Самостійна робота №5
з асимптотичної статистики
Варіант №4

Горбунов Даніел Денисович
4 курс бакалаврату
група "комп'ютерна статистика"

21 травня 2021 р.

1 Вступ.

У даній роботи ми побудували модель для класифікації номеру виноградника, де було вироблено вино, базуючись на його хімічному складі. Додатково наводиться інформація про аналіз якості прогнозування та оцінені числові характеристики параметрів моделі.

2 Хід роботи.

2.1 Початкові дані.

Ми працюємо із даними про хімічний склад та якостей вина, обраних із двох виноградників. Усього спостережень $N = 130$, серед яких 59 сопстережень містять інформацію про вино з першого виноградника, а інші 71 – з другого. Для обраного варіанту пропонується побудувати модель класифікації номеру виноградника за такими характеристиками: вміст спирту (Alcogol), проліну (Proline), кислот (Malic Acid) та золи (Ash). Побудуємо діаграму розсіювання, обчислимо коефіцієнти кореляції.

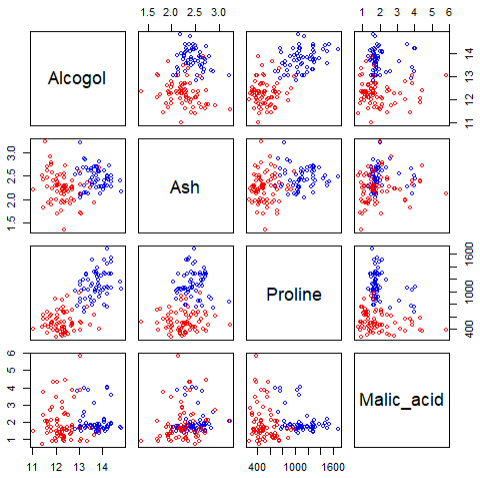

Рис. 1: Діаграма розсіювання початкових даних. Сині точки – дані про вино, що відноситься до першого виноградника, а червоні - до другого.

Серед усіх можливих пар змінних на діаграмі (2.1) ми можемо побачити, що найкраще лінійне розділення простору можливе для пари змінних Alcogol та Proline. В інших випадках бачимо помірне (пара Ash та Proline) або сильне змішування хмарин (наприклад, для пари Ash та Malic Acid). Наводимо матриці кореляцій, обчислені за коефіцієнтами кореляцій Пірсона та Спірмена.

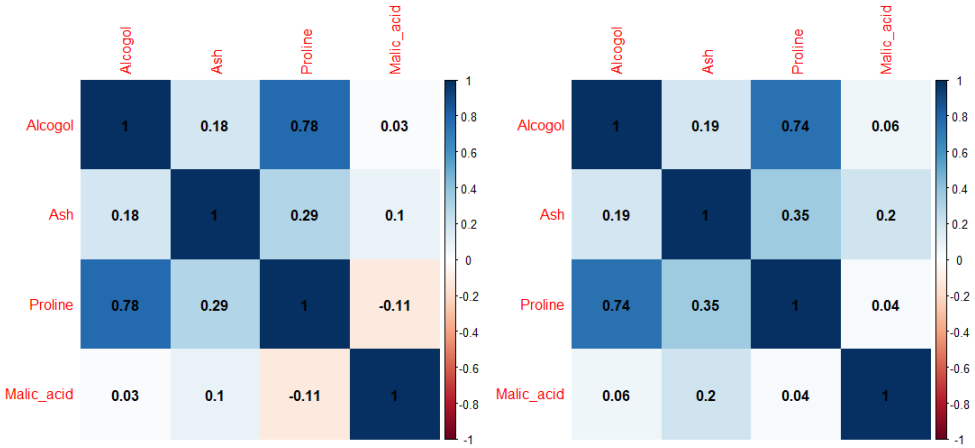

Рис. 2: Кореляційні матриці початкових даних. Зліва – за Пірсоном, справа – за Спірменом.

З результатів попередної діаграми розсіювання можна було зауважити, що для змінних Proline та Alcogol добре виражена лінійна залежність. Трохи слабшою є для пар Ash та Proline, Ash та Alcogol. Висунуте припущення про значущість залежності між Alcogol та Proline підсилюється отриманими значеннями кореляції, як видно з (2.1).

Здається, слушною думкою буде побудувати класифікаційну модель лише за змінними, в якість є добре виражена залежність то розділеність хмарин з різних категорій. Але перш ніж одразу будувати її, ми покажемо як буде поводитися модель, побудована на усіх досліджуваних змінних.

2.2 Логістична регресія та вино.

2.2.1 Теоретичні відомості.

