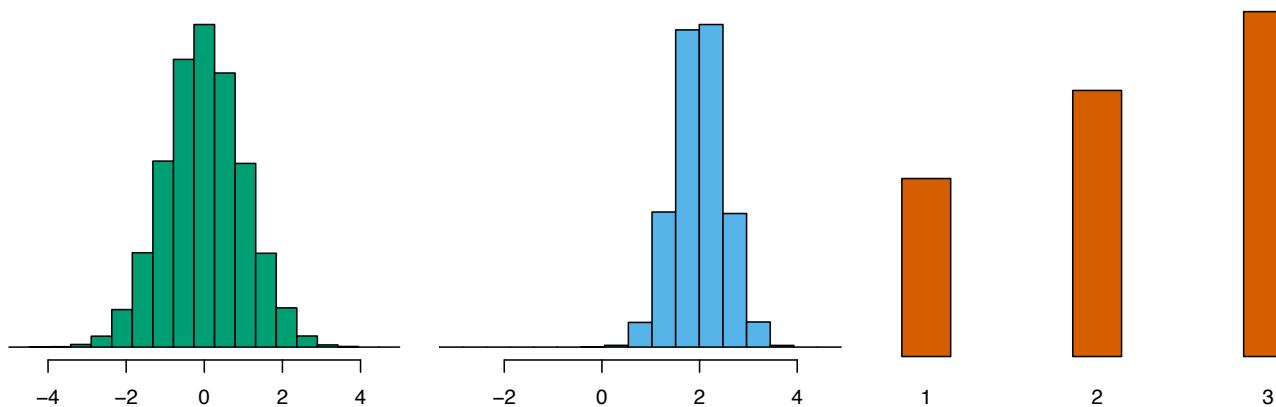
X_j es Cualitativo

- Contar la proporción de observaciones de entrenamiento para el j -ésimo predictor correspondiente a cada clase
- Ej.: $X_j \in \{1,2,3\}$, 100 observaciones en la k -ésima clase
 - Si X_j toma valores de 1, 2 y 3 en 32, 55 y 13 de esas observaciones, entonces podemos estimar f_{kj} como

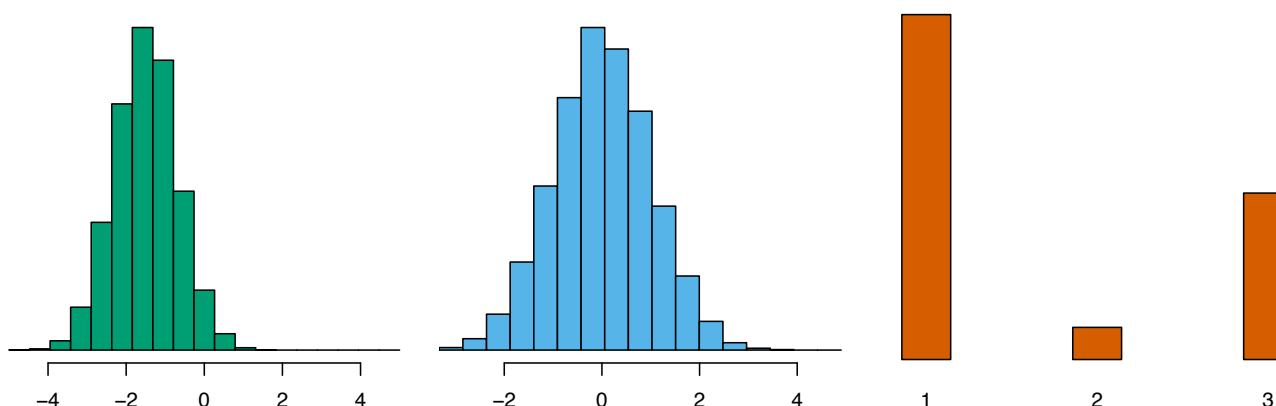
$$\hat{f}_{kj}(x_j) = \begin{cases} 0.32 & \text{if } x_j = 1 \\ 0.55 & \text{if } x_j = 2 \\ 0.13 & \text{if } x_j = 3 \end{cases}$$

Ejemplo Simple

Estimaciones de densidad para la clase $k = 1$



Estimaciones de densidad para la clase $k = 2$



$K = 2$
 $p = 3$
- 2 cuantitativos
- 1 cualitativo
 $\pi_1 = \pi_2$,
 $x = (0.4, 1.5, 1)$

$$\begin{cases} \hat{f}_{11}(0.4) = 0.368 \\ \hat{f}_{12}(1.5) = 0.484 \\ \hat{f}_{13}(1) = 0.226 \end{cases}$$

$$\begin{cases} \hat{f}_{21}(0.4) = 0.030 \\ \hat{f}_{22}(1.5) = 0.0130 \\ \hat{f}_{23}(1) = 0.616 \end{cases}$$

$\rightarrow p(x) = 94.4\%$
 $k = 1$

Incumplimiento de Pago

		<i>True default status</i>		Total
		No	Yes	
<i>Predicted default status</i>	No	9621	244	9865
	Yes	46	89	135
	Total	9667	333	10000

TABLE 4.8. Comparison of the naive Bayes predictions to the true default status for the 10,000 training observations in the **Default** data set, when we predict default for any observation for which $P(Y = \text{default}|X = x) > 0.5$.

		<i>True default status</i>		Total
		No	Yes	
<i>Predicted default status</i>	No	9339	130	9469
	Yes	328	203	531
	Total	9667	333	10000

TABLE 4.9. Comparison of the naive Bayes predictions to the true default status for the 10,000 training observations in the **Default** data set, when we predict default for any observation for which $P(Y = \text{default}|X = x) > 0.2$.

Bayes Ingenuo, falla?

- En este ejemplo, no sorprende que el *Bayes Ingenuo* no supere de manera convincente a LDA:
 - este conjunto de datos tiene $n = 10,000$ y $p = 2$, por lo que la reducción de la varianza resultante del *Bayes Ingenuo* no necesariamente vale la pena.
 - Se espera una mayor rentabilidad al usar *Bayes Ingenuo* en relación con LDA o QDA en los casos en que p es mayor o n es menor, por lo que reducir la varianza es muy importante.

Una Comparación de Métodos de Clasificación

- Regresión Logística
- LDA
- QDA
- Bayes Ingenuo
- KNN

Configuración Experimental

- Datos generados para 6 *escenarios* diferentes
 - problema de clasificación binaria (de dos clases).
- Límite de decisión de Bayes
 - *lineal* : 3 escenarios
 - *no lineal* : 3 escenarios
- Para cada escenario,
 - Se generaron 100 conjuntos de datos de entrenamiento aleatorios.
 - Se ajusta cada método en cada uno de estos conjuntos de entrenamiento
 - Se calcula la tasa de error de prueba resultante en un conjunto de pruebas grande

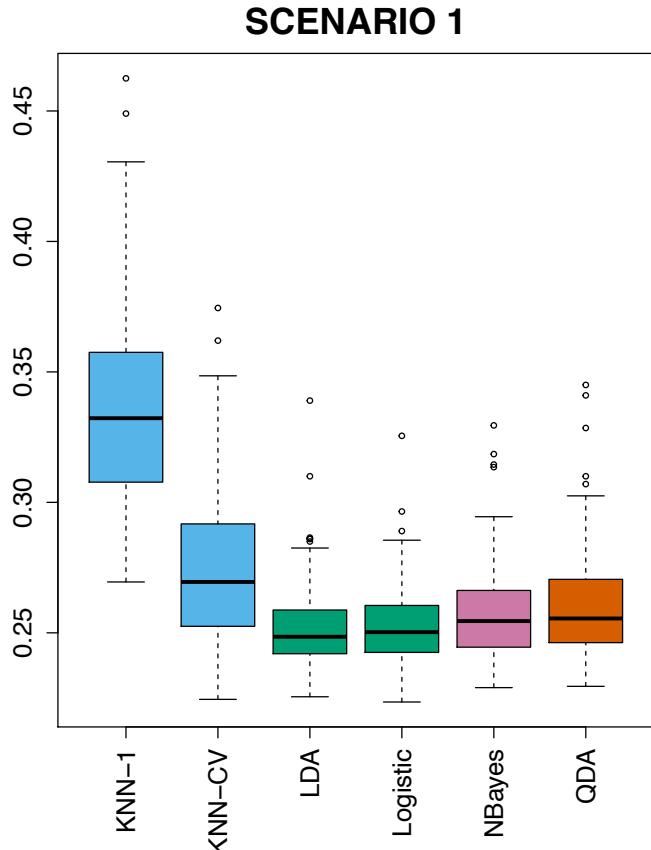
Conf. Exp. (2)

