Лекція 8. Логістична регресія [1].

Розглянемо випадок, який виразно ілюструє регресійні задачі з дискретними даними. Професором університету США штату Айова Кеном Коглером було встановлено, що на стать новонароджених черепашок значним чином впливає навколишнє середовище. Ним був проведений такий експеримент. Черепашачі яйця (одного виду) були перевезені з Іллінойсу та розподілені по декілька штук у спеціальні бокси, які вигрівалися при різних температурах — від 27,2 °C до 29,9 °C — по три бокси при кожній температурі. Коли черепашки народилися то стать новонароджених в залежності від температури розподілилася таким чином

Temp(°C)	Male	Female	Temp(°C)	Male	Female	Temp(°C)	Male	Female
27.2	1	9	27.2	0	8	27.2	1	8
27.7	7	3	27.7	4	2	27.7	6	2
28.3	13	0	28.3	6	3	28.3	7	1
28.4	7	3	28.4	5	3	28.4	7	2
29.9	10	1	29.9	8	0	29.9	9	0

Загальна кількість черепашок чоловічої статі становила $\frac{91}{136} = 0.67$. Коли температура була нижчою за $27.5\,^{o}$C, їхня частка становила $\frac{2}{27} = 0.07$, при температурі нижчій за $28\,^{o}$C — $\frac{19}{51} = 0.37$, коли температура була нижчою за 28.5 — $\frac{64}{108} = 0.59$ і при температурі, нижчій за 30.0 — $\frac{91}{136} = 0.67$ черепашок були чоловічої статі. Тобто було встановлено, що при підвищенні температури серед новонароджених переважають черепашки чоловічої статі.

Логістична регресія.

Тепер розглянемо безпосередньо логістичну регресію та схему її побудови. Для цього потрібна формалізація практичних даних.

Введемо позначення: $Y_i = 1$ - подія народження черепашки чоловічої статі; $Y_i = 0$ - подія народження черепашки жіночої статі; $P(Y_i = 1) = p_i$, $P(Y_i = 0) = 1 - p_i$ - імовірності подій $Y_i = 1$ та $Y_i = 0$; $E(Y_i) = 0 \cdot (1 - p_i) + 1 \cdot (p_i) = p_i$ - математичне сподівання події Y_i.

Задача полягає у передбаченні $P(Y_i = 1) = p_i$ на основі попередніх відомостей. Чи можна тут скористатися лінійною регресією $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i = p_i$? Нажаль, за умови дискретності даних виникають деякі проблеми, а саме:

1. Лінійна регресія не задовольняє умову $0 \le E(Y_i \mid X_i) = p_i \le 1$
2. Необхідною умовою для лінійної регресії є сталість дисперсії відгуків. У випадку ж дискретних даних маємо, що $D(Y_i) = p_i(1 - p_i)$, тобто її значення залежить від X_i.
3. Ще одна необхідна умова лінійної регресії – нормальний розподіл похибки : $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$. Коли ж $Y_i = 1$ маємо: $\varepsilon_i = 1 - (\beta_0 + \beta_1 X_i)$

Коли дані є дискретними графік очікуваних відгуків має вигляд деякої S-кривої, яку називають логістичною кривою. Модель логістичної регресії має вигляд:

$$E(Y_i \mid X_i) = p_i = \frac{e^{(\beta_0 + \beta_1 X_i)}}{1 + e^{(\beta_0 + \beta_1 X_i)}}$$

Досліджувані параметри знаходяться у степені експоненти, тому спочатку потрібно звести модель до вигляду, коли p_i будуть залежати від X_i лінійно, а потім повернутися до оригінального вигляду моделі щоб відобразити реальну залежність. Такі перетворення мають вигляд:

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

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

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

$$\frac{p}{1-p} \neq 1 \text{ тому } \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X$$

Тобто, маємо лінійну залежність від X.

Тепер наблизимо імовірність події p_i її частотою h_i. З рівності

$$\ln(\frac{h_i}{1-h_i}) = \beta_0 + \beta_1 X$$

знайдемо регресійні коефіцієнти і повернувшись до вихідної моделі зможемо обчислити передбачувану імовірність

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

Приклад 8.1. Нижче наведена таблиця є прикладом побудови логістичної регресії для випадку статі новонароджених черепашок.

