Labbrapport i Statistik

Laboration 1 - Komplettera

732G54

Zerui Wang
Viet Tien Trinh

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2024-11-28

Table of contents

1 Introduktion **1**

2 Labborationsuppgifter **2**

2.1 a) Dela upp hela datamaterialet i en tränings- och valideringsmängd med 2/3 av observationerna i träningsmängden. Presentera vilket seed som använts för uppdelningen. Ta sedan fram en frekvenstabell för responsvariabeln ur båda datamängderna och kommentera på resul-tatet; hur är balansen mellan klasserna, hur stor andel finns i varje klass i de två mängderna? 2

2.2 b) Anpassa följande logistiska regressionsmodeller 4

2.2.1 i.alla förklarande variabler som mäter frekvensen av antalet ord (variabler döpta Word i data) 4

2.2.2 ii.alla förklarande variabler som mäter frekvensen av antalet symboler (variabler döpta Char i data) 6

2.2.3 iii.alla förklarande variabler i data 7

2.3 c) Tolka parameterskattningen för den första variabeln i modell i. och ii. 10

2.4 d) Utgå från modell iii. och testa med ett test ifall modellen bör reduceras till modell i.. 11

2.5 e) Ta fram lämpliga utvärderingsmått på tränings- och valideringsmängden för de anpassade modellerna ur b). Presentera dessa i lämpliga tabeller för att underlätta jämförelsen. Kommentera och analysera era resultat. Vilken modell anser ni vara bäst för att modellera responsvariabeln och hur generaliserbar är modellerna? 12

3 Lärdomar, problem, övriga kommentarer **16**

1 Introduktion

Detta är ett arbete som handlar om logiska regressionen, vilket innehåller totalt fem uppgifter som geomfördes av oss.

2 Labborationsuppgifter

2.1 a) Dela upp hela datamaterialet i en tränings- och valideringsmängd med 2/3 av observationerna i träningsmängden. Presentera vilket seed som använts för uppdelningen. Ta sedan fram en frekvenstabell för responsvariabeln ur båda datamängderna och kommentera på resul-tatet; hur är balansen mellan klasserna, hur stor andel finns i varje klass i de två mängderna?

```
# Antalet observationer totalt
# Antalet observationer totalt
n <- nrow(data)

# Antalet som tilldelas till träningsmängden utifrån en andel på 2/3
nTrain <- n*(2/3)

# Sätter ett seed för reproducerbarhet
set.seed(355)

# Index (utvalda observationer) till träningsmängden
indexTrain <- sample(x = n, size = nTrain, replace = FALSE)

# Plockar ut utvalda observationer från materialet med positiv
# indexering (lägger till) för träning och negativ indexering
# (tar bort) för validering
dataTrain <- data[indexTrain,]
dataValid <- data[-indexTrain,]
```

Uppdelningen har gjorts med ett seed på 355. Mängden av både data är 333 observationer för träningsmängd respektive 167 observation för valideringsmängd.

```
### Frekvenstabell för spam på dataTrain
dataTrain_frek <- table(dataTrain$Spam) %>%
  kable(col.names = c("Spam", "Freq"), caption = "Frekvenstabell för spam på dataTrain")
#### 0 = 0.5796
#### 1 = 0.4204
dataTrain_frek
```

Tabell 1: Frekvenstabell för spam på dataTrain

Spam	Freq
0	193
1	140

Träningsmängd innehåller 42.04% observation tillhör klass 1 och 57.96% observation tillhör klass 0.

```
### Frekvenstabell för spam på dataValid
dataValid_frek <- table(dataValid$Spam) %>%
  kable(col.names = c("Spam", "Freq"), caption = "Frekvenstabell för spam på dataValid")
#### 0 = 0.5808
#### 1 = 0.4192
dataValid_frek
```

Tabell 2: Frekvenstabell för spam på dataValid

Spam	Freq
0	97
1	70

Valideringsmängd innehåller 41.92% observation tillhör klass 1 och 58.08% observation tillhör klass 0. Klasserna är relativt balanserade i både mängderna med en något högre andel av klass 0 på omkring 58% jämfört med andel av klass 1 som utgör omkring 42% i båda.