- El método *KNN* requiere la selección de K , el número de vecinos (no debe confundirse con el número de clases).
- Su usa *KNN* con dos valores de K :
 - $K = 1$
 - Un valor de K que se eligió automáticamente mediante un enfoque llamado *validación cruzada*

Conf. Exp. (3)

- Bayes Ingenuo se aplicó asumiendo
 - independencia de los predictores (por supuesto)
 - densidades gausianas uni-variadas para los predictores dentro de cada clase

Escenarios Lineales



Escenario 1

- 20 observaciones de entrenamiento en cada una de las dos clases
- Las observaciones dentro de cada clase fueron variables normales aleatorias no correlacionadas con una media diferente en cada clase

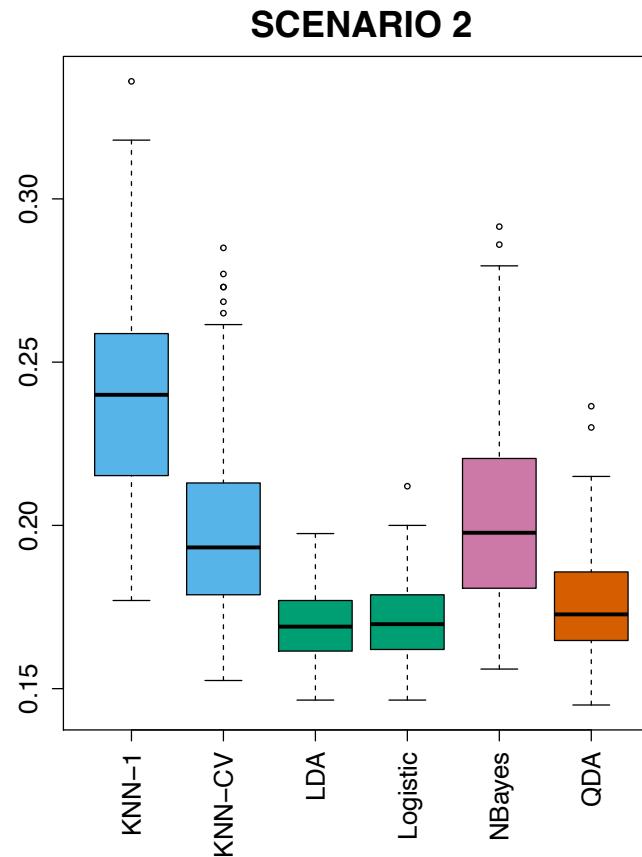
Observaciones

- *LDA* : buen desempeño, como era de esperar ya que este es el modelo asumido por LDA
- *Regresión logística* : funciona bastante bien, ya que supone un límite de decisión lineal
- *KNN* : desempeño deficiente. Pagó un precio en términos de varianza que no fue compensado por una reducción del sesgo
- *QDA* : peor desempeño que LDA, ya que se ajustaba a un clasificador más flexible de lo necesario
- *Bayes Ingenuo* : ligeramente mejor que QDA. La suposición de predictores independientes es correcta

Escenarios Lineales

Escenario 2

- Los detalles son los mismos que en el Escenario 1, excepto que dentro de cada clase, los dos predictores tiene una correlación de $-0,5$



Observaciones

- El desempeño de la mayoría de los métodos es similar al Escenario 1.
- Bayes Ingenuo* es la excepción notable
- Se desempeña muy mal aquí, ya que se viola el supuesto de predictores independientes.

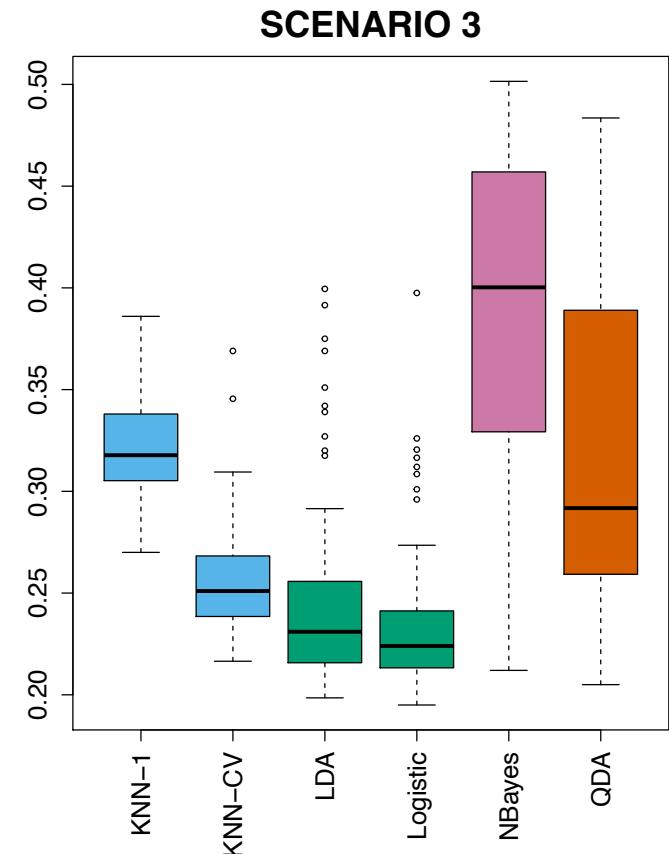
Escenarios Lineales

Escenario 3

- Como en el escenario 2, existe una correlación negativa sustancial entre los predictores dentro de cada clase. Sin embargo, X_1 y X_2 se generan a partir de la *distribución t*, con 50 observaciones por clase.
- La *distribución t* tiene una forma similar a la *distribución normal*, pero tiende a producir más puntos extremos, i.e. más puntos alejados de la media
- El límite de decisión todavía es lineal : encaja en el marco de la regresión logística.
- La configuración viola los supuestos de LDA: las observaciones no se extrajeron de una distribución normal.

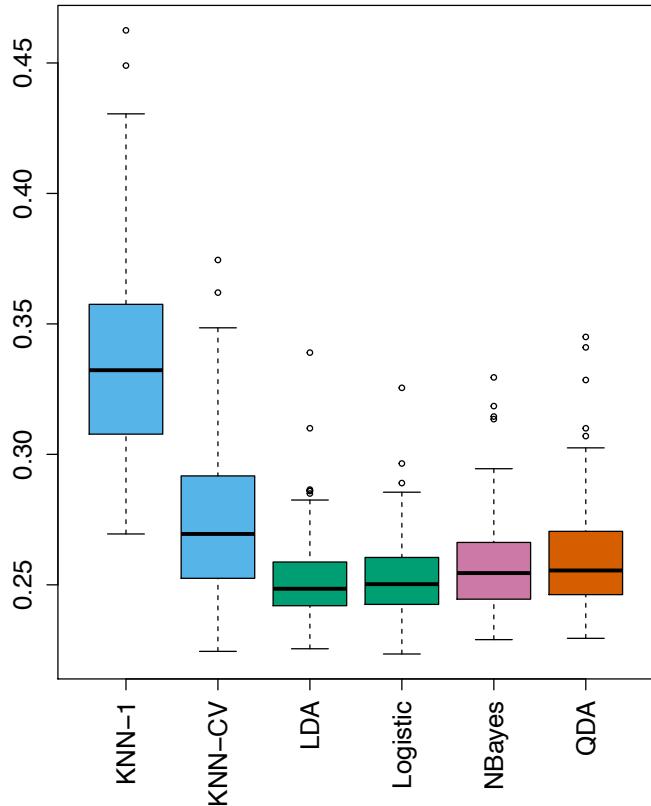
Observaciones

- Regresión logística supera a LDA. Ambos son superiores a los otros enfoques
- QDA deteriora considerablemente como consecuencia de la anormalidad
- Naive Bayes: desempeño muy pobre porque se viola el supuesto de independencia

