Моделі статистичного навчання: класифікація

Розглянемо три методи класифікації:

Розглянемо три методи класифікації:

Логістична регресія;

Лінійний дискримінантний аналіз;

Метод К-найближчих сусідів.

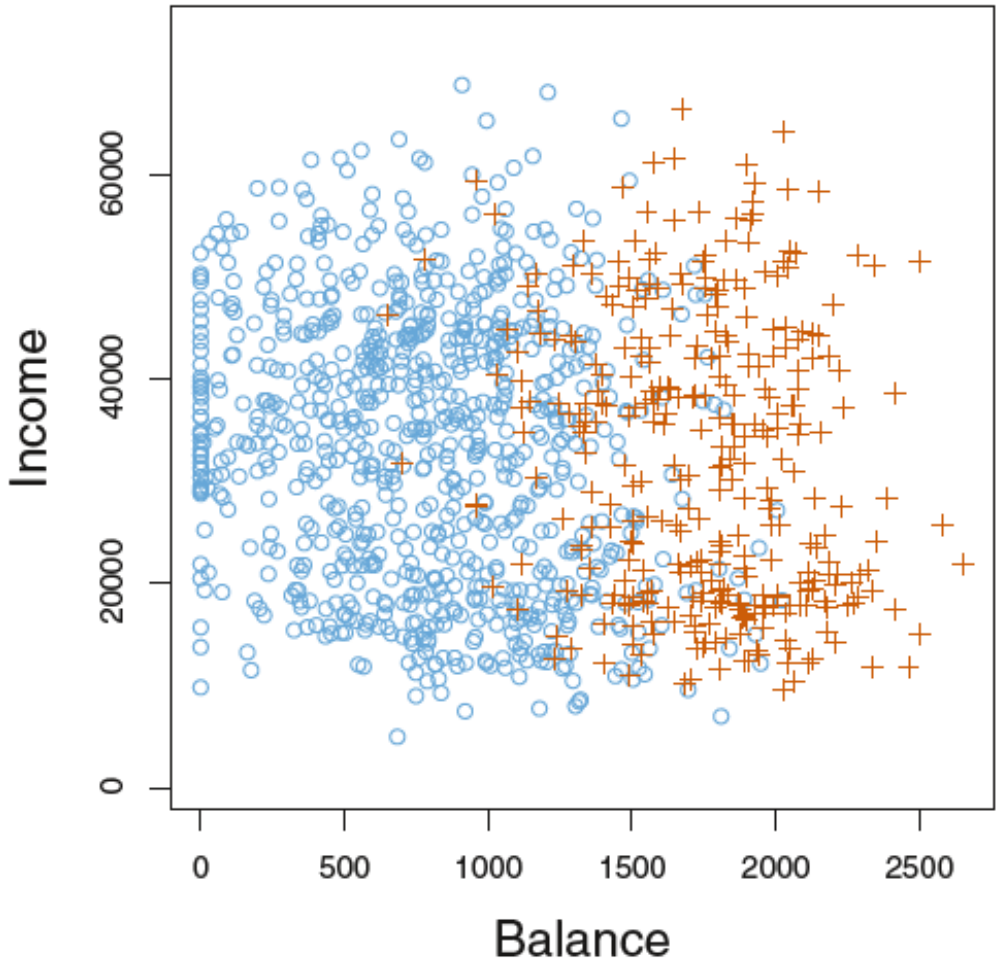
Розглянемо три методи класифікації:

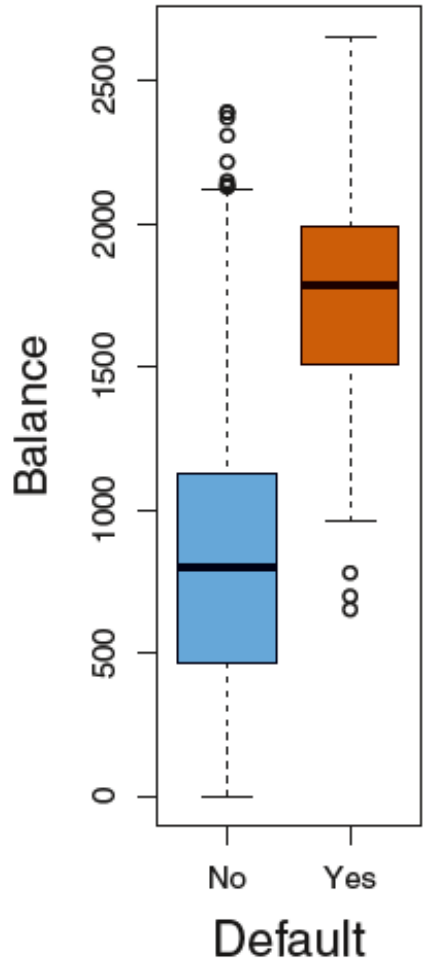
Логістична регресія;

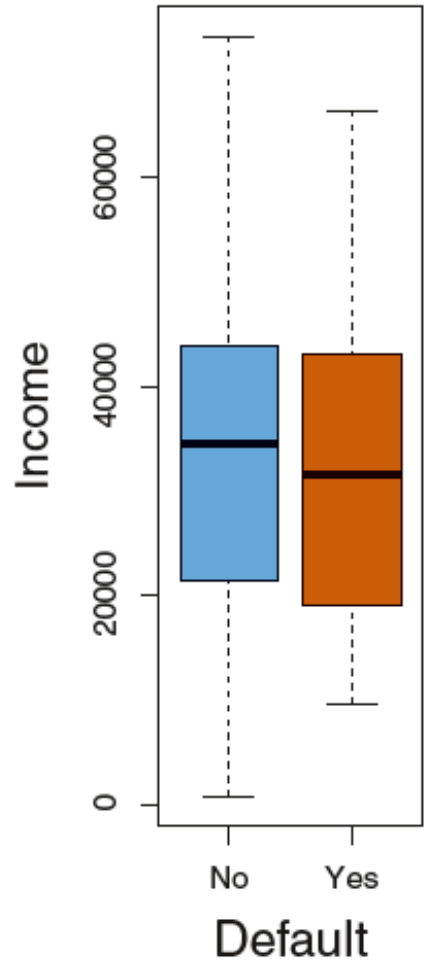
Лінійний дискримінантний аналіз;

Метод К-найближчих сусідів.

Нехай нам задано n даних, тобто пар (x_1, y_1), (x_2, y_2), …, (x_n, y_n), ми хочемо побудувати модель класифікації, яка б працювала добре не лише на навчальних даних, але й на тестових.


Income
60000
40000
20000
0
0
500
1000
1500
2000
2500
Balance


Balance
2500
2000
1500
1000
500
0
No
Yes
Default


Income
60000
40000
20000
0
No
Yes
Default

Чи можна використати метод фіктивних змінних для залежної мінної?

Чи можна використати метод фіктивних змінних для залежної мінної?

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

Чи можна використати метод фіктивних змінних для залежної мінної?

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

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

Чи можна використати метод фіктивних змінних для залежної мінної?

