Логистическая регрессия

Анастасия Миллер, Дмитрий Корчемкин

23 сентября 2015 г.

Задача

Предсказать значение номинативной переменной (здесь и далее – с двумя градациями). Будем предсказывать вероятность принятия с.в. одного из значений (второе будет появляться автоматически).

Y_i – бинарная предсказываемая переменная, $X = (X_1, \cdots, X_k)$ – предикторы. Модель:

$$\pi_i = P(Y_i = 1 | X_i = x_i) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i^1 + \cdots + \beta_k x_i^k)}$$

Иначе говоря – линейная модель относительно логит-функции от вероятности:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i^1 + \cdots + \beta_k x_i^k$$

Необходимо подобрать $(\beta_0, \cdots, \beta_k)$, максимизирующие функцию правдоподобия. Аналитического решения в общем случае не существует, так что используются итеративные алгоритмы.

Предположения

- Данные для различных индивидов независимы и одинаково распределены
- $Y_i \sim \text{Bin}(n_i, \pi_i)$, X_i обычно предполагаются из экспоненциального семейства
- Ошибки наблюдений должны быть независимы
- Для подбора коэффициентов используется MLE, то есть нужна достаточно большая выборка

Пример

Данные:

- Household Income (`Income` ; rounded to the nearest $1,000.00)
- Gender (`IsFemale` = 1 if the person is female, 0 otherwise)
- Marital Status (`IsMarried` = 1 if married, 0 otherwise)
- College Educated (`HasCollege` = 1 if has one or more years of college education, 0 otherwise)
- Employed in a Profession (`IsProfessional` = 1 if employed in a profession, 0 otherwise)
- Retired (`IsRetired` = 1 if retired, 0 otherwise)
- Not employed (`Unemployed` = 1 if not employed, 0 otherwise)
- Length of Residency in Current City (`ResLength` ; in years)
- Dual Income if Married (`Dual` = 1 if dual income, 0 otherwise)
- Children (`Minors` = 1 if children under 18 are in the household, 0 otherwise)
- Home ownership (`Own` = 1 if own residence, 0 otherwise)
- Resident type (`House` = 1 if residence is a single family house, 0 otherwise)
- Race (`White` = 1 if race is white, 0 otherwise)
- Language (`English` = 1 is the primary language in the household is English, 0 otherwise)

- Previously purchased a parenting magazine (`PrevParent` = 1 if previously purchased a parenting magazine, 0 otherwise).
- Previously purchased a children's magazine (`PrevChild` = 1 if previously purchased a children's magazine)
- Purchased "Kid Creative" (`Buy` = 1 if purchased "Kid Creative," 0 otherwise)

```
data <- read.csv("KidCreative.csv")
for (factor in c('Buy', 'Is.Female', 'Is.Married', 'Has.College', 'Is.Professional
', 'Is.Retired', 'Unemployed', 'Dual.Income', 'Minors', 'Own', 'House', 'White', '
English', 'Prev.Child.Mag', 'Prev.Parent.Mag')) {
  data[, factor] <- factor(data[, factor], levels=c('0','1'), labels=c('0','1'))
}
```

Разделим данные на тренировочную и тестовую выборку:

```
index <- which(1:length(data[,1])%%5 == 0)
train <- data[-index,] # обучающая выборка
test <- data[index,] # тестовая выборка
```

Обучимся по обучающей выборке:

```
model <- glm(Buy ~ ., train, family="binomial")
summary(model)
```

```
##
## Call:
## glm(formula = Buy ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.55992  -0.08580  -0.00961  -0.00101   2.44023
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.685e+01  2.401e+00  -7.016 2.28e-12 ***
## Obs.No.          -2.007e-04  1.130e-03  -0.178  0.85905
## Income            1.912e-04  2.548e-05   7.503 6.23e-14 ***
## Is.Female1        1.587e+00  5.246e-01   3.025  0.00249 **
## Is.Married1       9.082e-01  6.421e-01   1.414  0.15726
## Has.College1     -9.375e-03  4.842e-01  -0.019  0.98455
## Is.Professional1  2.549e-01  5.209e-01   0.489  0.62466
## Is.Retired1      -1.096e+00  1.004e+00  -1.091  0.27515
## Unemployed1      -1.060e+01  1.183e+03  -0.009  0.99285
## Residence.Length  2.289e-02  1.510e-02   1.516  0.12963
## Dual.Income1      3.461e-01  5.812e-01   0.595  0.55155
## Minors1           1.179e+00  5.105e-01   2.309  0.02096 *
## Own1              8.404e-01  6.052e-01   1.388  0.16499
## House1           -1.324e+00  6.837e-01  -1.937  0.05273 .
## White1            2.043e+00  6.414e-01   3.184  0.00145 **
## English1          1.175e+00  8.758e-01   1.341  0.17981
## Prev.Child.Mag1   2.168e+00  8.158e-01   2.657  0.00788 **
## Prev.Parent.Mag1  3.568e-01  6.562e-01   0.544  0.58660
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 525.86  on 538  degrees of freedom
## Residual deviance: 153.30  on 521  degrees of freedom
## AIC: 189.3
##
## Number of Fisher Scoring iterations: 17
```

Много незначимых признаков, выберем только значимые. Выбор происходит по тем же принципам (более того, вызывается точно та же функция), что и с линейной моделью:

```
model.significant <- step(model)
```

```
summary(model.significant)
```

```
##
## Call:
## glm(formula = Buy ~ Income + Is.Female + Is.Married + Is.Retired +
##     Residence.Length + Minors + Own + House + White + English +
##     Prev.Child.Mag, family = "binomial", data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.69799  -0.09362  -0.01129  -0.00184   2.21984
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.665e+01  2.254e+00  -7.386 1.51e-13 ***
## Income            1.915e-04  2.468e-05   7.757 8.69e-15 ***
## Is.Female1        1.516e+00  5.024e-01   3.017  0.00255 **
## Is.Married1       1.097e+00  5.209e-01   2.106  0.03522 *
## Is.Retired1      -1.565e+00  8.757e-01  -1.787  0.07397 .
## Residence.Length  2.200e-02  1.485e-02   1.481  0.13853
## Minors1           1.089e+00  4.898e-01   2.224  0.02617 *
## Own1              9.386e-01  5.836e-01   1.608  0.10779
## House1           -1.370e+00  6.687e-01  -2.049  0.04047 *
## White1            1.949e+00  6.181e-01   3.153  0.00162 **
## English1          1.285e+00  8.312e-01   1.546  0.12221
## Prev.Child.Mag1   2.191e+00  8.092e-01   2.708  0.00677 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 525.86  on 538  degrees of freedom
## Residual deviance: 154.45  on 527  degrees of freedom
## AIC: 178.45
##
## Number of Fisher Scoring iterations: 8
```

Проверим качество предсказания на тестовой выборке:

```
result <- predict(model, newdata=test, type='response') # вектор вероятностей
length(which(as.numeric(result > 0.4) == test$Buy)) / length(test$Buy)
```

```
## [1] 0.9402985
```

Параметр `type` показывает, в какой шкале будут построены предсказываемые значения: в шкале линейных предикторов (`"link"`), что в нашем случае означает вероятности в логистической шкале, или в шкале ответов (`"response"`), что для нас означает как раз искомые вероятности.