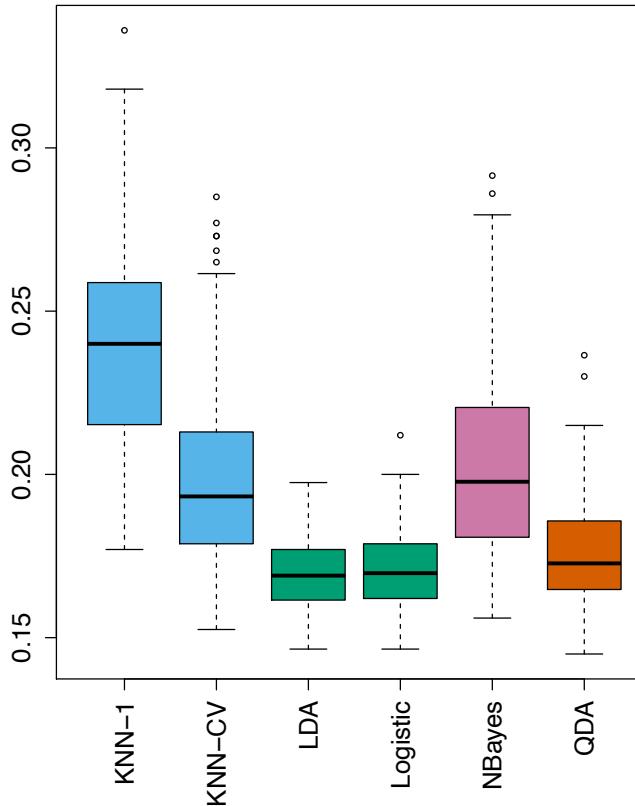


Escenarios Lineales

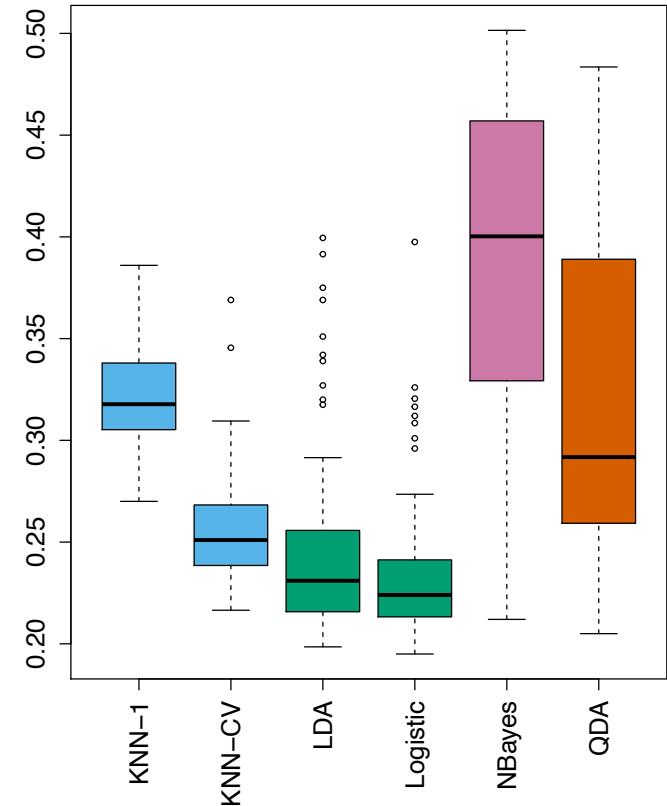
SCENARIO 1



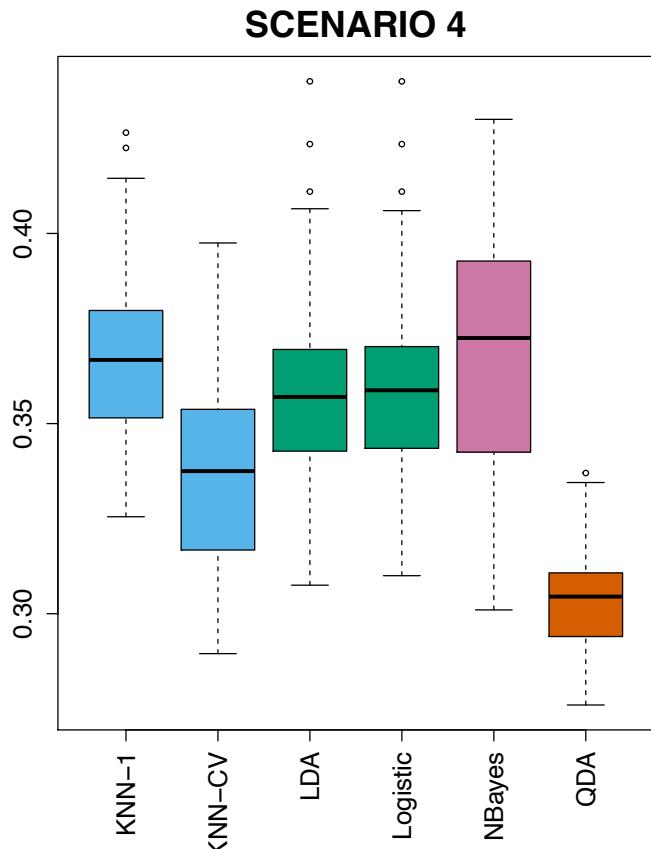
SCENARIO 2



SCENARIO 3



Escenarios No Lineales



Escenario 4

- Los datos se generan a partir de una distribución normal, con una correlación de 0,5 entre los predictores de la primera clase y una correlación de -0,5 entre los predictores de la segunda clase
- Esta configuración corresponde al supuesto QDA
- Resulta en límites de decisión cuadráticos

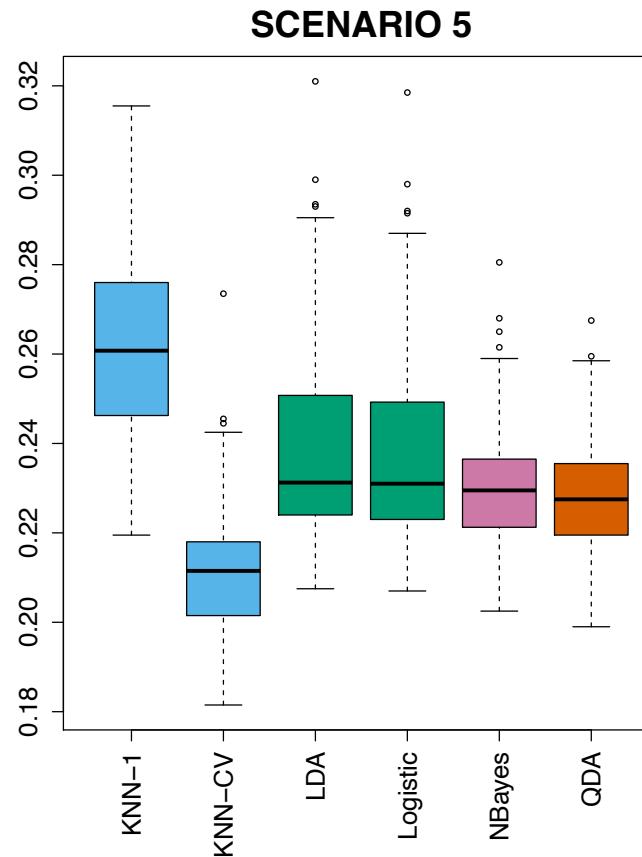
Observaciones

- QDA supera a todos los demás enfoques.
- Bayes Ingenuo tiene un desempeño deficiente : se viola el supuesto de predictores independientes

Escenarios No Lineales

Escenario 5

- Los datos se generan a partir de una distribución normal con predictores no correlacionados
- Luego, las respuestas se muestrearon de la función logística aplicada a una complicada función no lineal de los predictores



Observaciones

- QDA y *Bayes Ingenuo* da resultados ligeramente mejores que los métodos lineales
- KNN-CV, mucho más flexible, da los mejores resultados
- KNN con $K = 1$ da los peores resultados
- Incluso si los datos exhiben una relación no lineal compleja, un método no paramétrico como KNN puede dar malos resultados si el nivel de flexibilidad no se elige correctamente

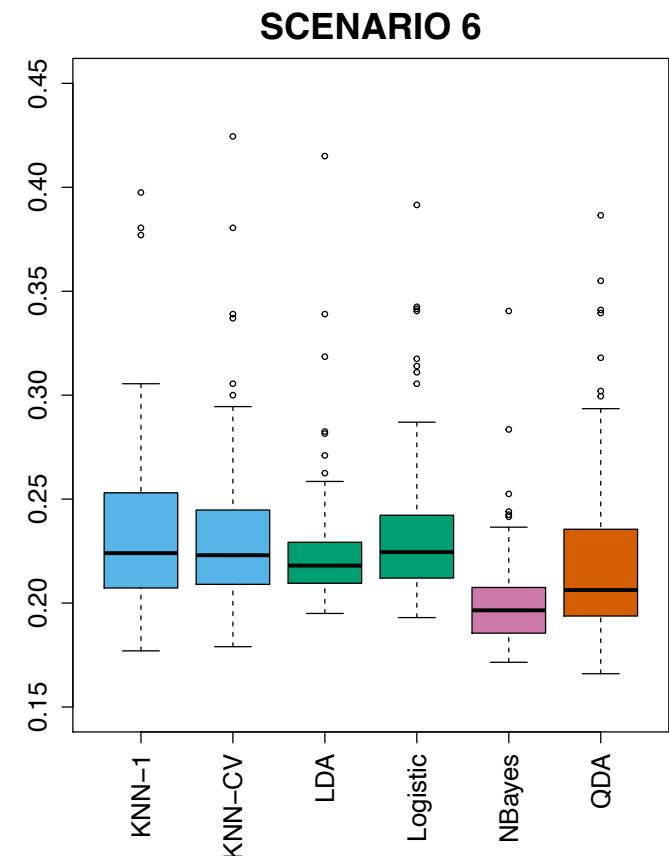
Escenarios No Lineales

Escenario 6

- Las observaciones se generan a partir de una distribución normal con una matriz de covarianza diagonal diferente para cada clase
- Sin embargo, el tamaño de la muestra es muy pequeño: sólo $n = 6$ en cada clase

