Budowanie modeli - zajęcia

Piotr J. Sobczyk

1 grudnia 2016

> Dobry chrześcijanin powinien wystrzegać się matematyków i tych wszystkich, którzy tworzą puste proroctwa. Istnieje niebezpieczeństwo, że matematycy zawarli przymierze z diabłem, aby zgubić duszę człowieka i wtrącić go w odmęty piekieł. – św. Augustyn

Co prawda św. Augustyn nie miał raczej na myśli matematyków w dzisiejszym rozumieniu, a raczej astrologów, ale weźmiemy sobie do serca ostrzeżenie o „pustych proroctwach".

Zauważmy, że do tej pory nasze postępowanie było następujące:

1. Budujemy model w oparciu o wszystkie dostępne dane
2. Sprawdzamy czy model został zbudowany poprawnie (diagnostyka)
3. Ewentualnie naprawiamy model

Poznaliśmy też jedną miarę dobroci dopasowania - R^2. Zobaczyliśmy, że nie jest ona idealna, im więcej zmiennych „wrzucimy" do modelu, tym wyższe będzie R^2. W zagadnieniu analizy wariancji poznaliśmy metodę porównywania modeli zagnieżdżonych.

Jaki jest problem z R^2 lub jakąkolwiek inną miarą? Oceniamy dopasowanie danych do modelu, który został zbudowany w oparciu o nie. Zależy nam na tym, żeby model był uniwersalny, i dawał dobrą predykcję na nowych danych. Każdy model ma tendencję do „przeuczania", to znaczy lepiej wypada na danych, które „widział" niż na nowych. To trochę jak kolokwium z zadań z list, coś mierzy, ale nie jest dobrą miarą wiedzy jaką opanowali studenci. Podsumowując, aby ocenić przydatność modelu musimy go ocenić na innym, nowym zbiorze danych.

Generalna zasada, za książką Hadelya Wickhama, jest następująca:

> Each observation can either be used for exploration or confirmation, not both.

Eksploracja, czyli przyglądanie się danym, budowanie modeli, transformacje zmiennych itd. Potwierdzenie (confirmation) to jedynie ocena jakości końcowego modelu. W momencie kiedy zostanie to zbadane, nasze szukanie i budowa modelu MUSZĄ się zakończyć. W przeciwnym razie obserwacja stała by się „esploracyjna".

Podział danych na 3 grupy

Poniżej zamieszczam ogólnie przyjętą praktykę. Nie jest ona jedyna możliwa, ale jest sensowna i warto jej się trzymać póki nie zdobędzie się więcej doświadczenia. Pochodzi z książki R for Data Science.

Przed rozpoczęciem analizy dzielimy zbiór danych na trzy części:

- 60% danych stanowi zbiór treningowy - czyli taki, na którym możemy robić wszystko. Od wizualizacji do budowy dowolnej liczby modeli
- 20% danych stanowi zbiór walidacyjny. Na tej części danych możemy porównywać między sobą modele, ale nie tworzymy nowych modeli w oparciu o te dane. Na podstawie porównania modeli na zbiorze walidacyjnym wybieramy ostateczny model
- 20% danych stanowi zbiór testowy. Można go użyć tylko raz, w celu sprawdzenia dokładności wybranego, ostatecznego modelu

Przykład - ceny diamentów

```
?diamonds
data(diamonds)
str(diamonds)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    53940 obs. of  10 variables:
##  $ carat  : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
##  $ depth  : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num  55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int  326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
##  $ y      : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z      : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

```
summary(diamonds)
```

```
##      carat               cut        color        clarity
##  Min.   :0.2000   Fair      : 1610   D: 6775   SI1    :13065
##  1st Qu.:0.4000   Good      : 4906   E: 9797   VS2    :12258
##  Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194
##  Mean   :0.7979   Premium   :13791   G:11292   VS1    : 8171
##  3rd Qu.:1.0400   Ideal     :21551   H: 8304   VVS2   : 5066
##  Max.   :5.0100                     I: 5422   VVS1   : 3655
##                                     J: 2808   (Other): 2531
##      depth           table           price             x
##  Min.   :43.00   Min.   :43.00   Min.   :  326   Min.   : 0.000
##  1st Qu.:61.00   1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710
##  Median :61.80   Median :57.00   Median : 2401   Median : 5.700
##  Mean   :61.75   Mean   :57.46   Mean   : 3933   Mean   : 5.731
##  3rd Qu.:62.50   3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540
##  Max.   :79.00   Max.   :95.00   Max.   :18823   Max.   :10.740
##
##        y               z
##  Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 4.720   1st Qu.: 2.910
##  Median : 5.710   Median : 3.530
##  Mean   : 5.735   Mean   : 3.539
##  3rd Qu.: 6.540   3rd Qu.: 4.040
##  Max.   :58.900   Max.   :31.800
##
```

```
set.seed(23) #MEGA WAŻNE!!!!!
train=sample(1:nrow(diamonds), floor(0.6 * nrow(diamonds)))
diamonds_train=diamonds[train,]
diamonds_rest=diamonds[-train,]
query=sample(1:nrow(diamonds_rest), floor(0.5 * nrow(diamonds_rest)))
diamonds_query=diamonds_rest[query,]
diamonds_test=diamonds_rest[-query,]
```

Zadanie

Zbuduj kilka modeli przewidywania ceny diamentów. Dokonaj diagnostyki i naprawy modelu regresji. Wybierz najlepszy model i sprawdź jego dokładność.

Referencje:

http://r4ds.had.co.nz/model-intro.html