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

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

$$Y = \begin{cases} 0 & \text{if } \texttt{stroke}; \\ 1 & \text{if } \texttt{drug overdose}. \end{cases}$$

Логістична регресія

Логістична регресія

Залежною змінною є умовна імовірність, наприклад,

$$P(\text{default=Yes} \mid \text{balance}) = P(Y = 1 \mid X) = P(X)$$

Логістична регресія

Залежною змінною є умовна імовірність, наприклад,

$$P(\text{default=Yes} \mid \text{balance}) = P(Y = 1 \mid X) = P(X)$$

Розглянемо модель:

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

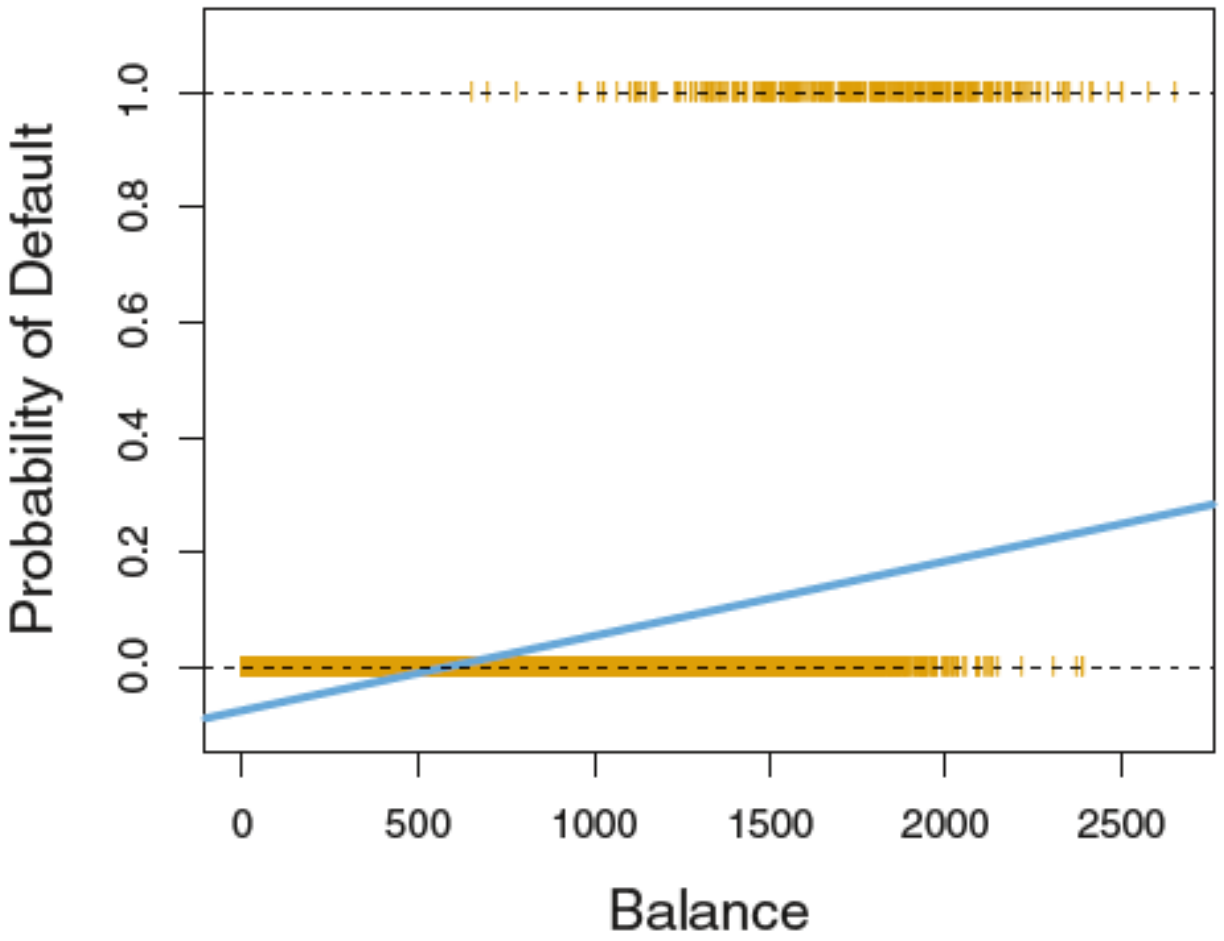
Логістична регресія

Залежною змінною є умовна імовірність, наприклад,

$$P(\text{default=Yes} \mid \text{balance}) = P(Y = 1 \mid X) = P(X)$$

Розглянемо модель:

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



Використаємо натомість логістичну функцію, тобто

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

Використаємо натомість логістичну функцію, тобто

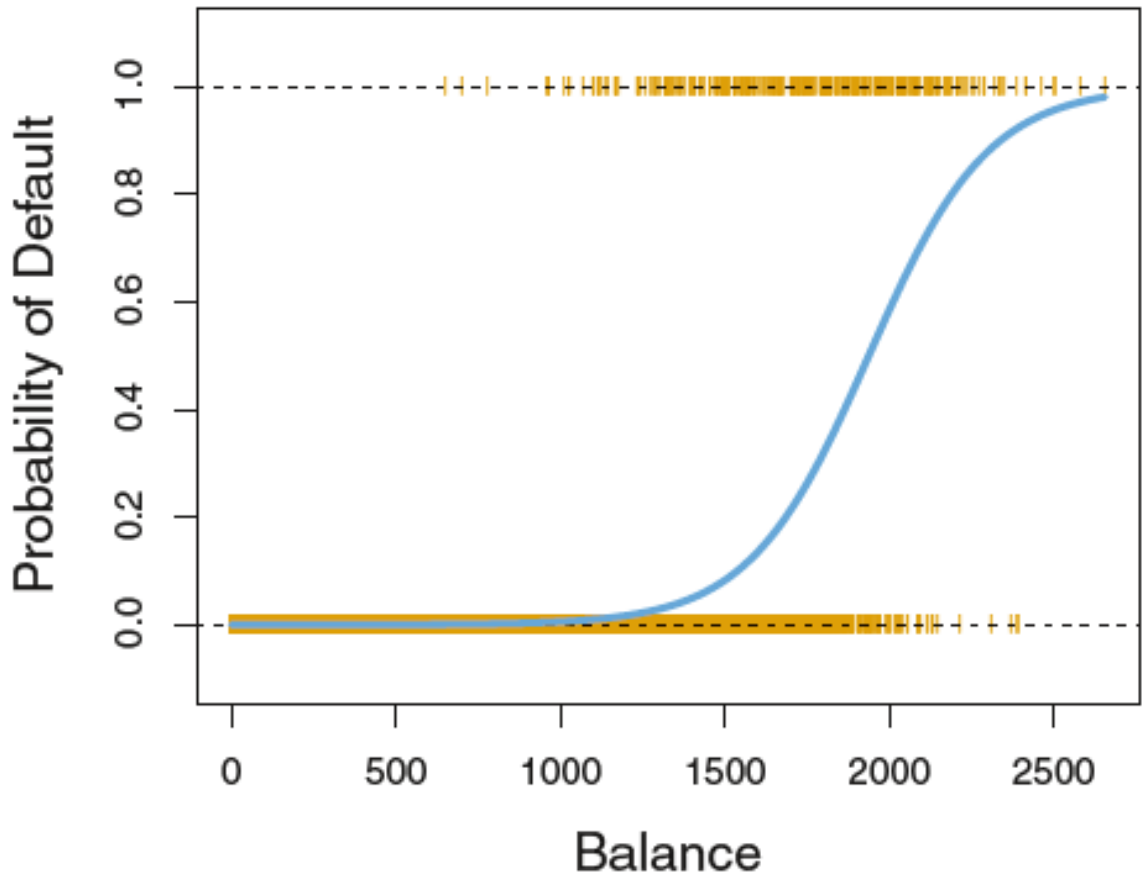
$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \qquad \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

Використаємо натомість логістичну функцію, тобто

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

Використаємо натомість логістичну функцію, тобто

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



Оцінювання моделі на основі функції правдоподібності

Оцінювання моделі на основі функції правдоподібності

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

Оцінювання моделі на основі функції правдоподібності

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

	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

Оцінювання моделі на основі функції правдоподібності

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

	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

Передбачення ймовірності дефолту при balance = 1000

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1{,}000}}{1+e^{-10.6513+0.0055\times 1{,}000}} = 0.00576$$

Оцінювання моделі на основі функції правдоподібності

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

	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

Передбачення ймовірності дефолту при balance = 1000

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1{,}000}}{1+e^{-10.6513+0.0055\times 1{,}000}} = 0.00576$$

при balance = 2000, отримаємо 0.586!