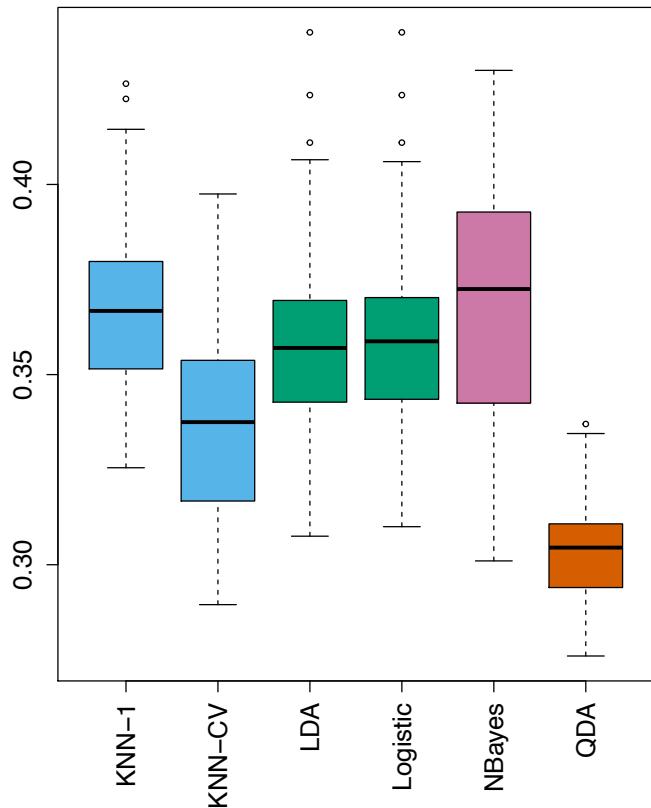
Observaciones

- *Bayes Ingenuo*: funciona muy bien; se cumplen sus supuestos.
- *LDA* y *Regresión Logística* : desempeño deficiente. El límite de decisión no es lineal, debido a las matrices de covarianza desiguales.
- *QDA* : desempeño un poco peor que Bayes Ingenuo. Dado el tamaño de muestra muy pequeño, QDA incurre en demasiada varianza al estimar la correlación entre los predictores dentro de cada clase.
- *KNN* : su desempeño también se ve afectado debido al tamaño muy pequeño de la muestra.

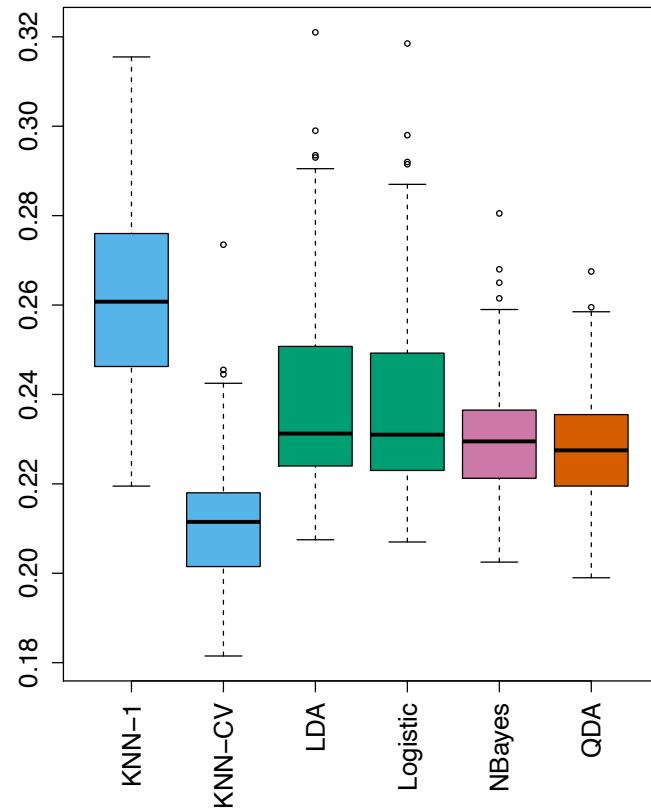


Escenarios No Lineales

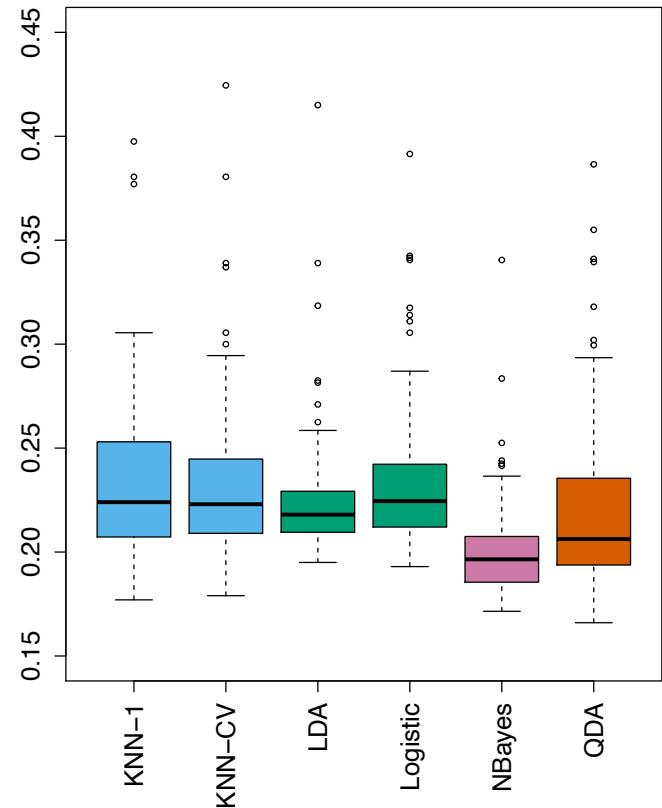
SCENARIO 4



SCENARIO 5



SCENARIO 6



Comparación : Conclusiones

- Estos seis ejemplos ilustran que **ningún método domina a los demás en todas las situaciones**
- Cuando los verdaderos *límites de decisión son lineales*, los enfoques *LDA y de Regresión Logística* tenderán a funcionar bien
- Cuando los *límites son moderadamente no lineales*, *QDA o Bayes Ingenuo* pueden dar mejores resultados
- Para *límites de decisión mucho más complicados*, un *enfoque no paramétrico como KNN* puede ser superior. Pero el nivel de flexibilidad para un enfoque no paramétrico debe elegirse con cuidado.
- En el próximo capítulo examinaremos una serie de enfoques para elegir el nivel correcto de flexibilidad y, en general, para seleccionar el mejor método general.

Relaciones no Lineales entre X y Y

- En Regresión : podemos acomodar una relación no lineal entre los predictores y la respuesta incluyendo transformaciones de los predictores
- Se podría adoptar un enfoque similar en el ámbito de la clasificación.
- Por ejemplo, podríamos crear una versión más flexible de la Regresión Logística incluyendo X^2 , X^3 e incluso X^4 como predictores
 - Esto puede o no mejorar el desempeño de la regresión logística. Depende de si el aumento de la varianza debido a la mayor flexibilidad se compensa con una reducción suficientemente grande del sesgo
- Se puede hacer lo mismo con LDA
 - Si agregamos todos los términos cuadráticos y productos cruzados posibles a LDA, la forma del modelo sería la misma que la del modelo QDA, aunque las estimaciones de los parámetros serían diferentes
 - Esto nos permite movernos entre un modelo LDA y un QDA.