Таблиця 8.1. — Дані про народження черепашок (температура повітря, стать тощо)

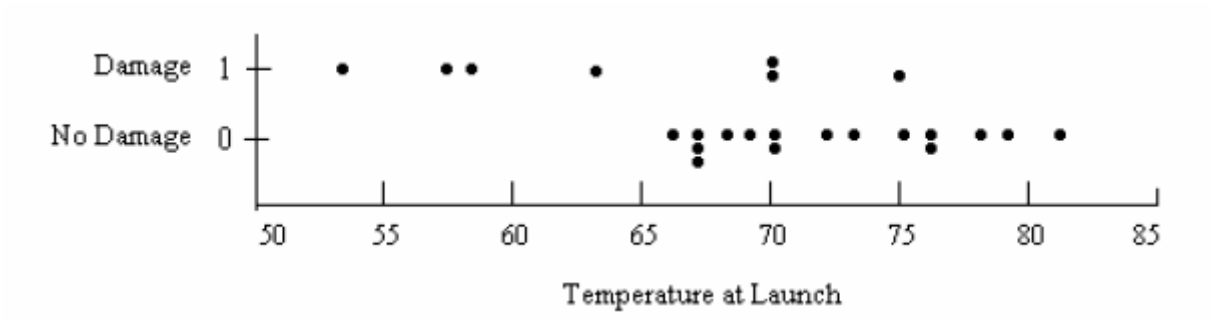
Temp	Male	Female	Total	h_i	$\ln(\frac{h_i}{1-h_i})$	Pmale
27.2	2	25	27	0.074	-2.5257	0.0741
27.7	17	7	24	0.708	0.8873	0.7083
28.3	26	4	30	0.867	1.8718	0.8667
28.4	19	8	27	0.704	0.8650	0.7037
29.9	27	1	28	0.964	3.2958	0.9643

Приклад 8.2.
28 грудня 1986 року космічний челнок Challenger вибухнув у повітрі внаслідок пошкодження кільця ущільнення на пальному баці. Це був 25-й старт челнока. На протязі попередніх 24 польотів було зафіксовано 7 випадків ушкодження кільця, у 16 випадках пошкодження не було, і в одному випадку дані отримані не були. Виникає питання: чи було пов'язане ушкодження кільця ущільнення із холодною погодою під час старту?

Наведені нижче дані були отримані із звіту про катастрофу Presidential Commission on the Space Shuttle Challenger Accident (1986). Ці дані містять номер польоту, температуру повітря під час польоту (° F), а також індикатор пошкодження кільця ущільненні паливного баку (Ні = 0, Так = 1). Температура повітря під час катастрофічного запуску челнока STS 51-L (Challenger) складала 31°F.

Таблиця 9.1. — Дані про старти (номер, температура повітря індикатор ушкодження)

Старт	^{0}F	Інд.	Старт	^{0}F	Інд.	Старт	^{0}F	Інд.
1	66	0	9	70	0	17	75	0
2	70	1	10	57	1	18	70	0
3	69	0	11	63	1	19	81	0
4	80	Не відомо	12	70	1	20	76	0
5	68	0	13	78	0	21	79	0
6	67	0	14	67	0	22	75	1
7	72	0	15	53	1	23	76	0
8	73	0	16	67	0	24	58	1