За наявності якісних незалежних змінних

За наявності якісних незалежних змінних, використовуємо метод фіктивних змінних.

За наявності якісних незалежних змінних, використовуємо метод фіктивних змінних.

	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

За наявності якісних незалежних змінних, використовуємо метод фіктивних змінних.

	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

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

Множинна (багатовимірна) логістична регресія

Множинна (багатовимірна) логістична регресія

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

Множинна (багатовимірна) логістична регресія

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

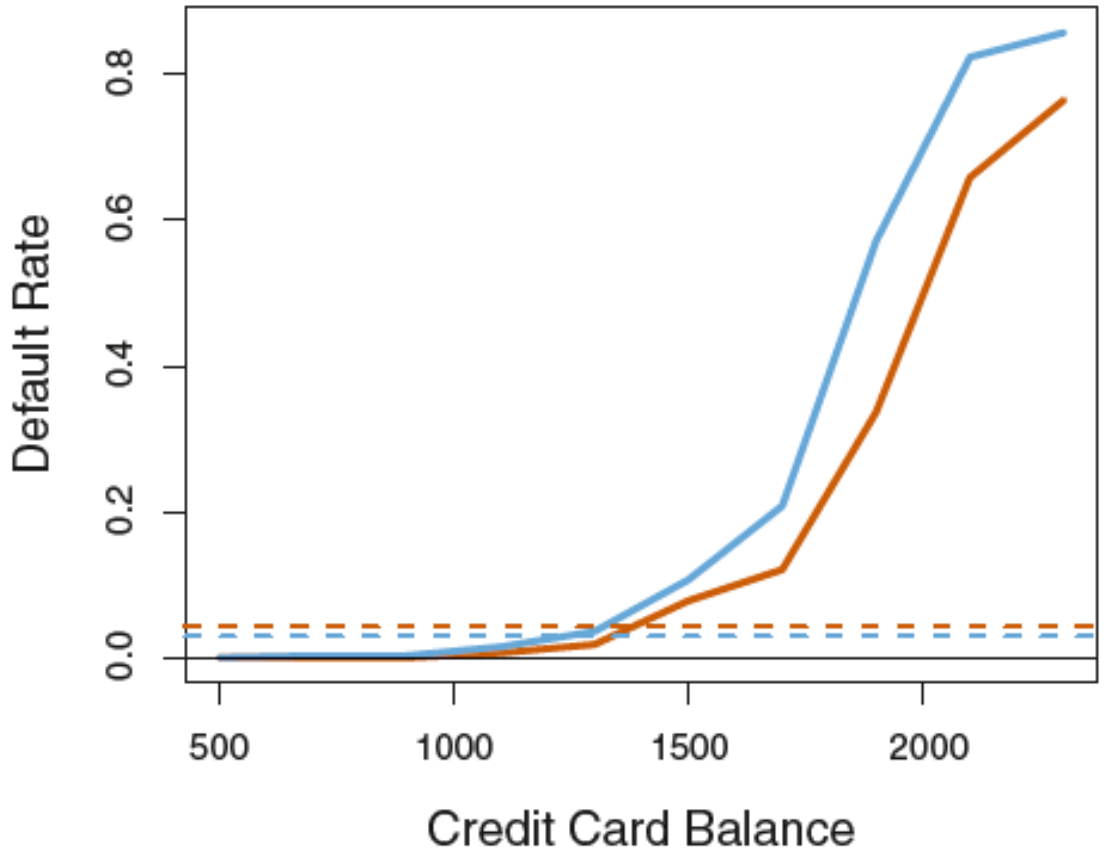
$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

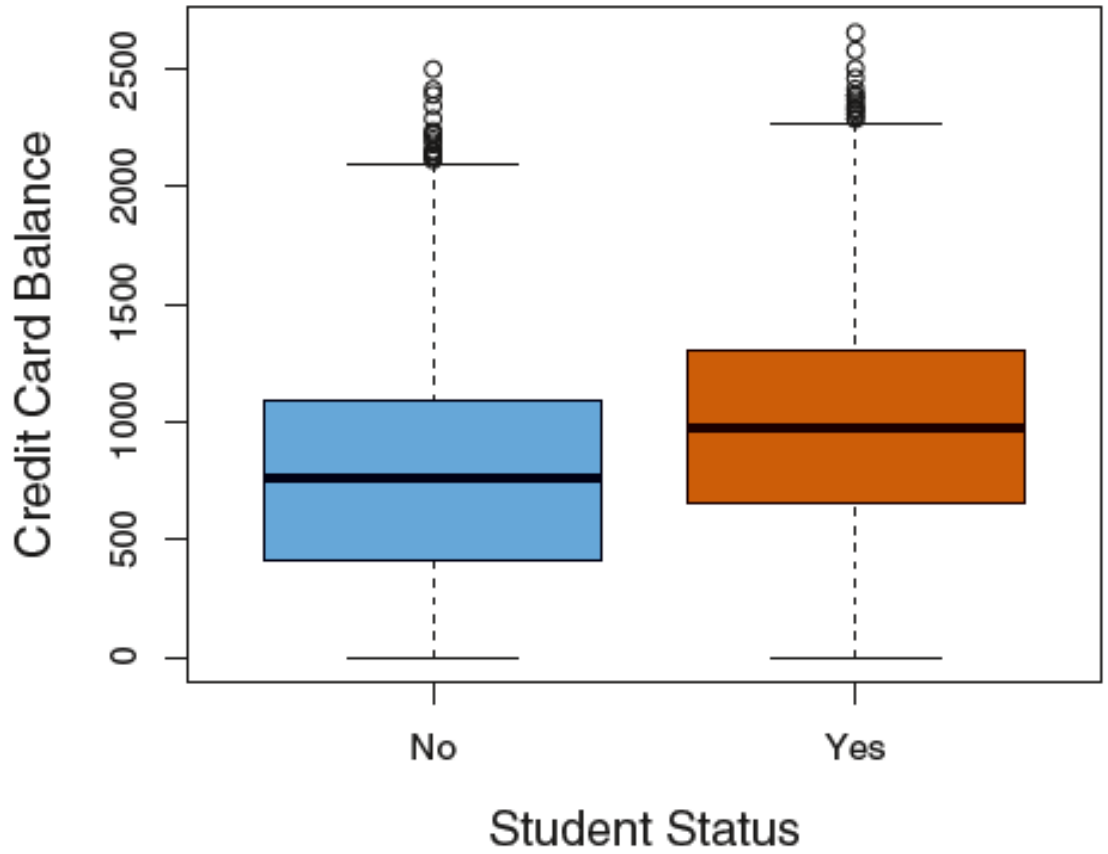
Множинна (багатовимірна) логістична регресія

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

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

	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


Default Rate
0.0
0.2
0.4
0.6
0.8
500
1000
1500
2000
Credit Card Balance


Credit Card Balance
0
500
1000
1500
2000
2500
No
Yes
Student Status

Прогнозне значення імовірності дефолту для студента з balance = 1500 та income = 40000 дорівнює

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}} = 0.058$$

Прогнозне значення імовірності дефолту для студента з balance = 1500 та income = 40000 дорівнює

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}} = 0.058$$

А для не студента з тими ж даними

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}} = 0.105$$

Лінійний дискримінантний аналіз

Лінійний дискримінантний аналіз

На відміну від логістичної регресії, тут моделюється розподіл предикторів в окремих класах відносно залежної змінної.

Лінійний дискримінантний аналіз

На відміну від логістичної регресії, тут моделюється розподіл предикторів в окремих класах відносно залежної змінної.

Обгрунтування

Лінійний дискримінантний аналіз

На відміну від логістичної регресії, тут моделюється розподіл предикторів в окремих класах відносно залежної змінної.