Modelos Lineales Generalizados (GLM)

1. Regresión lineal de los datos de bicicletas compartidas
2. Regresión de Poisson sobre los datos de bicicletas compartidas
3. Modelos lineales generalizados en mayor generalidad

Respuesta: Ni Cualitativa Ni Cuantitativa

- Respuesta *Y es cuantitativa*
 - Exploramos el uso de la *Regresión Lineal* para predecir Y
- Respuesta *Y es cualitativa*
 - Exploramos *Regresión Logística*, *LDM*, *QDM*, *Bayes Ingenuo*, *KNN*, para predecir la clase
- Hay situaciones en las que *Y no es ni cualitativa ni cuantitativa*
 - Ni la regresión lineal ni los métodos de clasificación que hemos visto son aplicables

Bicicleta Compartida (BC)

- La respuesta son los ciclistas *bikers*, el número de usuarios por horas de un programa de bicicletas compartidas en Washington, DC.
- Este valor de respuesta no es ni cualitativo ni cuantitativo: toma valores enteros no negativos o un *contador*.
- Consideraremos predecir los ciclistas *bikers* usando las covariables:
 - *mnth* (mes del año),
 - *hr* (hora del día, de 0 a 23),
 - *workingday* (día laborable, igual a 1 si no es fin de semana ni feriado),
 - *temp* (temperatura normalizada, en grados Celsius), y
 - *weathersit* (variable sobre el tiempo que toma uno de cuatro valores posibles: despejado; brumoso o nublado; lluvia ligera o nieve ligera; o lluvia intensa o nieve intensa).
 - El mes, la hora y el tiempo se tratan como variables cualitativas

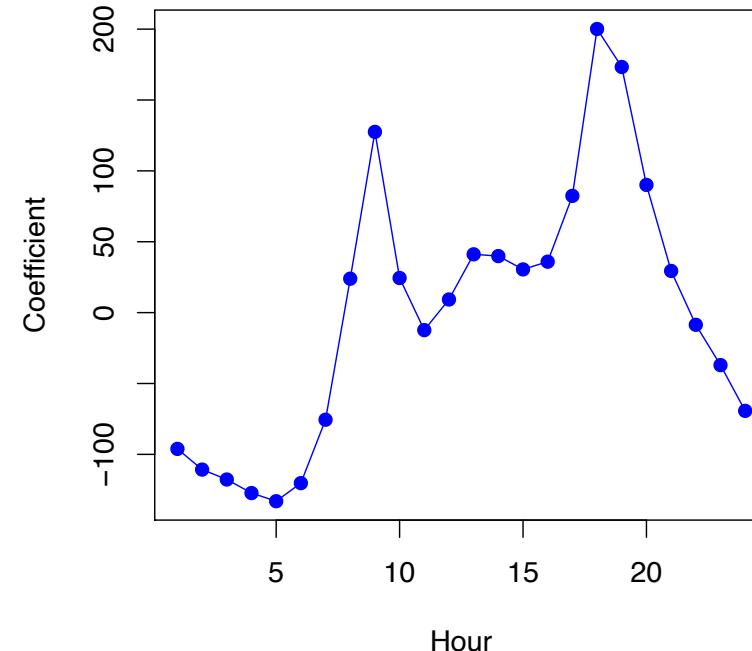
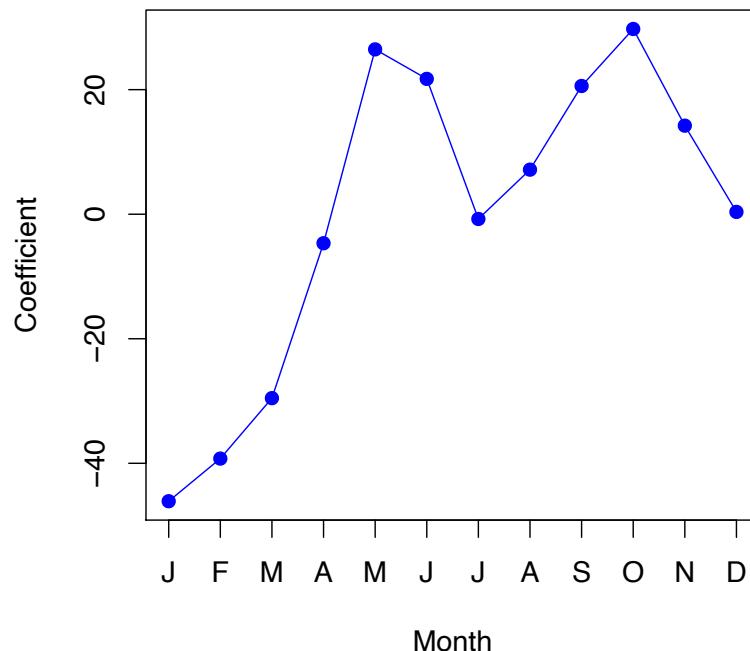
Regresión Lineal de BC

$$\text{bikers} \sim \beta_0 + \beta_1 \text{workingday} + \beta_2 \text{temp} + \dots$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	73.60	5.13	14.34	0.00
workingday	1.27	1.78	0.71	0.48
temp	157.21	10.26	15.32	0.00
weathersit [cloudy/misty]	-12.89	1.96	-6.56	0.00
weathersit [light rain/snow]	-66.49	2.97	-22.43	0.00
weathersit [heavy rain/snow]	-109.75	76.67	-1.43	0.15

Resultados
razonables e
intuitivos?

TABLE 4.10. Results for a least squares linear model fit to predict **bikers** in the Bikeshare data. The predictors **mnth** and **hr** are omitted from this table due to space constraints, and can be seen in Figure 4.13. For the qualitative variable **weathersit**, the baseline level corresponds to clear skies.



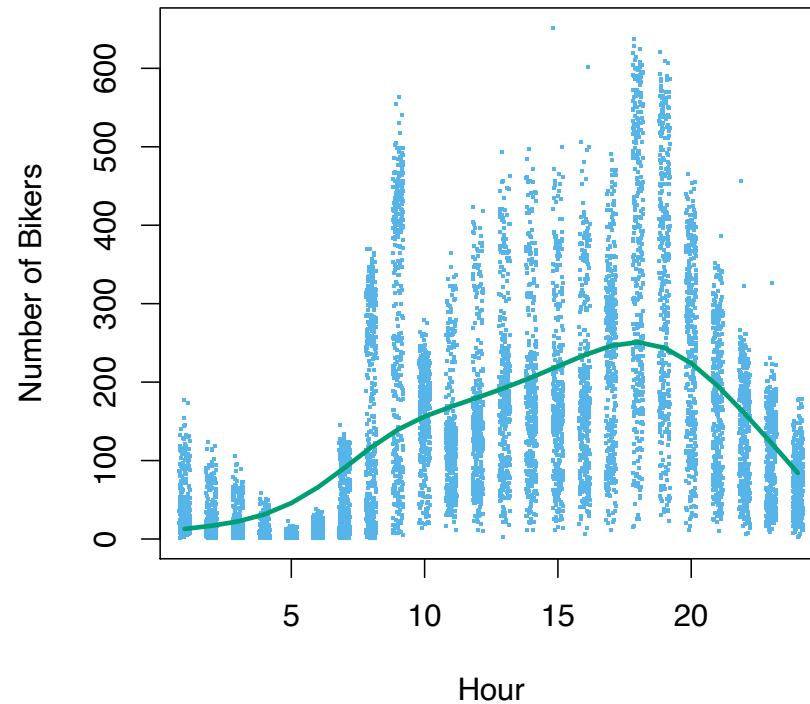
Problemas con LR de BC

- Tras una inspección cuidadosa, algunos problemas se vuelven evidentes
- El modelo de regresión lineal predice un número negativo de usuarios durante el 9,6% de las horas del conjunto de datos
- Pone en duda
 - capacidad para realizar predicciones significativas sobre los datos
 - precisión de las estimaciones de los coeficientes, los intervalos de confianza y otros resultados del modelo de regresión

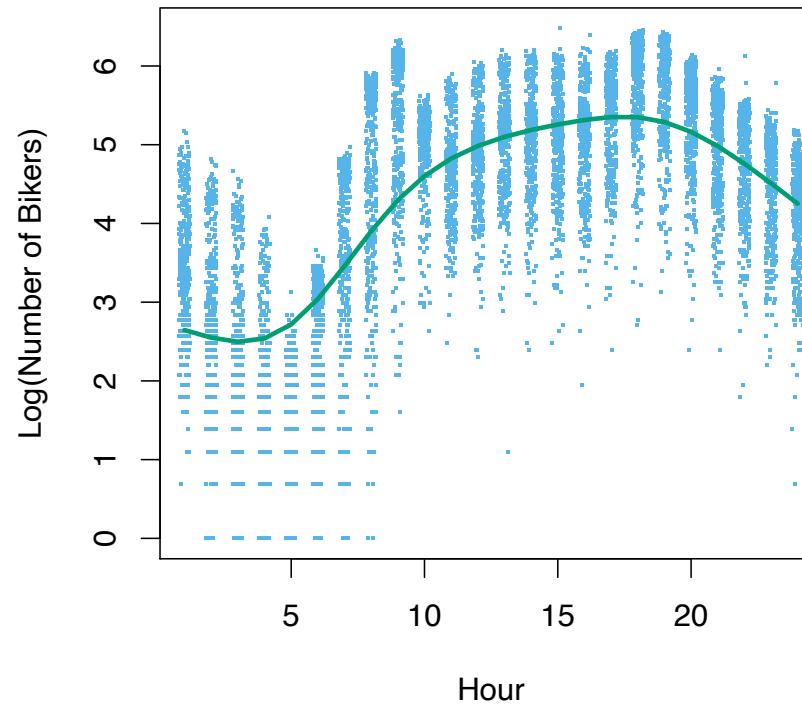
Problemas con LR de BC (2)

$$Y = \sum_{j=1}^p \beta_j X_j + \varepsilon$$

Relación media - varianza



Logaritmo



- **La varianza en el número de ciclistas aumenta con la media**
 - ▶ Viola supuestos importantes de un modelo lineal,
 - ε : media cero con una varianza σ^2 constante y no una función de las covariables
 - ▶ **Heterocedasticidad** de los datos pone en duda la idoneidad del modelo de regresión lineal