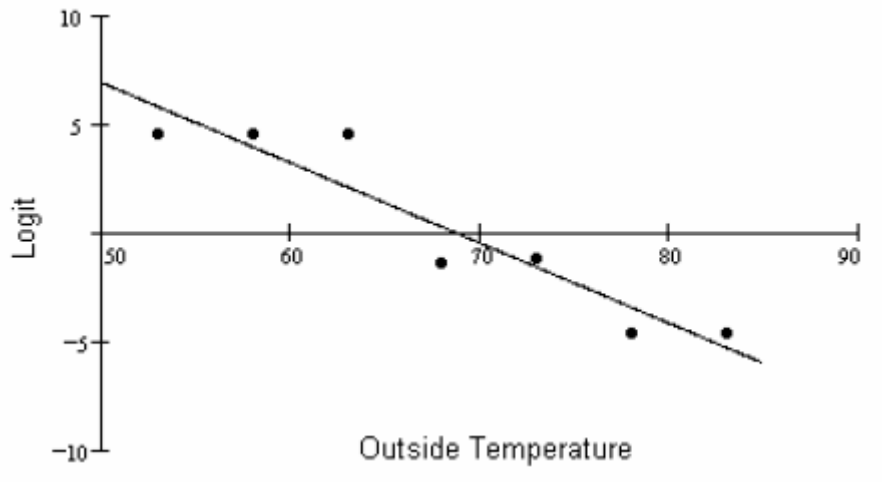
Розбивши діапазон температури на групи по 5 градусів, отримаємо таку таблицю.

Таблиця 9.3. Розрахунки за логістичною регресією

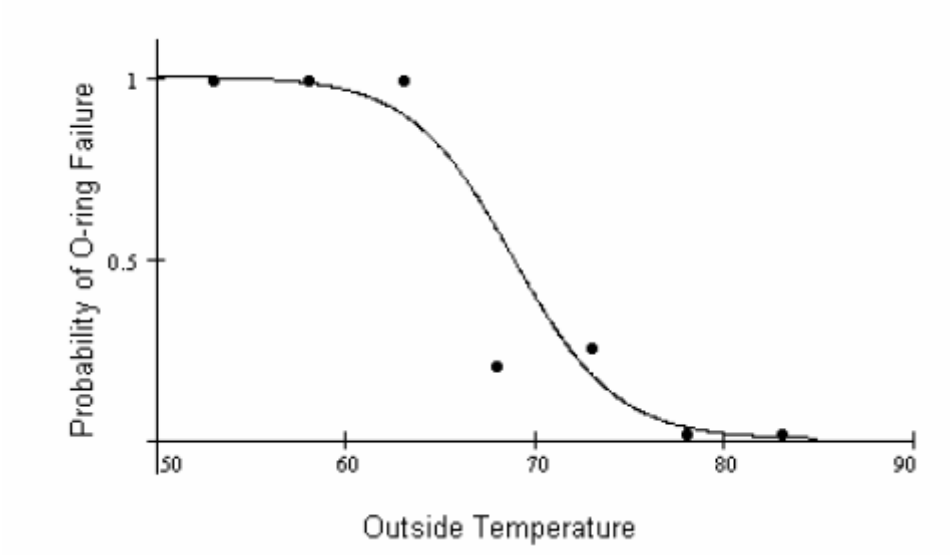
Інтервал	(51,55)	(56,60)	(61,65)	(66, 70)	(71,75)	(76,80)	(81, 85)
Температура	53	58	63	68	73	78	83
Ймовірність	0,99	0,99	0,99	0,20	0,25	0,01	0,01
Logit	4,595	4,595	4,595	-1,386	-1,099	-4,595	-4,595

За допомогою лінійної регресії будуємо аппроксимацію

$$\ln\frac{p}{1-p} = 25,386 - 0,369Temp$$



Переходячи до ймовірності, отримуємо $\hat{P} = \frac{e^{25,386-0,369Temp}}{1+e^{25,386-0,369Temp}}$.



Застосування логістичної регресії для класифікації: ROC-аналіз [2]

Для бінарної класифікації дуже часто використовується ROC-крива (Receiver Operator Characteristic), яка демонструє залежність кількості правильно класифікованих зразків від кількості невірно класифікованих зразків. Перша група висновків називаються хибнопозитивними, а друга — хибнонегативними. Вважається, що існує параметр, за допомогою якого можно отримати розбиття на два класи. Цей параметр називається точкою відсікання (cut-off value). Від цієї точки залежить величини помилок 1-го і 2-го роду.