Обгрунтування

Коли класи добре розділені, оцінки параметрів моделі логістичної регресії є нестійкі.

Лінійний дискримінантний аналіз

На відміну від логістичної регресії, тут моделюється розподіл предикторів в окремих класах відносно залежної змінної.

Обгрунтування

Коли класи добре розділені, оцінки параметрів моделі логістичної регресії є нестійкі.

Якщо *n* мале і розподіл предикторів X є приблизно нормальний для кожного з класів, то результати лінійної дискримінантної моделі також є стійкіші, ніж для моделі логістичної регресії.

Лінійний дискримінантний аналіз

На відміну від логістичної регресії, тут моделюється розподіл предикторів в окремих класах відносно залежної змінної.

Обгрунтування

Коли класи добре розділені, оцінки параметрів моделі логістичної регресії є нестійкі.

Якщо *n* мале і розподіл предикторів X є приблизно нормальний для кожного з класів, то результати лінійної дискримінантної моделі також є стійкіші, ніж для моделі логістичної регресії.

Лінійний дискримінантний аналіз використовується частіше коли ми маємо більше двох класів для залежної змінної.

Теорема Байеса для класифікації

Теорема Байеса для класифікації

Нехай π_k – імовірність, що випадково вибране спостереження належить класу k.

Нехай $f_k(X) = \mathrm{P}(X = x \mid Y = k)$ – густина розподілу X в межах класу k.

Теорема Байеса для класифікації

Нехай π_k – імовірність, що випадково вибране спостереження належить класу k.

Нехай $f_k(X) = \mathrm{P}(X = x \mid Y = k)$ – густина розподілу X в межах класу k.

З теореми Байеса отримуємо

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

Теорема Байеса для класифікації

Нехай π_k – імовірність, що випадково вибране спостереження належить класу k.

Нехай $f_k(X) = \mathrm{P}(X = x \mid Y = k)$ – густина розподілу X в межах класу k.

З теореми Байеса отримуємо

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

Нехай $p = 1$, припустимо, що $f_k(X)$ є густиною нормального розподілу, тобто

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

Припустимо, що $\sigma_1 = \sigma_2 = \ldots = \sigma_k = \sigma$, тоді

Припустимо, що $\sigma_1 = \sigma_2 = \ldots = \sigma_k = \sigma$, тоді

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

Припустимо, що $\sigma_1 = \sigma_2 = \ldots = \sigma_k = \sigma$, тоді

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

Спостереження $X = x$ попадає в той клас, для якого попередня імовірність є найбільшою.

Припустимо, що $\sigma_1 = \sigma_2 = \ldots = \sigma_k = \sigma$, тоді

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

Спостереження $X = x$ попадає в той клас, для якого попередня імовірність є найбільшою. Взявши логарифм, отримаємо

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Припустимо, що $\sigma_1 = \sigma_2 = \ldots = \sigma_k = \sigma$, тоді

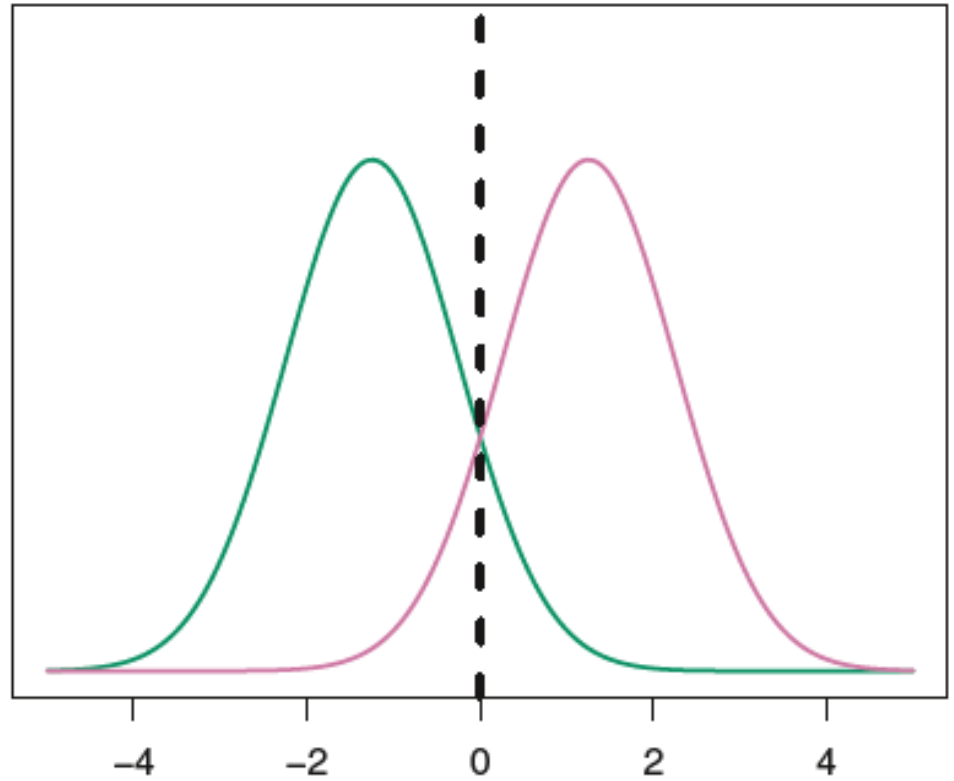
$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

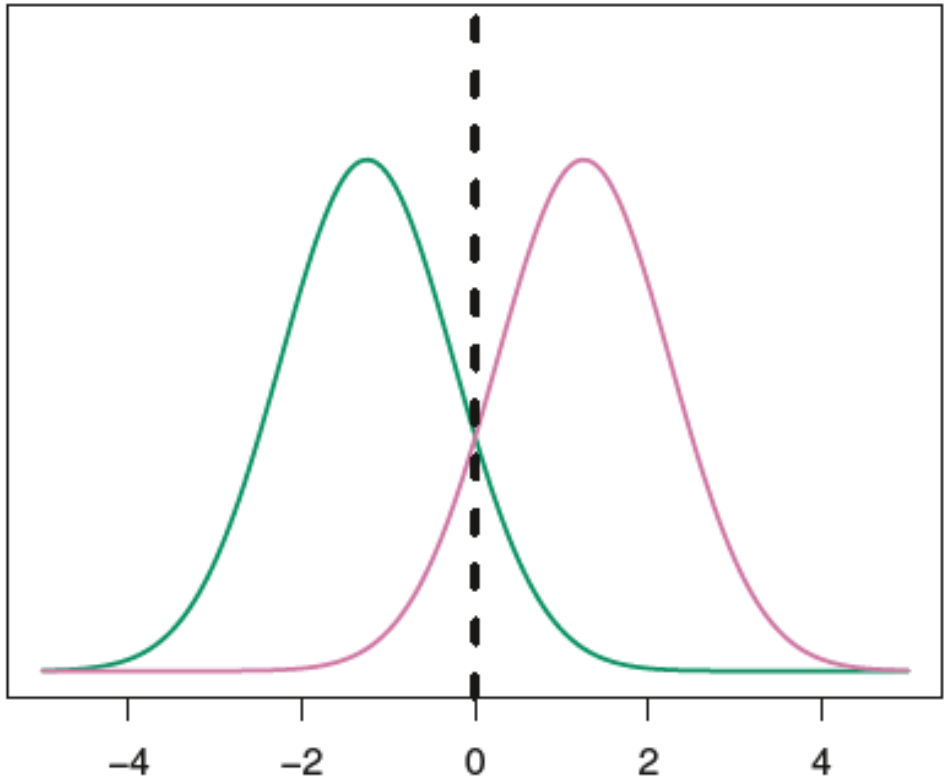
Спостереження $X = x$ попадає в той клас, для якого попередня імовірність є найбільшою. Взявши логарифм, отримаємо

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Наприклад, нехай $K = 2$, $\pi_1 = \pi_2 = 1/2$, тоді якщо $2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$ то спостереження попадає в клас 1, інакше – 2.