Problemas con LR de BC (3)

$$Y = \sum_{j=1}^p \beta_j X_j + \varepsilon$$

- La *respuesta*, *número de ciclistas*, tiene un *valor entero*
 - Pero bajo un modelo donde el término de *error* ε es *continuo*
- La *respuesta* Y es necesariamente *continua (cuantitativa)*
- La naturaleza entera de la respuesta sugiere que un modelo de regresión lineal no es del todo satisfactorio para este conjunto de datos

Transformar la Respuesta

- *Ej.*

$$\log(Y) = \sum_{j=1}^p \beta_j X_j + \varepsilon$$

- Evita predicciones Y negativas
- Supera en parte la *heterocedasticidad* en los datos no transformados
- *Ajustar un modelo lineal a una transformación de la respuesta puede ser un enfoque adecuado para algunos conjuntos de datos con valores de conteo*
- Pero, a menudo no es una solución del todo satisfactoria
 - Predicciones e inferencias en términos del logaritmo de la respuesta
 - “un aumento de una unidad en X_j está asociado con un aumento en la media del \log de Y en una cantidad β_j ”
 - Si $Y = 0$, no se puede aplicar el logaritmo

Regresión de Poisson de BC

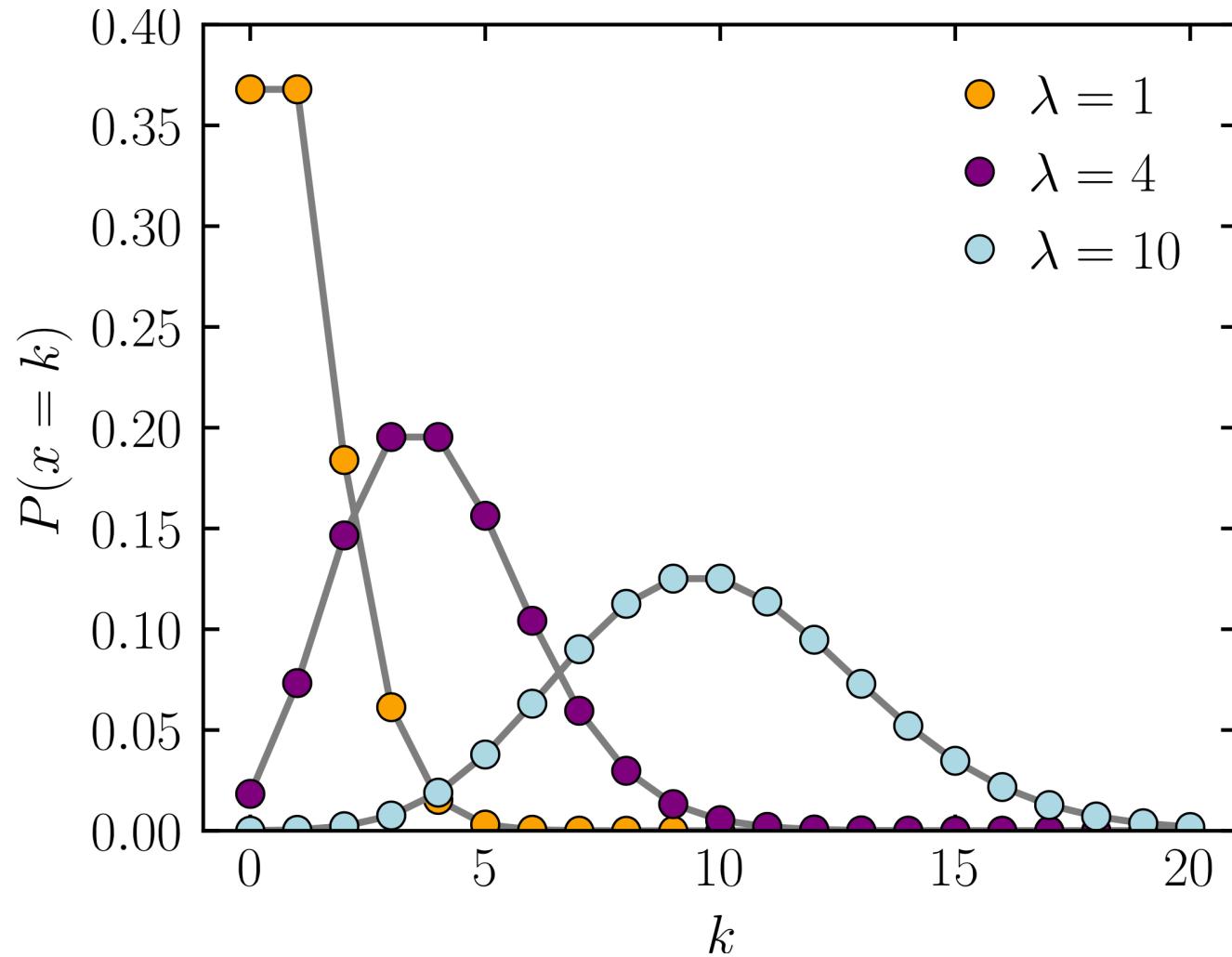
- Suponga que una variable aleatoria toma valores no negativos, $Y \in \{0,1,2,\dots\}$
- Si Y sigue una *distribución de Poisson*

$$Pr(Y = k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k = 0,1,2,\dots$$

$$\lambda = E(Y) = Var(Y)$$

- Cuanto mayor es la media de Y , mayor es su varianza!
- La *distribución de Poisson* se utiliza normalmente para modelar contadores

Distribución de Poisson



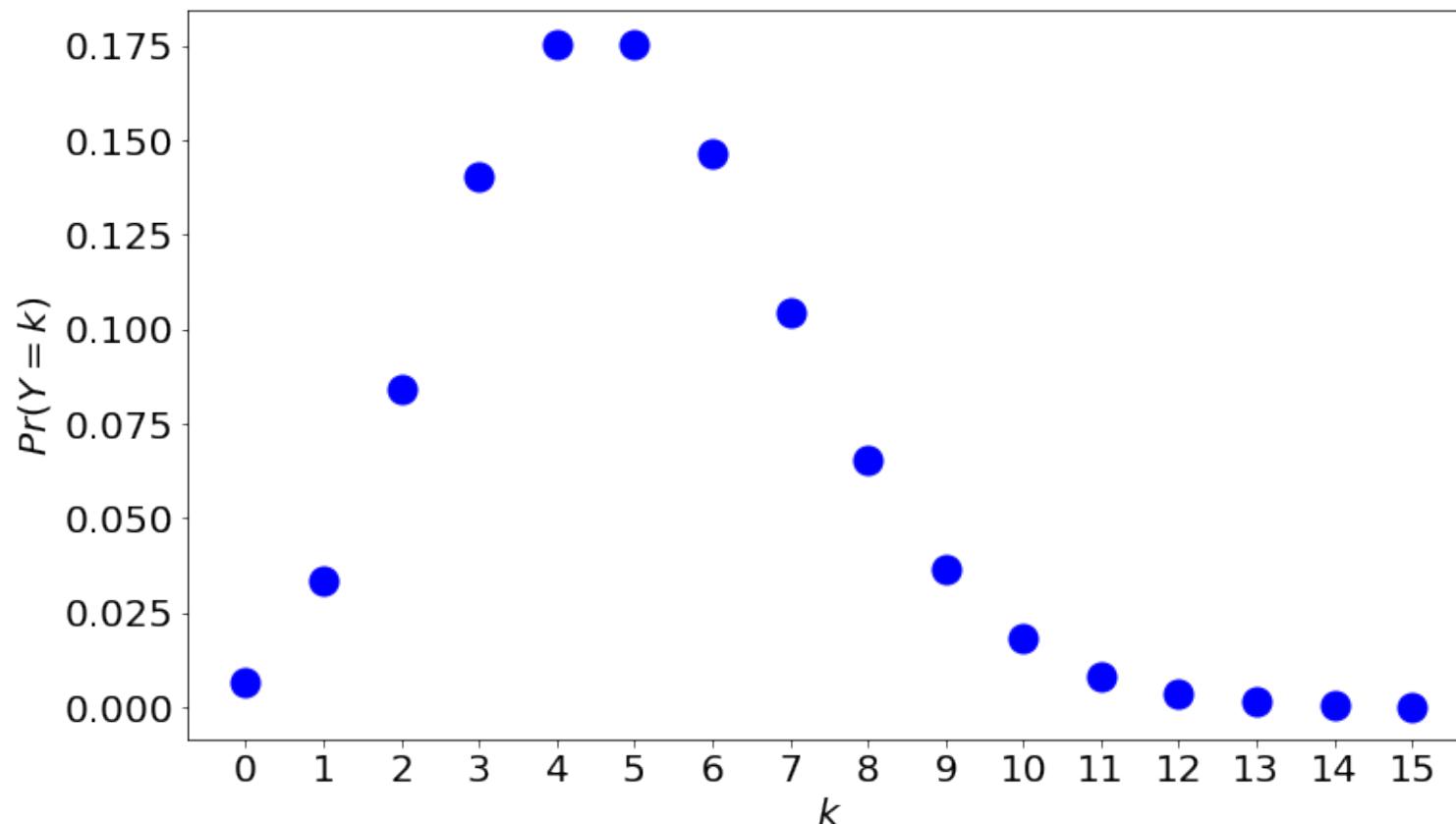
$$Pr(x = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