В усіх випадках ми розглядаємо модель логістичної регресії, але з різним комбінуванням змінних. У загальній постановці, ми хочемо дослідити вплив змінних $X^1, \dots, X^p$ на бінарну змінну Y (для простоти вважаємо, що $Y \in \{0, 1\}$). Для цього розглянемо умовну ймовірність того яке значення буде приймати Y при відомих (фіксованих) значеннях $X^1, \dots, X^p$:

$$\mathbb{P}\left(Y_j = 1 \mid X^1, \dots, X^p\right) = \varphi\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_j^p\right),$$

де $\varphi(t)$ - відома функція зв'язку, а параметри $\beta_0, \beta = (\beta_1, \dots, \beta_p)^\top$ потрібно оцінити. У нашому випадку ми розглядаємо логістичну модель регресії, тобто відома функція зв'язку φ задається у вигляді:

$$\varphi(t) = \text{Logist}(t) = \frac{1}{1 + e^{-t}}$$

Для оцінювання β_0, β використовується метод найбільшої вірогідності. Припускаючи сталість значень регресорів $X_j = (X_1^p, \dots, X_n^p)^\top$, ми можемо розглянути функцію вірогідності для випадкових спостережень $Y = (Y_1, \dots, Y_n)^\top$:

$$L(Y, \beta_0, \beta) = \prod_{m=1}^{n} \left(\varphi\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_j^p\right)\right)^{Y_j} \left(1 - \varphi\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_j^p\right)\right)^{1-Y_j}$$

Тоді оцінки максимальної вірогідності β_0, β - це просто

$$(\hat{\beta}_0, \hat{\beta}) = \arg \max_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} L(Y, \beta_0, \beta)$$

2.2.2 Підгонка та прогноз.

Для досліджуваної задачі нам цікаво не лише підігнати модель на запропонованих даних, але й спробувати зробити прогноз для так званих нових, невідомих даних. Для цього ми розіб'ємо сукупні дані $(X_j \mid Y_j)_{j=1}^n$ на дві підмножини – тренувальну та тестову. На тренувальній буде проведена підгонка моделі, а на тестовій – прогноз.

Тренувальна вибірка буде складатися із ста елементів. Ми хочемо мати більш-менш збалансовану вибірку, тому відбір елементів з генеральної вибірки будемо проводити за допомогою стратифікації із випадковим відбором без повернення, де в якості факторної змінної фігурує номер виноградника, до якого належать дані про вино. Із кожної страти будемо обирати по 50 елементів у тренувальну вибірку. В R такий відбір нескладно описати:

```
set.seed(0)
# Зчитування даних, вибыр необхідних змінних
wine.data <- read.csv2(file=path.to.wine)
columns.to.use <- c("Site", "Alcogol", "Ash", "Proline", "Malic_acid")
wine.data <- wine.data[wine.data$Site != 3, columns.to.use]
X <- wine.data[,-1]
Y <- wine.data[,1] - 1
# Стратифікований відбір
N <- nrow(X)
n.0 <- 50
idx <- 1:N
# Відбір 50-ти елементів із першої страти
idx.0 <- sample((1:N)[Y==0], size=n.0)
# Відповідно з другої
idx.1 <- sample((1:N)[Y==1], size=n.0)
idx.train <- c(idx.0, idx.1)
# Формування тренувальної вибірки
# Ті елементи, які не належать їй, будуть належати до тестової
X.train <- X[idx.train,]
Y.train <- Y[idx.train]
```

2.2.3 Перша модель.

У першій моделі ми спробуємо задіяти всі змінні, що досліджуються. Тобто модель матиме вигляд:

$$\begin{aligned}&\mathbb{P}\left(Site_j = 2 \mid Alcogol,\ Ash,\ Proline,\ MalicAcid\right) = \\ &= \text{Logist}\left(\beta_0 + \beta_1 Alcogol_j + \beta_2 Ash_j + \beta_3 Proline_j + \beta_4 (MalicAcid)_j\right)\end{aligned}$$

Підгонку ми будемо робити, використовуючи функцію glm в R.

```
model.full <- glm(Y.train ~ Alcogol + Ash + Proline + Malic_acid, data=X.train,
                  family="binomial")
print(summary(model.full))
```

Підгонка за всіма змінними вийшла невдалою: алгоритм оптимізації не зійшовся, оцінка дисперсії коефіцієнтів моделі досить велика, а тест про значущість коефіцієнтів додатково вказує про неадекватність отриманих результатів.

```
Deviance Residuals:
       Min          1Q      Median          3Q         Max
-4.643e-04  -2.000e-08   0.000e+00   2.000e-08   4.715e-04
Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.367e+04  1.479e+06   0.016    0.987
Alcogol     -1.620e+03  1.014e+05  -0.016    0.987
Ash          4.131e+02  2.577e+04   0.016    0.987
Proline     -3.096e+00  1.902e+02  -0.016    0.987
Malic_acid  -3.829e+02  2.362e+04  -0.016    0.987
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1.3863e+02  on 99  degrees of freedom
Residual deviance: 7.5748e-07  on 95  degrees of freedom
```