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$


−4
−2
0
2
4



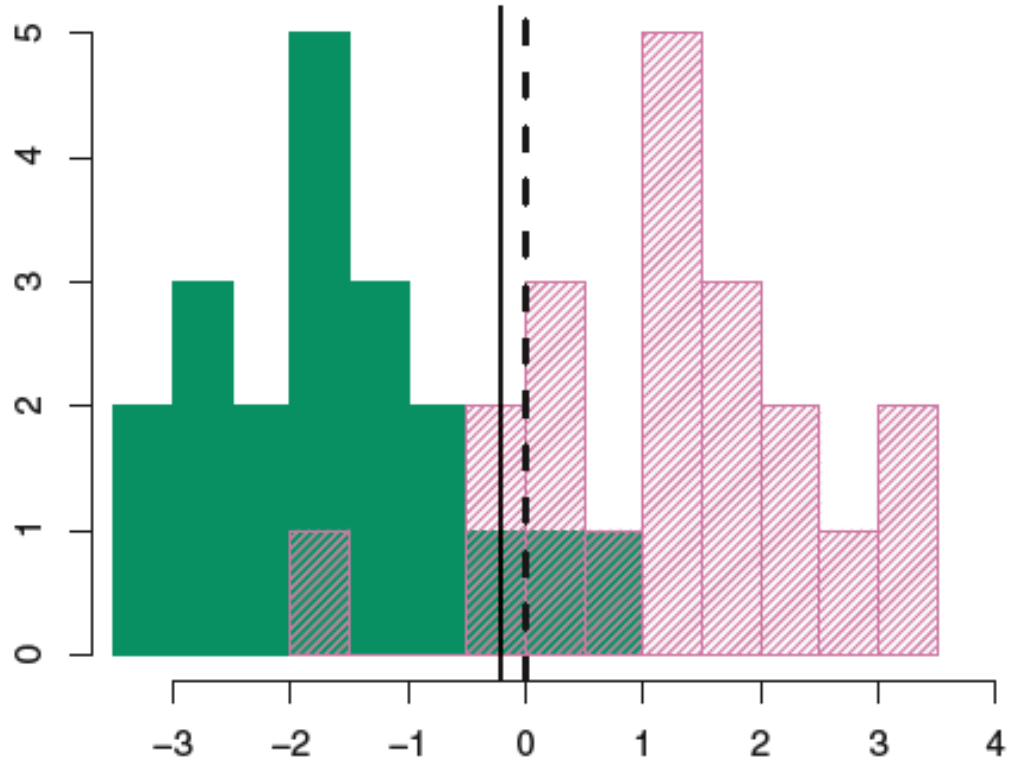
Оцінки невідомих параметрів отримуємо з:

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

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

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

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

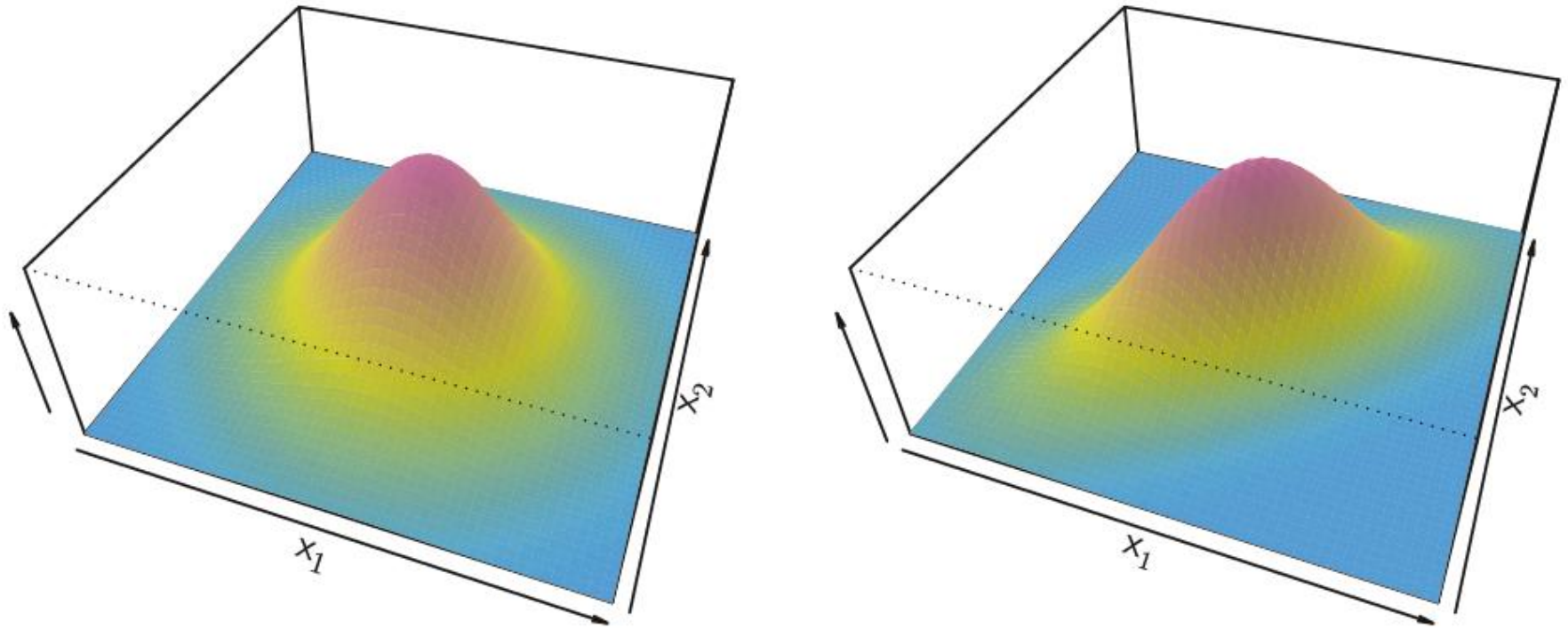


Тестова помилка становить 11,1 % в порівнянні з мінімальним значенням 10,5 %

Нехай тепер $p > 1$.

Нехай тепер $p > 1$.

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



Спостереження $X = x$ попадає в той клас, для якого наступна величина є найбільшою.

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

Спостереження $X = x$ попадає в той клас, для якого наступна величина є найбільшою.

