Proper approach for modeling multi-class probabilities (proportions/compositional data)

Asked 1 year, 2 months ago Modified 1 year, 1 month ago Viewed 297 times

2

I have a dataset in which the target values are counts of objects within each class, example (Xs are input features, Ys are values to be predicted):

```
example_id| X1, X2, X3, X4 | Y1,   Y2,   Y3,   Y4
----------+----------------+-----------------------
         1|  1,  2,  3,  4 |  1,    0,    5,   8
         2|  5,  6,  7,  8 |  3,   10,    0,   0
         3|  8,  9,  0,  1 |  1,    1,    1,   1
```

What I'm actually interested in predicting are the proportions of each classes within the sample, that is:

```
example_id| X1, X2, X3, X4 |   Y1,     Y2,     Y3,     Y4
----------+----------------+------------------------------
         1|  1,  2,  3,  4 | 1/14,   0/14,    5/14,   8/14
         2|  5,  6,  7,  8 | 3/13,  10/13,    0/13,   0/13
         3|  8,  9,  0,  1 |  1/4,    1/4,     1/4,    1/4
```

My initial idea was to transform the original data into the proportions (so model directly what I'm interested in) and then train the GBM-based model on it with a softmax-based objective that would allow passing in multi-dimensional and continuous input.

What I've found out is that neither of the common GBM libraries actually support that kind of setup, meaning it's either not very common or just inappropriate.

Clearly there are several ways I can try solving that problem, including:

- using neural nets instead of GBMs to train using the objective I initially wanted to use
- modeling the counts directly and then getting the proportions in the post-processing step
- use MultiCrossEntropy objective and then transform the scores so they sum up to 1, which seems quite dumb and very inappropriate

I can obviously try all the approaches and see which one works best, but I was wondering whether there's any recommended/theory-driven approach to this kind of problem that I failed to find?

regression machine-learning classification multiple-regression compositional-data

Share Cite Improve this question Follow

edited Nov 30, 2022 at 11:35

asked Nov 29, 2022 at 16:02

Matek
941 1 8 16

Are you looking for Dirichlet Regression? – dimitriy Nov 30, 2022 at 6:38

@dimitriy I don't think I am, assuming I understand its definition correctly. The Y values are not directly related as a sum of the X values. I.e. that part "It is practically a case where there are multiple dependent 'Y' variables and one predictor X variable, whose sum is distributed among the Ys" doesn't apply to my case. – Matek Nov 30, 2022 at 10:09

My understanding is that you do not need Ys to be directly related to Xs and you could have multiple Xs as predictors. Where is that quote from? – dimitriy Nov 30, 2022 at 10:15

The other common approach is a fractional multinomial logit. – dimitriy Nov 30, 2022 at 10:42

@dimitriy The quote is from r-statistics.co/Dirichlet-Regression-With-R.html (sorry, should have posted that before). – Matek Nov 30, 2022 at 10:59

2 Answers

Sorted by: Highest score (default)

2

If I understand the data correctly, what you have is a set of 4 mutually exclusive outcome classes. Under reasonable assumptions (in particular, the "independence of irrelevant alternatives") that can be handled by multinomial logistic regression, which can be implemented via a neural net as shown in the link. You then convert to probabilities after modeling.

You have to model counts rather than proportions, as a proportion based on 1000 observations is much more reliable than one based on 10 observations. Although the data format in the linked example above has a single outcome value for each data row, multinomial regression can handle aggregated data like you show. The `multinom()` function in the `nnet` package can accept an outcome that is "a matrix with K columns, which will be interpreted as counts for each of K classes" (from the help page). That seems to be just what you have. Then you use your `Xs` as the predictors for the regression, and ultimately represent the results in the probability scale.

Interfacing with other software like the `emmeans` package works better if you reformulate the data into a long form. Each data row is transformed into a number of rows equal to the number of outcome categories. The outcome is represented as levels of a categorical variable, and the number of counts with that outcome is provided to the `weights` argument of the function. There are a couple of worked-through examples on this page.

I'm not sure about how to implement multinomial outcomes with a gradient boosted machine, however.

Share Cite Improve this answer Follow

edited Nov 30, 2022 at 13:08

answered Nov 29, 2022 at 21:39

EdM
89k 10 88 260

I might have been not clear enough when writing the example, but I don't think that problem fits to a multinomial logistic regression. It's close, because the format of the output is pretty much what I'm looking for, but the input is different. Would you mind looking at the updated description of the question? – Matek Nov 29, 2022 at 23:38

1 @Matek as far as I can tell from the updated description, your data format can fit quite nicely into multinomial regression, perhaps with a bit of simple reformatting. I expanded the answer a bit to describe the approaches. – EdM Nov 30, 2022 at 13:09

Thanks, multinomial regression it is then (it was generally my primary choice/intuition). My format is unfortunately not supported by any of the GBMs so I'll either try working around that or just use NN. – Matek Nov 30, 2022 at 13:42

1 @Matek check whether any of your GBMs can accept a multi-level categorical outcome with weights. It's pretty simple to reformat your data that way. If weights don't work, you could go into a super-long form and make one data row for each individual count. For example, the first row of your data example would become 1 row with `Y1` as outcome, 5 rows with `Y3` as outcome, and 8 rows with `Y4` as outcome. Each of those rows would have the same set of `Xs`. – EdM Nov 30, 2022 at 15:40

These are all very good hints. Thanks a lot Sir/Madam – Matek Nov 30, 2022 at 23:12

0

You can do this using a fractional multinomial logit model. It is a multivariate generalization of the fractional logit model proposed by Papke and Wooldridge (1996).