2.2 b) Anpassa följande logistiska regressionsmodeller

2.2.1 i.alla förklarande variabler som mäter frekvensen av antalet ord (variabler döpta Word i data)

$$\text{logit}(P(Y=1)) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Word}_1 + \hat{\beta}_2 \cdot \text{Word}_2 + ... + \hat{\beta}_{48} \cdot \text{Word}_{48}$$

Där:
P är sannolikhet för spam ord.
$\hat{\beta}_0$ är interceptet.
$\hat{\beta}_1 - \hat{\beta}_{48}$ är koefficienter för respektive förklarande variabler (word1 - word48).

```
modell1 <- glm(formula = Spam ~.-Char1-Char2-Char3-Char4-Char5-Char6-Capitalrun1-
                 Capitalrun2-Capitalrun3,data = dataTrain,family = "binomial",
               control = glm.control(maxit = 150))
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(modell1) %>%
  coef() %>%
  kable(digits = 3, caption = "Modellens skattade koefficienter.",
        col.names=c("Variabel", "Skattning", "Medelfel", "z-värde", "p-värde"))
```

Tabell 3: Modellens skattade koefficienter.

Variabel	Skattning	Medelfel	z-värde	p-värde
(Intercept)	-3.263888e+12	8727436	-373980.1	0
Word1	-1.306950e+14	11601543	-11265314.9	0
Word2	-4.582731e+14	2620442	-174883921.7	0
Word3	2.783295e+14	9113105	30541678.9	0
Word4	2.122719e+14	1662367	127692528.6	0
Word5	-5.517909e+13	6064200	-9099154.3	0
Word6	5.412937e+14	17719066	30548657.1	0
Word7	2.566995e+14	13484150	19037126.1	0
Word8	1.172784e+14	9750781	12027587.1	0
Word9	7.135481e+14	15021156	47502876.3	0
Word10	3.126474e+14	7105490	44000822.6	0
Word11	-5.513911e+14	25742196	-21419738.9	0
Word12	-1.255229e+13	6061708	-2070751.9	0
Word13	-5.213070e+14	14981941	-34795690.6	0
Word14	3.539259e+13	20454185	1730334.8	0

Variabel	Skattning	Medelfel	z-värde	p-värde
Word15	7.096462e+14	24824499	28586527.6	0
Word16	3.452158e+14	4958541	69620453.6	0
Word17	6.343007e+14	8074325	78557730.7	0
Word18	1.171746e+14	8036650	14580033.8	0
Word19	-3.431940e+13	2622363	-13087203.3	0
Word20	2.096455e+15	7426644	282288335.0	0
Word21	7.462763e+13	4287259	17406839.8	0
Word22	1.722137e+13	3331105	5169866.8	0
Word23	9.078761e+14	9483686	95730299.7	0
Word24	-1.800910e+14	6053344	-29750669.8	0
Word25	-1.054518e+15	2798575	-376805425.1	0
Word26	-1.079959e+15	9066013	-119121773.1	0
Word27	-5.420988e+14	1512472	-358419060.4	0
Word28	2.252042e+14	11990797	18781424.7	0
Word29	-2.304244e+14	13644518	-16887693.4	0
Word30	5.098583e+14	11932348	42729081.5	0
Word31	2.558328e+15	36752100	69610398.0	0
Word32	9.461806e+14	43616546	21693157.0	0
Word33	-5.320227e+13	11427023	-4655829.0	0
Word35	-1.525021e+15	16831521	-90605042.7	0
Word36	3.501576e+14	22205943	15768643.3	0
Word37	-5.513641e+13	10557748	-5222364.1	0
Word38	2.111937e+14	40563883	5206446.6	0
Word39	-6.016119e+14	12061802	-49877446.1	0
Word40	-3.109134e+14	26400721	-11776699.8	0
Word41	-3.733696e+15	26022261	-143480849.7	0
Word42	-1.157578e+15	11840786	-97761894.9	0
Word43	-1.866564e+15	13902757	-134258587.6	0
Word44	-3.898527e+14	15259195	-25548705.1	0
Word45	-3.881767e+14	6224011	-62367620.0	0
Word46	-3.323186e+14	10328967	-32173456.2	0
Word47	-5.843689e+15	64813963	-90160966.1	0
Word48	-1.026707e+15	31417271	-32679705.9	0

```
AIC(modell1)
```

```
[1] 2402.794
```

Modellen har många stora parameterskattningar men alla verkar signifikanta. Något orimliga resultat på grund av modellens onödiga komplexitet eller överanpassning. Modellens AIC är 2402,794. Detta höga värde är ytterligare en indikation att modellen är överanpassad.