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

$$\hat{\mu}_k = \sum_{g_i=k} x_i / N_k;$$

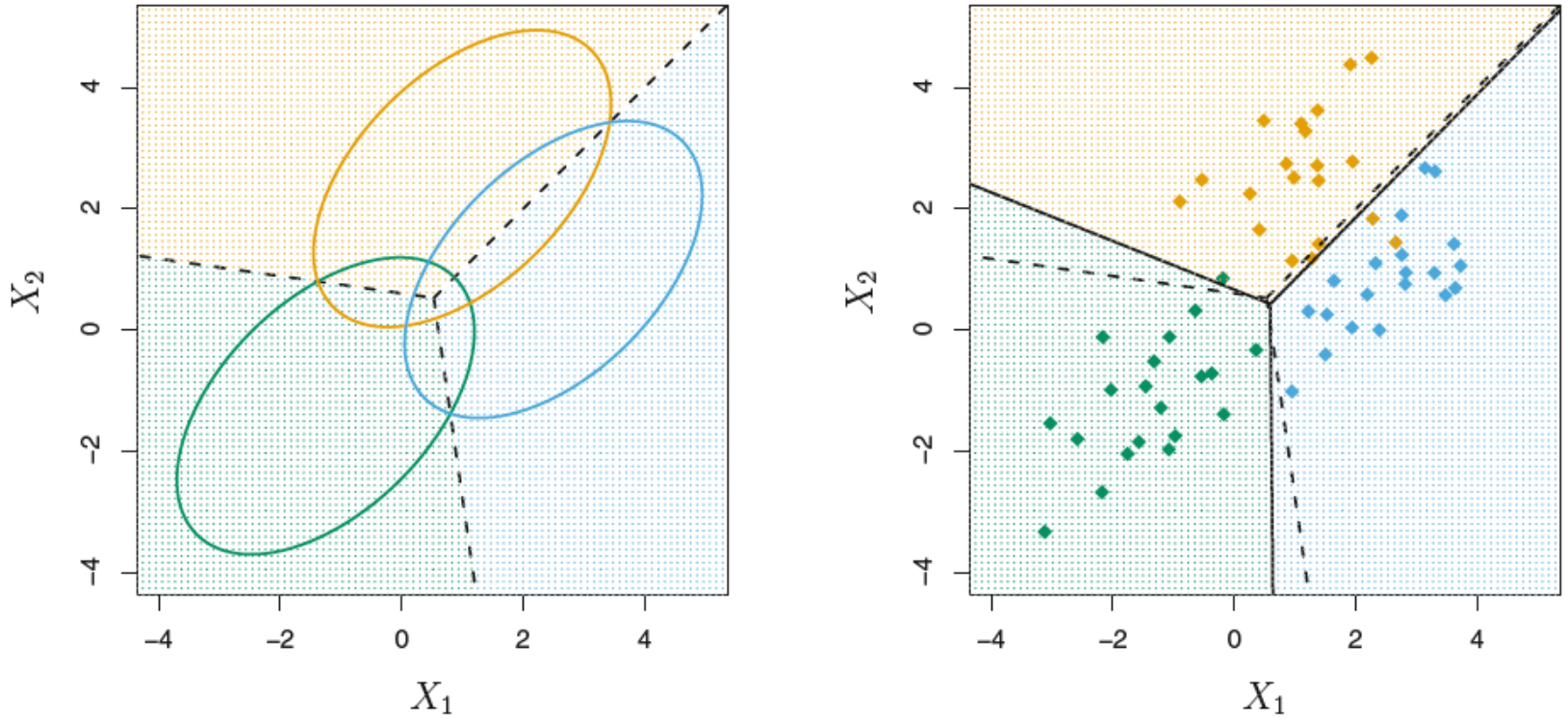
$$\hat{\mathbf{\Sigma}} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K)$$

Спостереження $X = x$ попадає в той клас, для якого наступна величина є найбільшою.

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

$$\hat{\mu}_k = \sum_{g_i=k} x_i / N_k;$$

$$\hat{\mathbf{\Sigma}} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K)$$



Приклад з даними Default.

Приклад з даними Default. Отримали помилки на навчальних даних на рівні 2.75%.

Перш за все, рівень помилок у навчальній вибірці зазвичай буде нижчим, ніж помилки у тестових даних.

Приклад з даними Default. Отримали помилки на навчальних даних на рівні 2.75%.

Перш за все, рівень помилок у навчальній вибірці зазвичай буде нижчим, ніж помилки у тестових даних.

По-друге, оскільки лише 3,33% осіб дефолтували у вибірці для навчання, то простий класифікатор, який завжди передбачає, що кожна особа не матиме дефолту, незалежно від даних, призведе до рівня помилок 3,33%.

Приклад з даними Default. Отримали помилки на навчальних даних на рівні 2.75%.

Перш за все, рівень помилок у навчальній вибірці зазвичай буде нижчим, ніж помилки у тестових даних.

По-друге, оскільки лише 3,33% осіб дефолтували у вибірці для навчання, то простий класифікатор, який завжди передбачає, що кожна особа не матиме дефолту, незалежно від даних, призведе до рівня помилок 3,33%.

		True default status		
		No	Yes	Total
Predicted	No	9,644	252	9,896
default status	Yes	23	81	104
	Total	9,667	333	10,000

Чутливість моделі дорівнює відсотку ідентифікованих осіб, що дефолтнули серед тих, що справді дефолтнули і становить всього 24,3 %.

Чутливість моделі дорівнює відсотку ідентифікованих осіб, що дефолтнули серед тих, що справді дефолтнули і становить всього 24,3 %.

Специфікація моделі дорівнює відсотку ідентифікованих осіб, що не дефолтнули серед тих, що справді не дефолтнули і становить аж 99,8 %.

Чутливість моделі дорівнює відсотку ідентифікованих осіб, що дефолтнули серед тих, що справді дефолтнули і становить всього 24,3 %.

Специфікація моделі дорівнює відсотку ідентифікованих осіб, що не дефолтнули серед тих, що справді не дефолтнули і становить аж 99,8 %.

Для покращення результатів передбачення дефолту доцільно змінити межу імовірності, тобто замість

$$\Pr(\texttt{default} = \texttt{Yes}|X = x) > 0.5$$

використати, наприклад,

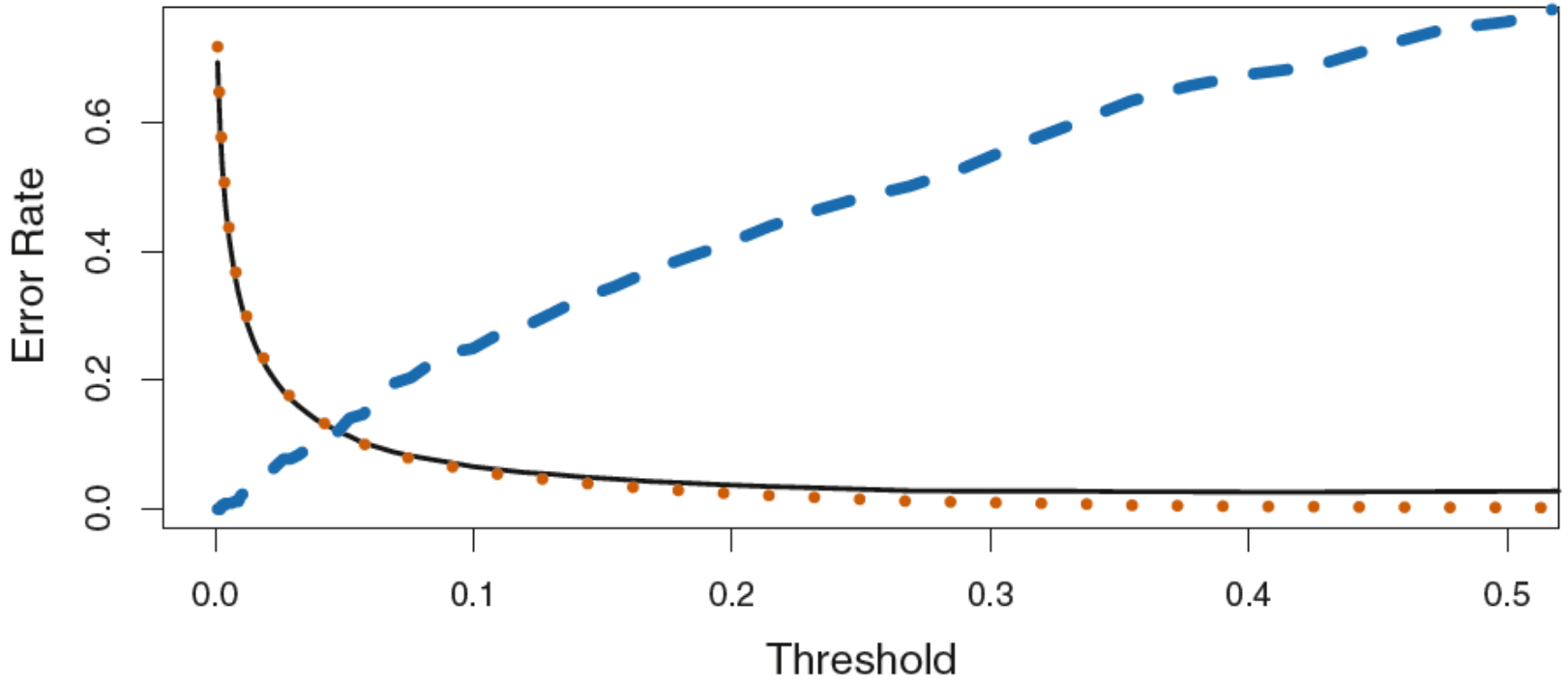
$$P(\texttt{default} = \texttt{Yes}|X = x) > 0.2$$

		True default status		
		No	Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

Тепер помилково ідентифіковані особи, що дефолтнули становлять 41,4 % в порівнянні з 75,7 % в попередньому випадку.