$$\lambda = E(Y) = Var(Y)$$

Distribución de Poisson en la Práctica

- Sea Y el número de usuarios del programa de bicicletas compartidas
 - durante una hora particular del día,
 - bajo un conjunto particular de condiciones climáticas y
 - durante un mes particular del año.
- Podríamos modelar Y como una distribución de Poisson con media $E(Y) = \lambda = 5$

Poisson $E(Y) = \lambda = 5$



$$Pr(x=k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

Distribución de Poisson en la Práctica (2)

- En realidad esperamos que el número medio de usuarios del programa de bicicletas compartidas, $\lambda = E(Y)$, varíe en función de la hora del día, el mes del año, las condiciones climáticas, etc.

$$\log(\lambda(X_1, \dots, X_p)) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

$$\lambda(X_1, \dots, X_p) = e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}$$

- $\beta_0, \beta_1, \dots, \beta_p$ son parámetros a ser estimados

Regresión de Poisson

$$Pr(x = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

$$\log(\lambda(X_1, \dots, X_p)) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

- Tomamos
 - \log de $\lambda(X_1, \dots, X_p)$ como lineal en X_1, \dots, X_p
 - en lugar de que $\lambda(X_1, \dots, X_p)$ sea lineal en X_1, \dots, X_p
- Asegura que $\lambda(X_1, \dots, X_p)$ tome valores no negativos para todos los valores de las covariables

Máxima Verosimilitud

- Datos $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- Se estima $\beta_0, \beta_1, \dots, \beta_p$ similar a Regresión Logística

$$l(\beta_0, \beta_1, \dots, \beta_p) = \prod_{i=1}^n \frac{e^{-\lambda(x_i)} \lambda(x_i)^{y_i}}{y_i!}$$

$$\lambda(x_i) = e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}$$

- Se estima los coeficientes que maximizan $l(\beta_0, \beta_1, \dots, \beta_p)$

Regresión de Poisson de BC

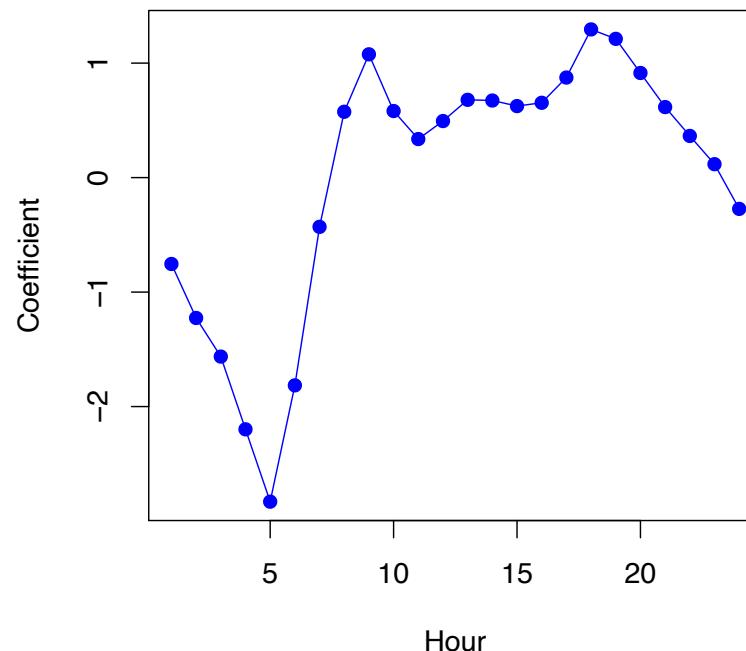
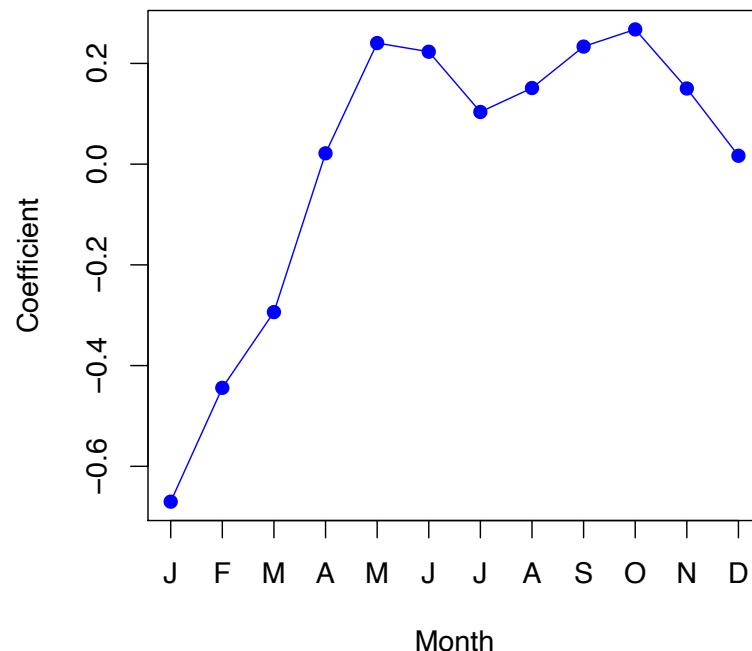
$$\log(E(\text{bikers})) \sim \beta_0 + \beta_1 \text{workingday} + \beta_2 \text{temp} + \dots$$

Cualitativamente:
resultados similares
a regresión lineal

	Coefficient	Std. error	z-statistic	p-value
Intercept	4.12	0.01	683.96	0.00
workingday	0.01	0.00	7.5	0.00
temp	0.79	0.01	68.43	0.00
weathersit [cloudy/misty]	-0.08	0.00	-34.53	0.00
weathersit [light rain/snow]	-0.58	0.00	-141.91	0.00
weathersit [heavy rain/snow]	-0.93	0.17	-5.55	0.00



TABLE 4.11. Results for a Poisson regression model fit to predict **bikers** in the Bikeshare data. The predictors **mnth** and **hr** are omitted from this table due to space constraints, and can be seen in Figure 4.15. For the qualitative variable **weathersit**, the baseline corresponds to clear skies.



Lineal vs Poisson

$$\text{bikers} \sim \beta_0 + \beta_1 \text{workingday} + \beta_2 \text{temp} + \dots$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	73.60	5.13	14.34	0.00
workingday	1.27	1.78	0.71	0.48
temp	157.21	10.26	15.32	0.00
weathersit [cloudy/misty]	-12.89	1.96	-6.56	0.00
weathersit [light rain/snow]	-66.49	2.97	-22.43	0.00
weathersit [heavy rain/snow]	-109.75	76.67	-1.43	0.15



TABLE 4.10. Results for a least squares linear model fit to predict **bikers** in the **Bikeshare** data. The predictors **mnth** and **hr** are omitted from this table due to space constraints, and can be seen in Figure 4.13. For the qualitative variable **weathersit**, the baseline level corresponds to clear skies.

$$\log(E(\text{bikers})) \sim \beta_0 + \beta_1 \text{workingday} + \beta_2 \text{temp} + \dots$$

	Coefficient	Std. error	z-statistic	p-value
Intercept	4.12	0.01	683.96	0.00
workingday	0.01	0.00	7.5	0.00
temp	0.79	0.01	68.43	0.00
weathersit [cloudy/misty]	-0.08	0.00	-34.53	0.00
weathersit [light rain/snow]	-0.58	0.00	-141.91	0.00
weathersit [heavy rain/snow]	-0.93	0.17	-5.55	0.00



TABLE 4.11. Results for a Poisson regression model fit to predict **bikers** in the **Bikeshare** data. The predictors **mnth** and **hr** are omitted from this table due to space constraints, and can be seen in Figure 4.15. For the qualitative variable **weathersit**, the baseline level corresponds to clear skies.