2.2.2 ii.alla förklarande variabler som mäter frekvensen av antalet symboler (variabler döpta Char i data)

$$\text{logit}(P(Y=1)) = \hat{\beta}_0 + \hat{\beta}_1 \cdot Char_1 + \hat{\beta}_2 \cdot Char_2 + \hat{\beta}_3 \cdot Char_3 + \hat{\beta}_4 \cdot Char_4 + \hat{\beta}_5 \cdot Char_5 + \hat{\beta}_6 \cdot Char_6$$

Där:
P är sannolikhet för spam ord.
$\hat{\beta}_0$ är interceptet.
$\hat{\beta}_1 - \hat{\beta}_6$ är koefficienter för respektive förklarande variabler (Char1 - Char6).

```
modell2 <- glm(formula = Spam ~ Char1 + Char2 + Char3 + Char4 + Char5 + Char6,
               data = dataTrain, family = "binomial")
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(modell2) %>%
  coef() %>%
  kable(digits = 3, caption = "Modellens skattade koefficienter.",
        col.names=c("Variabel", "Skattning", "Medelfel", "z-värde", "p-värde"))
```

Tabell 4: Modellens skattade koefficienter.

Variabel	Skattning	Medelfel	z-värde	p-värde
(Intercept)	-1.120	0.214	-5.231	0.000
Char1	-0.164	0.508	-0.322	0.747
Char2	-3.681	1.078	-3.414	0.001
Char3	1.386	2.871	0.483	0.629
Char4	3.503	0.607	5.770	0.000
Char5	7.902	1.603	4.929	0.000
Char6	2.000	1.173	1.705	0.088

```
AIC(modell2)
```

```
[1] 309.9088
```

Modellen innehåller några icke-signifikanta parametrar. Modellens AIC är 309,9088.

2.2.3 iii.alla förklarande variabler i data

$$\text{logit}(P(\text{Spam} = 1)) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Word}_1 + \hat{\beta}_2 \cdot \text{Word}_2 + ... + \hat{\beta}_{48} \cdot \text{Word}_{48} + \hat{\beta}_{49} \cdot \text{Char}_1 + ... + \hat{\beta}_{54} \cdot \text{Char}_6 + \hat{\beta}_{55} \cdot \text{Capitalrun}_1 + \hat{\beta}_{56} \cdot \text{Capitalrun}_2 + \hat{\beta}_{57} \cdot \text{Capitalrun}_3$$

Där:
P är sannolikhet för spam ord.
$\hat{\beta}_0$ är interceptet.
$\hat{\beta}_1 - \hat{\beta}_{57}$ är koefficienter för alla förklarande variabler som innehålls i data.

```
modell_full <- glm(formula = Spam ~ ., data = dataTrain, family = "binomial")
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(modell_full) %>%
  coef() %>%
  kable(digits = 3, caption = "Modellens skattade koefficienter.",
        col.names=c("Variabel", "Skattning", "Medelfel", "z-värde", "p-värde"))
```

Tabell 5: Modellens skattade koefficienter.