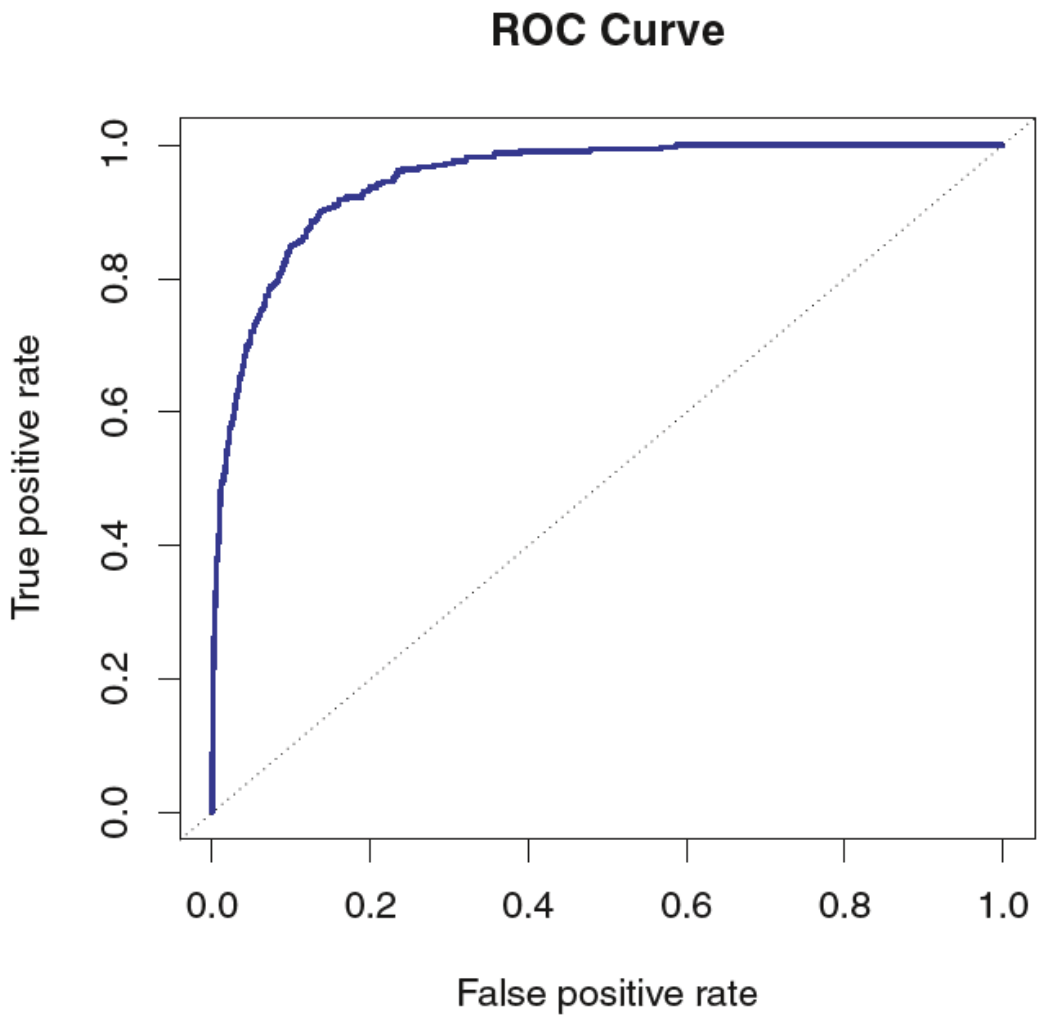
		True default status		
		No	Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

Тепер помилково ідентифіковані особи, що дефолтнули становлять 41,4 % в порівнянні з 75,7 % в попередньому випадку.



ROC крива для класифікатора LDA для Default даних. Відображає два типи помилок. True positive rate - це чутливість: частка осіб, що дефолтнули, які правильно визначені. False positive rate - 1 – специфічність: частка осіб, що не дефолтнули, яких ми класифікуємо, як особи, що дефолтнули.

ROC крива для класифікатора LDA для Default даних. Відображає два типи помилок. True positive rate - це чутливість: частка осіб, що дефолтнули, які правильно визначені. False positive rate - 1 – специфічність: частка осіб, що не дефолтнули, яких ми класифікуємо, як особи, що дефолтнули.



		Predicted class		
		− or Null	+ or Non-null	Total
True	− or Null	True Neg. (TN)	False Pos. (FP)	N
class	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

		Predicted class		
		− or Null	+ or Non-null	Total
True	− or Null	True Neg. (TN)	False Pos. (FP)	N
class	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