Поняття помилок першого та другого роду будується на основі таблиці спряженості:

	Насправді	
За моделлю	**Так**	**Ні**
Так	TP	FP
Ні	FN	TN

- TP (*True Positives*) – правильно класифіковані зразки (істинно-позитивні випадки);
- TN (*True Negatives*) — правильно класифіковані негативні приклади (істинно негативні випадки);
- FN (*False Negatives*) – позитивні зразки, класифіковані як негативні. Це помилка 1-го роду (хибно-позитивні зразки);
- FP (*False Positives*) – негативні зразки, класифіковані як позитивні. Це помилка 2-го роду (хибно позитивні випадки).

Позитивним випадком називають підтвердження нульвої (основної) гіпотези, а вибір цієї гіпотези залежить від дослідника.

Характеристикою якості класифікації вважаються такі показники.

Доля істинно-позитивних зразків (True Positives Rate): $TRP = \frac{TP}{TP + FN} \cdot 100\%$

Доля істинно-позитивних зразків (False Positives Rate): $FRP = \frac{FP}{TN + FP} \cdot 100\%$

Чутливість (Sensitivity) = Доля істинно-позитивних зразків TRP.

$$Se = TRP = \frac{TP}{TP + FN} \cdot 100\%$$

Специфічність (Specificity) – доля істинно-негативних зразків , що були класифіковані правильно

$$Sp = \frac{TN}{TN + FP} \cdot 100\%$$

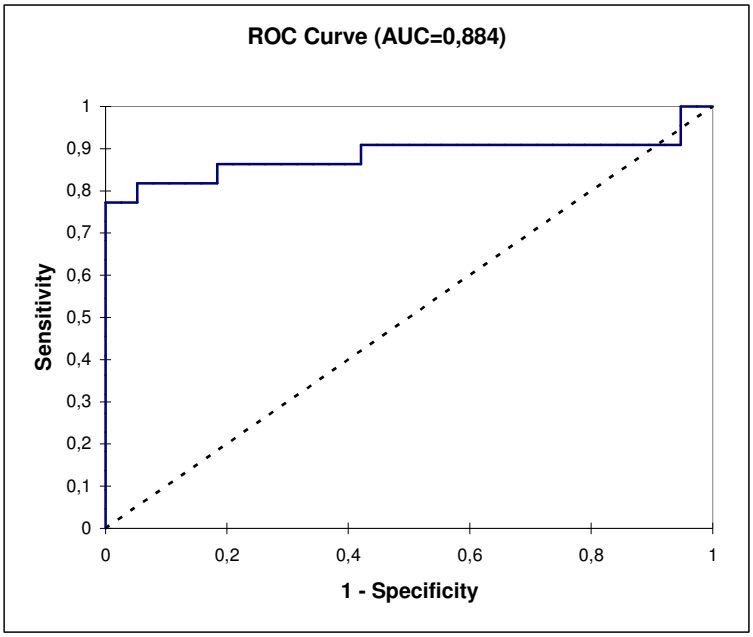
Висока чутливість ще не гарантія високої якості класифікатора: якщо просто оголостити усіх без винятку хворими (гіпердіагностика), можна виявити усіх хворих, тобо чутливість буде 100%, але усі здорові люди будуть оголошеня хворими, тобто специфічність буде дорівнювати 0%.

ROC-крива будується як графік залежності TRP від FRP.

Приклад 8.3. Прогностична значущість показника CA-125.

Розглянемо реальний приклад із медичної практики. Під час клінічних досліджень хворих на рак сечевої системи було виявлено, що генетичний маркер CA-125 можна використовувати як прогностичний фактор для оцінки якості життя хворого після операції. Для цього у кожного пацієнта вимірювали показник CA-125, а потім певний час спостерігали, виникнуть у нього метастази чи ні (1 або 0). Решта показників вважалися незначущими, тобто групи за фншими факторами вважаються статистично однорідними (це окрема складна тема, що називається рандомізацією). Отже, за значенням CA-125 можна визначити ймовірність того, що у лімфатичних вузлах пацієнта виникнуть метастази.