Variabel	Skattning	Medelfel	z-värde	p-värde
(Intercept)	-1.368496e+14	9649674.73	-14181787.3	0
Word1	-4.413109e+13	11681341.75	-3777913.1	0
Word2	-2.658775e+14	2645510.11	-100501429.6	0
Word3	2.511774e+14	9628765.02	26086144.8	0
Word4	1.334430e+14	1668562.79	79974815.8	0
Word5	-2.474257e+14	6129006.02	-40369630.3	0
Word6	5.576330e+14	17844406.76	31249733.8	0
Word7	2.519924e+14	13578098.98	18558737.1	0
Word8	2.552874e+14	9870992.52	25862380.3	0
Word9	3.767980e+14	15793936.20	23857128.2	0
Word10	1.071383e+14	7239503.66	14799119.3	0
Word11	-1.274574e+15	26110780.33	-48814072.2	0
Word12	-2.183655e+13	6140573.70	-3556109.2	0
Word13	-6.160612e+13	15315674.07	-4022423.2	0
Word14	2.657600e+14	24401070.73	10891323.4	0
Word15	5.593407e+14	28896701.37	19356557.7	0
Word16	2.561489e+14	4998099.14	51249266.0	0
Word17	6.596773e+14	8176665.63	80678029.6	0
Word18	-1.215196e+12	8304818.14	-146324.2	0

Variabel	Skattning	Medelfel	z-värde	p-värde
Word19	-4.097714e+12	2703164.39	-1515895.2	0
Word20	1.021557e+15	7667960.16	133224132.8	0
Word21	-8.961243e+13	4328856.76	-20701177.0	0
Word22	1.356225e+14	3869194.33	35051861.8	0
Word23	8.079454e+14	9906650.34	81555861.2	0
Word24	4.242097e+13	6067518.49	6991485.4	0
Word25	-5.160202e+14	2833404.50	-182120203.8	0
Word26	-1.063231e+14	9284616.69	-11451529.8	0
Word27	-2.651965e+14	1524161.64	-173995022.0	0
Word28	-2.320304e+14	12340382.65	-18802528.8	0
Word29	-4.110542e+14	13762449.56	-29867808.6	0
Word30	3.334060e+14	12100679.45	27552668.9	0
Word31	-5.264105e+14	36950737.87	-14246278.5	0
Word32	6.877663e+14	43922752.26	15658542.4	0
Word33	-7.406622e+13	11669472.36	-6347006.6	0
Word35	-2.918369e+14	16906015.16	-17262313.6	0
Word36	8.677129e+14	22347082.70	38828912.1	0
Word37	-2.891109e+13	10827315.98	-2670199.4	0
Word38	1.405651e+14	40801724.97	3445077.3	0
Word39	-3.071064e+14	12103780.04	-25372766.8	0
Word40	3.292079e+14	26553931.98	12397709.8	0
Word41	-8.893111e+14	26377070.95	-33715308.3	0
Word42	-6.702789e+14	11864177.78	-56496028.4	0
Word43	-5.594158e+14	14399664.47	-38849224.0	0
Word44	-3.065073e+14	15354155.49	-19962495.5	0
Word45	-3.644286e+14	6285751.99	-57976926.0	0
Word46	-2.430924e+14	10408970.68	-23354122.3	0
Word47	-4.594021e+15	64986428.94	-70691998.1	0
Word48	-3.091213e+15	31848192.50	-97060858.7	0
Char1	-4.124496e+14	16556740.21	-24911282.6	0
Char2	8.984323e+12	23115710.00	388667.4	0
Char3	-1.205184e+15	84996494.78	-14179221.7	0
Char4	4.876520e+14	6915659.02	70514178.4	0
Char5	1.102862e+15	15739390.91	70070200.9	0
Char6	-5.066937e+12	16052413.92	-315649.5	0
Capitalrun1	-1.402591e+12	159063.25	-8817817.9	0
Capitalrun2	8.646508e+11	59064.94	14638984.8	0
Capitalrun3	-1.926510e+11	13255.71	-14533438.1	0

```
AIC(modell_full)
```

```
[1] 2492.881
```

Modellen har många stora parameterskattningar men alla verkar signifikanta. Något orimliga resultat på grund av modellens onödiga komplexitet eller överanpassning. Modellens AIC är 2492,881. Detta höga värde är ytterligare en indikation att modellen är överanpassad.

2.3 c) Tolka parameterskattningen för den första variabeln i modell i. och ii.