We will model the proportion of spending on 6 different categories by 392 Dutch cities in 2005. The categories are governing, safety, education, recreation, social, and urban planning. These shares sum to one.

The explanatory variables are

- the average value of a house (in 100K euros)
- population density (thousands of persons per square km)
- a dummy for no left party in city government
- a dummy for left parties being a minority in city government

First, we load the data and fit the FML model in Stata:

```
. use http://fmwww.bc.edu/repec/bocode/c/citybudget.dta, clear
(Spending on different categories by Dutch cities in 2005)

. fmlogit governing safety education recreation social urbanplanning, ///
> eta(i.minorityleft i.noleft c.houseval c.popdens) nolog

ML fit of fractional multinomial logit                 Number of obs =    392
                                                       Wald chi2(20) = 275.23
Log pseudolikelihood = -673.12025                      Prob > chi2   = 0.0000

--------------------------------------------------------------------------------
                 |               Robust
                 | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-----------------+--------------------------------------------------------------
eta_safety       |
  1.minorityleft |   .1893638   .0596067     3.18   0.001     .0725368    .3061908
        1.noleft |    .082542   .0616854     1.34   0.181    -.0383592    .2034432
        houseval |  -.1400078   .0558587    -2.51   0.012    -.2494889   -.0305266
         popdens |   .0115814   .0212536     0.54   0.586    -.0300748    .0532377
           _cons |     .74898    .092535     8.09   0.000     .5676147    .9303453
-----------------+--------------------------------------------------------------
eta_education    |
  1.minorityleft |   .0387367   .1181969     0.33   0.743    -.1929249    .2703983
        1.noleft |  -.3648018   .1185739    -3.08   0.002    -.5972024   -.1324013
        houseval |  -.6371485   .1248264    -5.10   0.000    -.8818037   -.3924933
         popdens |   .0927616   .0374607     2.48   0.013     .0193399    .1661832
           _cons |   1.215266   .1979107     6.14   0.000     .8273682    1.603164
-----------------+--------------------------------------------------------------
eta_recreation   |
  1.minorityleft |   .2226632    .071707     3.11   0.002     .0821201    .3632062
        1.noleft |   .0138519   .0757628     0.18   0.855    -.1346405    .1623443
        houseval |  -.2308754   .0705698    -3.27   0.001    -.3691897   -.0925611
         popdens |   .0720411   .0256728     2.81   0.005     .0217234    .1223588
           _cons |   .4208606   .1160496     3.63   0.000     .1934076    .6483136
-----------------+--------------------------------------------------------------
eta_social       |
  1.minorityleft |    .136064   .0895568     1.52   0.129    -.0394641     .311592
        1.noleft |  -.1467066   .0928848    -1.58   0.114    -.3287575    .0353442
```

The magnitude of the coefficients is hard to interpret, but we can calculate the average marginal effects of no left parties on the share spent education, social programs, and safety:

```
. margins, dydx(noleft) predict(outcome(education))

Average marginal effects                                Number of obs = 392
Model VCE: Robust

Expression: predicted proportion for outcome education, predict(outcome(education))
dy/dx wrt:  1.noleft

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    1.noleft |  -.0352453   .0090812    -3.88   0.000    -.0530442   -.0174464
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.

. margins, dydx(noleft) predict(outcome(social))

Average marginal effects                                Number of obs = 392
Model VCE: Robust

Expression: predicted proportion for outcome social, predict(outcome(social))
dy/dx wrt:  1.noleft

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    1.noleft |   -.022242   .0106828    -2.08   0.037    -.0431799   -.0013041
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

```
. margins, dydx(noleft) predict(outcome(safety))

Average marginal effects                                   Number of obs = 392
Model VCE: Robust
```

This means that having no left parties in government is associated with a 3.5 percentage point reduction in the education budget share, a 2.2 percentage point reduction in social spending, and a 2.5 percentage point increase in the safety budget share. These effects are also statistically significant.

These are technically finite differences rather than derivatives. All six effects should sum to zero, since spending more/less on one category means spending less/more elsewhere, but I did not want to show all six marginal effects to save space.

You might also try using Dirichlet regression, as I suggested in the comment above. The results are very similar for the Dutch cities:

```
. quietly dirifit governing safety education recreation social urbanplanning, ///
> mu(minorityleft noleft houseval popdens)

. margins, dydx(noleft) predict(outcome(education))

Average marginal effects                                   Number of obs = 392
Model VCE: OIM

Expression: predicted proportion for outcome education, predict(outcome(education))
dy/dx wrt:  noleft

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      noleft |  -.0279187   .0068151    -4.10   0.000    -.0412759   -.0145614
------------------------------------------------------------------------------

. margins, dydx(noleft) predict(outcome(social))

Average marginal effects                                   Number of obs = 392
Model VCE: OIM

Expression: predicted proportion for outcome social, predict(outcome(social))
dy/dx wrt:  noleft

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      noleft |  -.0238505   .0100093    -2.38   0.017    -.0434683   -.0042326
------------------------------------------------------------------------------

. margins, dydx(noleft) predict(outcome(safety))

Average marginal effects                                   Number of obs = 392
Model VCE: OIM
```

However, I don't think this will work for your problem since this approach cannot handle zero or all-in-one-bucket shares. I don't have such cities here, but your data example has this feature.

In short, the FML model is a good option. There is an R implementation here, not sure about Python or other stats packages.

Share Cite Improve this answer Follow

edited Dec 1, 2022 at 6:55

answered Dec 1, 2022 at 5:50
dimitriy
35k 6 76 157