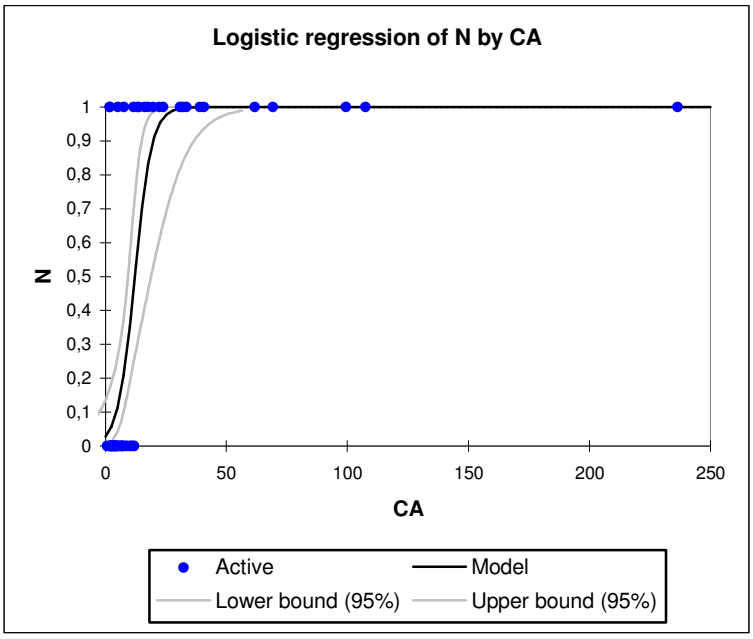
В дослідженні брали участь 60 пацієнтів, тому повну таблицю ми не наводимо. Звернімо увагу на побудовану криву ROC.



Якість класифікації визначається площею фигури під кривою ROC, яка позначається як AUC (area under curve). Чим вище показник AUC, тим вища прогностична цінність моделі. Проте слід пам'ятати, що показник AUC не может характеризувати чутливість і специфічність (тобто він не може їх замінити). Втім, використовується така характеристика якості моделей: від 0,9 до 1,0 — відмінна якість класифікації, від 0,8 до 0,9 — дуже добра, від 0,7 до 0,8 — добра, від 0,6 до 0,7 — середня, від 0,5 до 0,6 - незадовільна). Поріг класифікації (точка усічення cut-off) вибирається довільно. В нашій задачі AUC = 0,884, тобто якість моделі дуже добра.

Теоретично, слід прагнути до моделі, що має абсолютні чутливість і специфічність (тобто стовідсоткові). На практиці це неможливо, тому для пошуку балансу між ними використовується оптимальний поріг класифікації (точка усічення). Ця точка дозволяє розв'язувати задачу класифікації, тобто класифікувати нового пацієнта як такого, що має високу ймовірність виникнення метастазів,чи ні (належить до групи ризику чи ні). Як практичну рекомендацію можна прийняти побажання 1) мати чутливість не менше 80%, а за нею на кривій ROC шукати відповідну специфічність; 2) досягти максимальної сумарної чутливості і специфічності; 3) добиватися майже однакових чутливості і специфічності.

Для класифікації нових пацієнтів можна вибрати ймовірність (cut-off), орієнтуючись на наступний графік.



Поклавши Cut-off = 0,5 (рекомендація за замовченням), маємо, що пороговий CA-125 = 12,171.

Таблиця класифікації за навчальною вибіркою Точка усічення= 0,5

from \ to	0	1	Total	% correct	
0	38	0	38	100,00%	Специфічність
1	5	17	22	77,27%	Чутливість
Total	43	17	60	91,67%	Точність

Посилання

1. http://courses.ncssm.edu/math/Stat_Inst/PDFS/REG3_LOG.pdf
2. http://www.basegroup.ru/library/analysis/regression/logistic/
3. Яковлев П.Г., Сакало В.С., Клюшин Д.А., Мрачковський В.В., Григоренко В.М., Григорук А.В. Фактори прогнозу уротеліального раку сечоводу // Урологія. — 2010. — №2. — С. 22-29.
4. Сакало В.С., Яковлев П.Г., Мрачковський В.В., Клюшин Д.А.. Клінічні та морфологічні фактори прогнозу уротеліального раку ниркової миски // Урологія. — 2010. — №4. — С. 38-45.