För modellen 1: parameterskattningen för första variabeln är $-2,488 \cdot 10^{14}$ som tabellen ovan visar och det kan transformeras som $e^{-2,488191 \cdot 10^{14}} = 0$,vilket innebär att när frekvensen av Word1 ökar med en enhet, minskar oddset för att mejlet klassas som Spam (Y = 1) nästan till noll, vilket i sin tur betyder också att sannolikheten att mejlet klassas som spam minskar till nästan noll.

För modellen 2: parameterskattningen för första variabeln är -0.164 som tabellen ovan visar och det kan transformeras som $e^{-0.164}$, blir det 0.85. Då man kan säga att när Char1 öka med en enhet, så är oddset att mejlet är spam mindre 15%.

2.4 d) Utgå från modell iii. och testa med ett test ifall modellen bör reduceras till modell i..

För att testa att modell iii (innehåll alla förklarande variabler) som kan reducera till modell i (innehåller alla förklarande variabler som mäter frekvensen av antalet ord) likelihoodkvottest beräknas:

$$H_0 : \beta_{49} = \beta_{50} = \beta_{51} = \beta_{52} = \beta_{53} = \beta_{54} = \beta_{55} = \beta_{56} = \beta_{57} = 0$$
$$H_a : \text{Minst en av } \beta_{49}, \beta_{50}, \beta_{51}, \beta_{52}, \beta_{53}, \beta_{54}, \beta_{55}, \beta_{56}, \beta_{57} \neq 0$$

Test formeln

$$LR_{test} = -2ln(\frac{L_R}{L_F}) = -2lnL_R - (-2lnL_F)$$

Med hjälper från R

```
LR <- (-2*logLik(modell1) - (-2*logLik(modell_full))) %>% c()
LR
```

```
[1] -72.08731
```

```
pValue <- pchisq(q = LR, df = 9, lower.tail = FALSE)
pValue
```

```
[1] 1
```

Med testet i träningsmängden fick vi att LR-värden är ganska stora och P-värden är mindre än signifikansnivån på 5%, vilket innebär att vi kan förkasta H_0. Detta innebär att den fullständiga modellen (som innehåller alla förklarande variabler) är bättre än modell i (som innehåller alla förklarande variabler som mäter antal ord) i träningsmängden.

2.5 e) Ta fram lämpliga utvärderingsmått på tränings- och valideringsmängden för de anpassade modellerna ur b). Presentera dessa i lämpliga tabeller för att underlätta jämförelsen. Kommentera och analysera era resultat. Vilken modell anser ni vara bäst för att modellera responsvariabeln och hur generaliserbar är modellerna?

```
evalTrainModel1 <-
  classEvaluation(newData = dataTrain, model = modell1, trueY = dataTrain$Spam)

evalValidModel1 <-
  classEvaluation(newData = dataValid, model = modell1, trueY = dataValid$Spam)

evalTrainModel2 <-
  classEvaluation(newData = dataTrain, model = modell2, trueY = dataTrain$Spam)

evalValidModel2 <-
  classEvaluation(newData = dataValid, model = modell2, trueY = dataValid$Spam)

evalTrainModel3 <-
  classEvaluation(newData = dataTrain, model = modell_full, trueY = dataTrain$Spam)

evalValidModel3 <-
  classEvaluation(newData = dataValid, model = modell_full, trueY = dataValid$Spam)

tibble(
  Modell = c("Modell i", "Modell ii", "Modell iii"),
  `Träffsäkerhet Träning` = c(evalTrainModel1$overall[1], evalTrainModel2$overall[1],
                              evalTrainModel3$overall[1]),
  `Felkvot Träning` = c(evalTrainModel1$overall[2], evalTrainModel2$overall[2],
                        evalTrainModel3$overall[2]),
  AIC = c(AIC(modell1), AIC(modell2), AIC(modell_full)),
  `Träffsäkerhet Validering` = c(evalValidModel1$overall[1], evalValidModel2$overall[1],
                                 evalValidModel3$overall[1]),
  `Felkvot Validering` = c(evalValidModel1$overall[2], evalValidModel2$overall[2],
                           evalValidModel3$overall[2])
) %>%
  kable(digits = 3, caption = "Enkla utvärderingsmått för respektive modell.")
```

Tabell 6: Enkla utvärderingsmått för respektive modell.