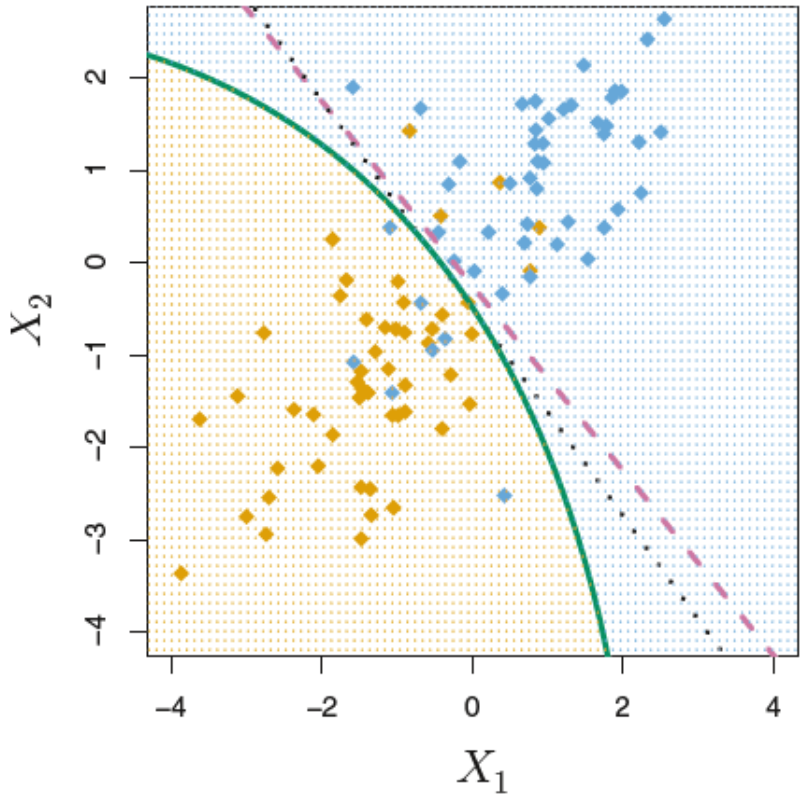
Квадратичний дискримінантний аналіз

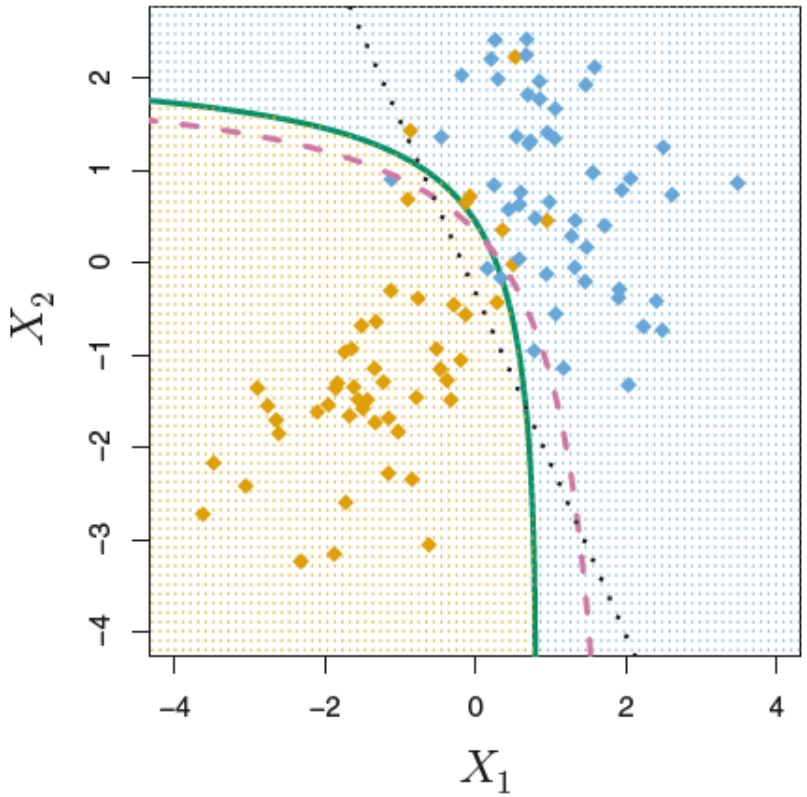
Квадратичний дискримінантний аналіз

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

Квадратичний дискримінантний аналіз

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





Порівняння логістичної регресії, лінійного та квадратичного дискримінантного аналізу та методу К-найближчих сусідів.

Порівняння логістичної регресії, лінійного та квадратичного дискримінантного аналізу та методу К-найближчих сусідів.

Для лінійного дискримінантного аналізу маємо

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x$$

Порівняння логістичної регресії, лінійного та квадратичного дискримінантного аналізу та методу К-найближчих сусідів.

Для лінійного дискримінантного аналізу маємо

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x$$

Для логістичної регресії:

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 x$$

Розглянемо 6 сценаріїв. У трьох сценаріях межа рішення Баєса є лінійною, а в інших сценаріях – нелінійною.

Для кожного сценарію згенеруємо 100 випадкових навчальних наборів даних. На кожному з цих навчальних наборів ми оцінимо кожен з методів та обчислимо результуючу тестову помилку на великому наборі тестових даних.

Ми використаємо метод К-найближчих сусідів з двома значеннями K: $K = 1$, та оціненим значенням.

Сценарій 1: У кожному з двох класів є 20 навчальних спостережень. Спостереження в кожному класі є некорельованими випадковими нормальними величинами з різним середнім значенням у кожному класі.

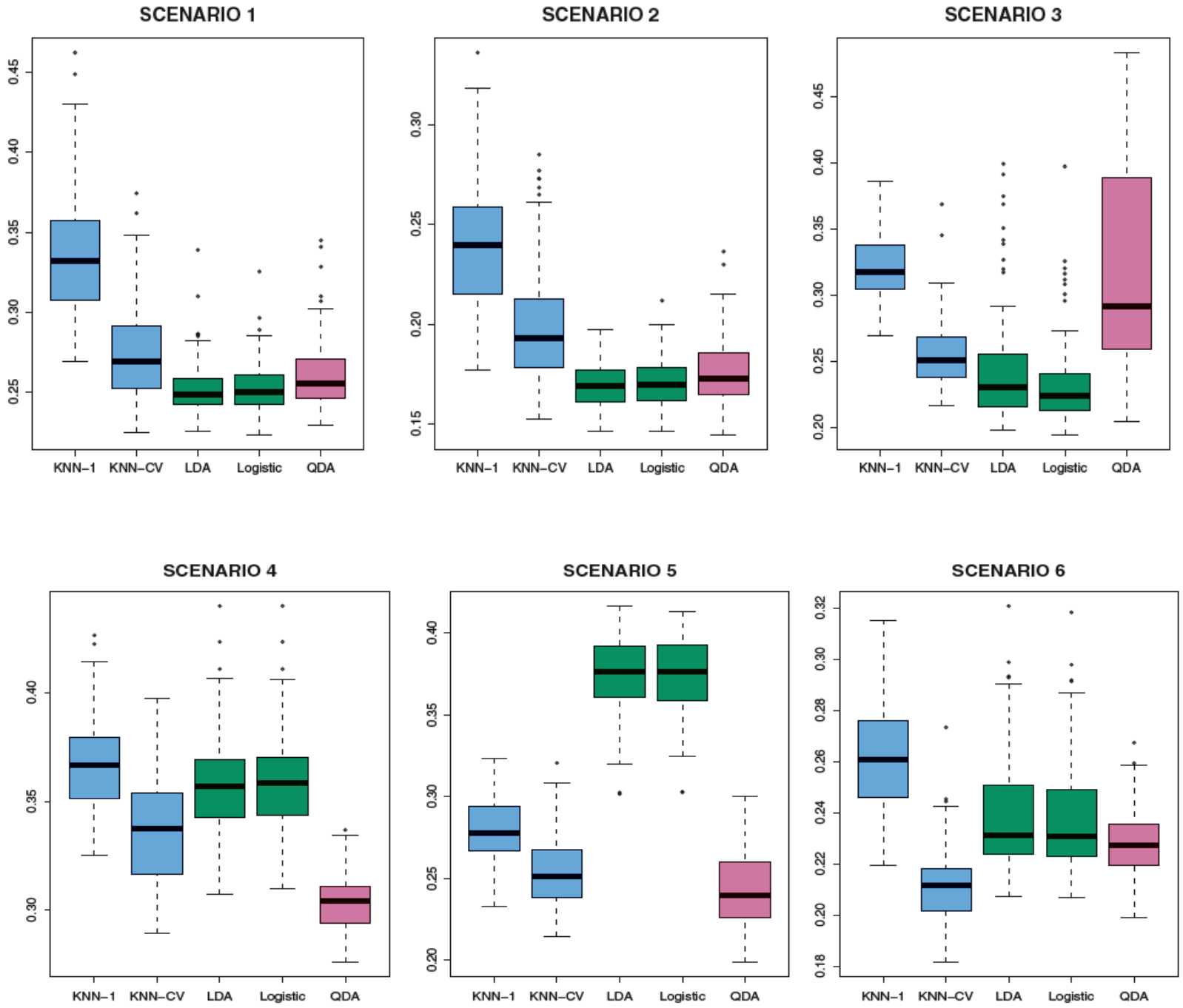
Сценарій 2: Вся як і в сценарії 1, за винятком, що в межах кожного з класів предиктори мають кореляцію -0,5.

Сценарій 3: Обидва предиктори згенеруємо з t-розподілу по 50 спостережень на клас.

Сценарій 4: Дані згенеровані з нормального розподілу, з кореляцією 0,5 між предикторами першого класу, і кореляцією −0,5 між предикторами другого класу.

Сценарій 5: В межах кожного класу спостереження згенеровані на основі нормального розподіл з некорельованими предикторами. Однак значення залежної змінної взяті з логістичної функції з предикторами X_1^2, X_2^2, $X_1 \times X_2$.

Сценарій 6: Як і в сценарії 5, але значення залежної змінної отримані на основі більш складної нелінійної функції.


SCENARIO 1
SCENARIO 2
SCENARIO 3
SCENARIO 4
SCENARIO 5
SCENARIO 6
KNN−1
KNN−CV
LDA
Logistic
QDA