Покажемо якість класифікації на тренувальних та тестових даних, вказавши таблицю спряженості між справжніми значеннями відгука та прогнозу, та наводимо значення характеристик Precision та Recall:

$Y_{Train} \setminus \hat{Y}_{Train}$	1	2
1	50	0
2	0	50

$Y_{Test} \setminus \hat{Y}_{Test}$	1	2
1	6	3
2	1	20

$$Precision_{Train} = 1,\ Recall_{Train} = 1;\ Precision_{Test} = 0.6666667,\ Recall_{Test} = 0.8571429$$

Рис. 3: Деякі відомості про якість класифікації за першою моделлю.

З такої хорошої ноти почали на тренувальних даних і так сумно закінчили на тестових. У даному випадку ми спіткнулися з ефектом перенавчання моделі. Давайте спробуємо виправити дану проблему, розглянувши більш розумну комбінацію змінних, запропоновану у розділі про початкові дані.

2.2.4 Друга модель.

З урахуванням попережніх міркувань, поточна модель логістичної регресії запишеться у вигляді:

$$\mathbb{P}\left(Site_j = 2 \mid Alcogol,\ Proline\right) = \text{Logist}\left(\beta_0 + \beta_1 Alcogol_j + \beta_2 Proline_j\right)$$

Результати підгонки стають у певному сенсі кращими, порівнюючи із попередньою моделлю:

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.53927  -0.04111   0.00193   0.03345   2.05641
Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) 67.167797  26.312390   2.553   0.0107
Alcogol     -4.427878   1.898485  -2.332   0.0197
Proline     -0.011813   0.004282  -2.759   0.0058
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 138.629  on 99  degrees of freedom
Residual deviance:  13.804  on 97  degrees of freedom
```

Як видно у звіті, то для стандартного рівня значущості $\alpha = 0.05$ слід прийняти гіпотезу про значущість коефіцієнтів. Отриманий розкид оцінок незадовільний, але набагато менший, ніж було для першої моделі. Результати підгонки на тренувальних даних та прогнозу на тестових показують наступне:

$Y_{Train} \setminus \hat{Y}_{Train}$	1	2
1	48	2
2	2	48

$Y_{Test} \setminus \hat{Y}_{Test}$	1	2
1	9	0
2	1	20

$$Precision_{Train} = 1,\ Recall_{Train} = 0.9;\ Precision_{Test} = 0.6666667,\ Recall_{Test} = 0.8571429$$

Рис. 4: Деякі відомості про якість класифікації за другою моделлю.

З таблиць спряженості ми можемо побачити, що на досліджуваних даних класифікуюча лінія відносить п'ять спостережень до помилкових виноградників. Але зауважимо, що на тестових даних прогнозування спрацювало краще.

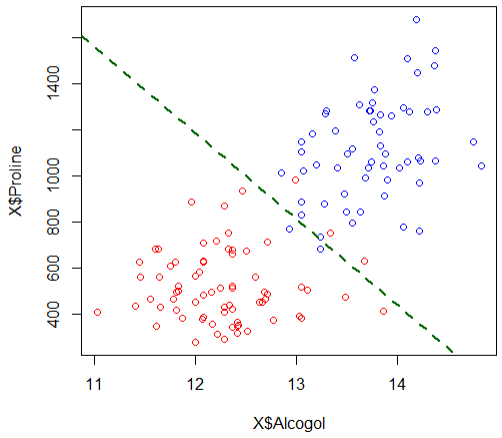

Рис. 5: Діаграма розсіювання даних за змінними Alcogol та Proline. Зелена пунктирна пряма позначає межу класифікації. Візуально можна побачити які саме спостереження віднесли до помилкових категорій.

Для завершення покажемо побудовані довірчі інтервали для коефіцієнтів моделі рівня $1 - \alpha = 1 - 0.05 = 0.95$:

	2.5 %	97.5 %
(Intercept)	30.6330443	144.91853883
Alcogol	-9.8971673	-1.65839090
Proline	-0.0234091	-0.00528301

Рис. 6: Довірчі інтервали для коефіцієнтів другої моделі.

3 Висновки.

Хорошу модель класифікації побудувати вдалося. Остаточна модель має вигляд:

$$\mathbb{P}\left(Site_j = 2 \mid Alcogol,\ Proline\right) = \text{Logist}\left(67.167797 - 4.427878 Alcogol_j - 0.011813 Proline_j\right)$$