Modell	Träffsäkerhet Träning	Felkvot Träning	AIC	Träffsäkerhet Validering	Felkvot Validering
Modell i	0.904	0.096	2402.794	0.820	0.180
Modell ii	0.823	0.177	309.909	0.760	0.240
Modell iii	0.901	0.099	2492.881	0.826	0.174

Även om modell iii med högst träffsäkerhet värden i valideringsmängden villket är 0.918 men det är också modellen med högst AIC värden villket är 2492,881 och det betyder att modellen är överanpassad och kan inte vara den lämpligast modellen. Å andra sidan Modell ii med lägst AIC värden villket är 309.909, så vi kan anser att modellen ii är den som bäst på att generalisera data.

```
cbind(
  evalTrainModel1$classWise, evalValidModel1$classWise
) %>%
  kable(digits = 3, caption = "Klassvisa utvärderingsmått för träning och
        valideringsmängden i modell 1") %>%
  add_header_above(
    c(
      " " = 1,
      "Träning" = 2,
      "Validering" = 2)
  )
```

Tabell 7: Klassvisa utvärderingsmått för träning och valideringsmängden i modell 1

	Träning		Validering	
	0	1	0	1
sensitivitet	0.948	0.843	0.814	0.829
specificitet	0.843	0.948	0.829	0.814

```
cbind(
  evalTrainModel2$classWise, evalValidModel2$classWise
) %>%
  kable(digits = 3, caption = "Klassvisa utvärderingsmått för träning och
        valideringsmängden i modell 2") %>%
  add_header_above(
    c(
      " " = 1,
```

```
      "Träning" = 2,
      "Validering" = 2)
  )
```

Tabell 8: Klassvisa utvärderingsmått för träning och valideringsmängden i modell 2

	Träning		Validering	
	0	1	0	1
sensitivitet	0.907	0.707	0.907	0.557
specificitet	0.707	0.907	0.557	0.907

```
cbind(
  evalTrainModel3$classWise, evalValidModel3$classWise
) %>%
  kable(digits = 3, caption = "Klassvisa utvärderingsmått för träning och
        valideringsmängden i modell 3") %>%
  add_header_above(
    c(
      " " = 1,
      "Träning" = 2,
      "Validering" = 2)
  )
```

Tabell 9: Klassvisa utvärderingsmått för träning och valideringsmängden i modell 3

	Träning		Validering	
	0	1	0	1
sensitivitet	0.948	0.836	0.845	0.800
specificitet	0.836	0.948	0.800	0.845

Enligt uppgiften d bedöms modell iii vara bättre än modell i, men detta gäller endast för träningsmängden, vilket syns tydligt genom träffsäkerheten på 90,7% jämfört med 82,3%. Det betyder dock inte att modell iii är bäst för att modellera responsvariabeln. De klasspecifika måtten ligger jämnt mellan 80–90% för både träningsmängden och valideringsmängden för både modell i och iii, vilket antyder att båda modellerna är lika bra på att prediktera klass 1 som klass 0. För modell iii är skillnaden i träffsäkerhet mellan träningsmängden och valideringsmängden relativt stor (90,7% jämfört med 82,6%). Detta är en större än skillnaden i träffsäkerhet mellan träningsmängden och valideringsmängden för modell i, vilket indikerar att modell i kan ge mer stabila värden även med olika data. För modell ii är träffsäkerheten för träningsmängden relativt hög, nämligen 82,3%, men

den sjunker till 76% för valideringsmängden, vilket indikerar att modellen överanpassar. Dessutom är modell ii inte lika bra på att prediktera klass 1 som klass 0; för träningsmängden är sensitiviteten 0,707% jämfört med 0,907%, och för valideringsmängden är den 0,557% jämfört med 0,907%. Sammanfattningsvis kan vi dra slutsatsen att modell I är bäst.

3 Lärdomar, problem, övriga kommentarer

Efter vi utförde arbetet har vi fått grundläggande koll på hur logiska regression fungerar/tillämpas i samband med träning-valideringsmängder för modells anpassning och utvärdering.