Interpretación

- Un aumento en X_j en una unidad está asociado con un cambio en $E(Y) = \lambda$ por un factor de $\exp(\beta_j)$
 - Un cambio en el tiempo de cielo despejado a nublado se asocia con un cambio en el uso medio de la bicicleta por un factor de
 - $\exp(-0,08) = 0,923$
 - en promedio, sólo el 92,3 % de la gente utilizará la bicicleta cuando esté nublado, en relación con cuando está claro
 - Si el clima empeora aún más y comienza a llover, entonces el uso medio de la bicicleta cambiará aún más en un factor de
 - $\exp(-0,5) = 0,607$
 - en promedio, sólo el 60,7% de las personas usarán bicicletas cuando llueve, en comparación con cuando está nublado

Relación Media-Varianza

- Lineal : la varianza del uso de la bicicleta siempre adquiere un valor constante.
- Poisson : $\lambda = E(Y) = Var(Y)$
 - Asumimos implícitamente que el uso medio de la bicicleta en una hora determinada = a la varianza del uso de la bicicleta durante esa hora
- Datos de Bicicleta Compartida: la media y la varianza en el uso de bicicletas son mucho más altas cuando en condiciones favorables que en condiciones desfavorables
- El modelo de regresión de Poisson es capaz de manejar la relación media-varianza observada en los datos
- El modelo de regresión lineal no lo es

Valores Ajustados no Negativos

- No existen predicciones negativas utilizando el modelo de regresión de Poisson
 - el modelo de Poisson sólo permite valores no negativos
- Cuando ajustamos un modelo de regresión lineal al conjunto de datos de Bicicleta Compartida,
 - casi el 10% de las predicciones son negativas

4.6.3

Modelos Lineales Generalizados en Mayor Generalidad

- Hasta ahora hemos analizado tres tipos de modelos de regresión:
 - Lineal,
 - Logístico y
 - de Poisson.
- Comparten algunas características comunes:
 - Utilizan predictores X_1, \dots, X_p para predecir una respuesta Y
 - Modela la media de Y en función de los predictores

Distribución de la Respuesta Y

- Suponemos que, condicionado a X_1, \dots, X_p , Y pertenece a una determinada familia de distribuciones
- *Regresión Lineal*
 - Y : distribución normal o gaussiana
- *Regresión Logística*
 - Y : distribución de Bernoulli
- *Regresión de Poisson*
 - Y : distribución de Poisson

Media de la Respuesta Y

- *Regresión Lineal*

$$E(Y|X_1, \dots, X_p) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

- *Regresión Logística*

$$\begin{aligned} E(Y|X_1, \dots, X_p) &= Pr(Y = 1 | X_1, \dots, X_p) \\ &= \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}} \end{aligned}$$

- *Regresión de Poisson*

$$E(Y|X_1, \dots, X_p) = \lambda(X_1, \dots, X_p) = e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}$$

Función de Enlace η

- η aplica una transformación a $E(Y|X_1, \dots, X_p)$ de modo que la media transformada sea una función lineal de los predictores

$$\eta(E(Y|X_1, \dots, X_p)) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

- *Regresión Lineal :* $\eta(\mu) = \mu$

- *Regresión Logística :* $\eta(\mu) = \log\left(\frac{\mu}{1 - \mu}\right)$

- *Regresión de Poisson :* $\eta(\mu) = \log(\mu)$

Familia Exponencial

- Las distribuciones Gaussiana, Bernoulli y Poisson son miembros de una clase más amplia de distribuciones, conocida como *familia exponencial*.
- Otros miembros conocidos de esta familia son la distribución
 - Exponencial,
 - Gamma y
 - Binomial Negativa

Modelo Lineal Generalizado (GLM)

- En general
 - podemos realizar una regresión modelando la respuesta Y como proveniente de un miembro particular de la familia exponencial
 - y luego transformando la media de la respuesta de manera que la media transformada sea una función lineal de los predictores

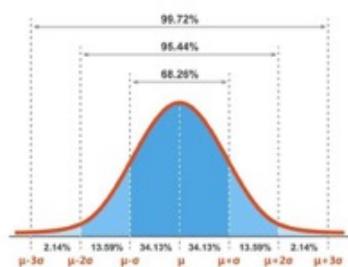
$$\eta(E(Y|X_1, \dots, X_p)) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

- Cualquier enfoque de regresión que siga esta receta general se conoce como *Modelo Lineal Generalizado (GLM)*

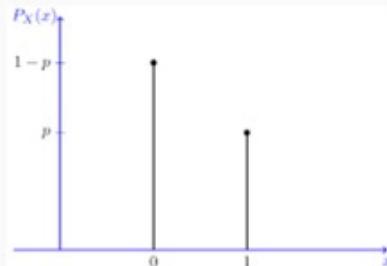
Modelo Lineal Generalizado (GLM) (2)

- Tres ejemplos de GLM
 - Regresión Lineal
 - Regresión Logística
 - Regresión de Poisson
- Otros ejemplos que no se tratan aquí incluyen
 - Regresión Gamma
 - Regresión Binomial Negativa

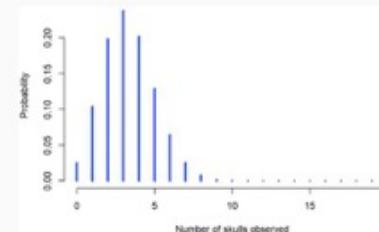
Algunas Distribuciones



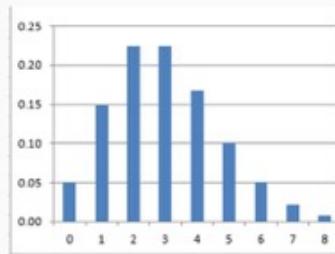
Normal Distribution



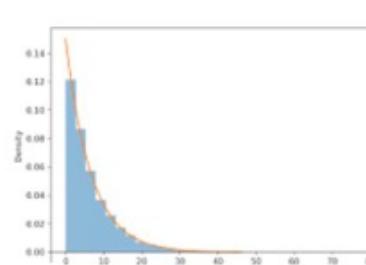
Bernoulli Distribution



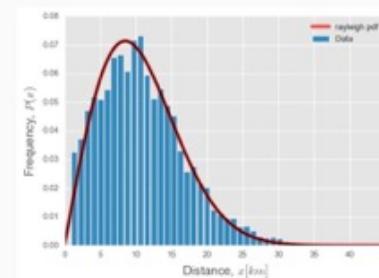
Binomial Distribution



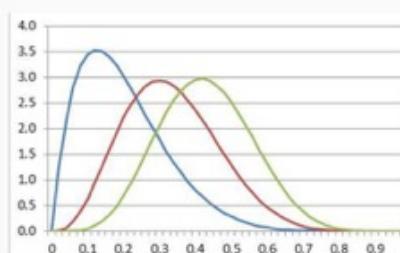
Poisson Distribution



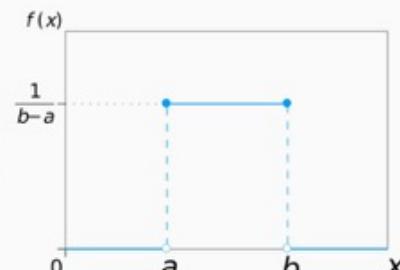
Exponential Distribution



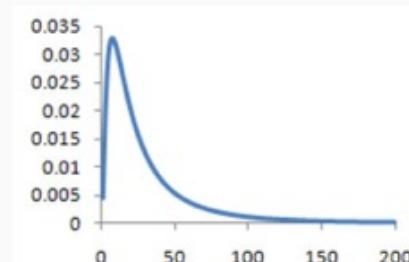
Gamma Distribution



Beta Distribution



Uniform Distribution



Log Normal Distribution

Lab: Regresión Logística, LDA, QDA y KNN

- Los datos del Mercado de Valores
- Regresión Logística
- Análisis Discriminante Lineal (LDA)
- Análisis Discriminante Cuadrático (QDA)
- Bayes Ingenuo
- K-vecinos más Cercanos (KNN)
- Regresión Lineal y de Poisson en los datos de Bicicletas Compartidas

Ejercicios