IMT574 Take-Home Final Exam/Winter 2020
100 pts/20% of the final grade

Your name:

Deadline: Mon, Mar 16th, 11:59pm

Introduction

This is the final take-home exam. Rules:

- you may use all available materials, including internet sources
- you **must not communicate with the others**. This is an individual exam.
- References must be appropriately cited. Links to solutions copied from websites, such as StackOverflow, must be provided in code comments. If you are re-using code from your previous problem sets, mention it.
- You have to sign the compliance form, attached on the last page, and return it together with the rest of the exam (a cellphone photo of the signed form is OK).
- Please explain your answers and show all your work; a complete argument must be presented to obtain full credit.
- All plots must be appropriately labeled, and appropriate colors/labels/font sizes must be used. Here I mean appropriate *for this exam*, it does not have to conform to the looong explanations in the *Science* journal.
- you can ask us if you are stuck with something. I want you to spend most of the time on ML problems, not on an obscure technical issue. However, it's somewhat limited how much help we can offer.

I expect you to use existing libraries for most of the work, but you are also welcome to implement some of the stuff yourself, or re-use you own code from a problem set.

The exam contains 2 questions. The first one uses Earned Income Tax Credit (EITC) data to estimate the causal effect of changing the EITC limits in 1994, and the second one is about categorizing Amazon reviews. The first dataset is probably new to you, the second one is the same as in PS7.

As always, please submit your results in two ways: as a code (notebook) and as a final version (html/pdf/...).

Deadline: Monday, Mar 16th midnight.

Good luck!

Ze, Li, and Ott

1 Causal effect estimation (50pt)

Your first task is to estimate the effect of increasing the Earned Income Tax Credit (EITC).

EITC is one of the major social support programs in US, intended to both provide support and encourage low-wage parents to search for work. It is designed as a tax credit on *earned income* (mainly wage). In case of low low income, and dependent children, you can get back some of your taxes (up to $6,500 in 2019). However, if you do not pay any taxes (i.e. you do not work), you also cannot get anything back. It is also closely related to the number of children in family, in mid-1990s, the period we look at here, the families without children could claim very little EITC. It was originally introduced in 1975, and has been continuously amended afterwards. Your task here is to analyze the reform, conducted in 1993 that noticeably increased the maximum tax credit amount (from approximately $2,000 to $2,700) for families with children. Consult Adireksombat (2010), in particular figures 1, 2 for more information.

1. (2pt) Load EITC data (called *eitc.csv* on canvas). Familiarize yourself a little with it. It contains data about single-mother families:

   year

   urate local unemployment rate

   children number of children in family

   nonwhite racial dummy

   finc total family income, including earned and non-earned income, such as capital income and social support

   earn earned income, mainly wage

   age

   ed school education of mother, in years (college not reflected here)

   work dummy: 1 if working

2. (5pt) First, let's do some graphical exploration. Plot the probability of working as a function of number of children (let's say only categories: 0, 1, 2+) over years. Comment the plot: do you see any changes in families with children, compared to those without children between 1993 and 1994? Do you see any differences between 1 and 2+-children families? Note: the latter also experienced the EITC increase in 1995 and 1996.

Next, use differences-in-differences method to estimate the effect of increasing the EITC. Let's treat families with no children as the control group, as very little changed for those at that time. Let's ignore income brackets for simplicity. First a few general questions:

3. (5pt) which families constitute a valid treatment group for the reform of 1993? Which years are valid *before*-years and which years are valid *after* years?

And now to the business. Estimate the effect of EITC increase by DiD:

4. (2pt) select the correct subsample (correct years, correct number of children).

5. (10pt) estimate the effect using DiD. Do this using linear regression. Do not include any other covariates (except constant!)

6. (2+2pt) comment your results:

   (a) did the growth of EITC increase or decrease working by the affected families? By how much?

   (b) is the effect statistically significant?

7. (4+4pt) Comment on the counterfactual assumption here:
   (a) what is the counterfactual assumption?
   (b) do you find it convincing?

8. (5pt) above you estimated the DiD without any covariates. What do you think about this setup: should we include covariates?

9. (5pt) include all relevant covariates you have in the data and re-estimate your model. What do you find?

10. (4pt) Based on the two estimates, give your final word about the efficacy of EITC reform. Feel free to comment the methodology too.

2 Text Classification (50p)

This question is rather similar to the PS5 Naive Bayes. Your task is to take the same Amazon reviews you used in PS7, but this time not to cluster those, but to identify whether a review has less than 5 stars. But the good news is that you are free to use all libaries you wish (you can also use your own implementation).

The data is downloaded from http://jmcauley.ucsd.edu/data/amazon/ and contain four variables, *date*, *summary*, *review*, and *rating*. In the current context you only need *review* and *rating*.

First, let's load data and take a closer look at it.

1. (1pt) Load the data. Depending how you work, you may need to take a subset (but processing the whole dataset as I did takes ~700M RAM and 5 mins only). I'd recommend to start with 1000 reviews and scale it up. Note: please take a random sample, not just first n lines. If you want to make your results replicable, use `np.random.seed`.

2. (2pt) Remove all the missing and empty observations of *review* and *rating*.

3. (2pt) Create your outcome variable: is rating less than 5-stars?

4. (4pt) Take a look at some of the reviews. Just be looking at those, think if you can tell if these are less than 5-star reviews. What do you think, how well can one predict the review stars based on these texts?

Now it is time to turn to modeling.

5. (2pt) Convert your reviews into bag-of-words (BOW). You can use either binary (word is present/absent) or actual counts.

   Hint: check out the `CountVectorizer` options.

6. (35pt) Your main task is to create a model that predicts the outcome (5-stars or less). The model should be as good as you can do, so you should tune different parameters, and show how did you try these.

   It is your responsibility to use appropriate models, appropriate parameters, and an appropriate way to tune those. You should also explain why you use the approach you are using.

   Note: you may also adjust parameters for `CountVectorizer`.

Finally, comment your results.

7. (1pt) Which model(s) is the best and how well does it perform?

8. (2pt) Do your results broadly correspond to what you guessed based on reading the reviews?

9. (0pt) and finally-finally, just for curiosity: how much time did you spend on this exam?

You are done! Submit, and enjoy summer ☺!

References

Adireksombat, K. (2010) The effects of the 1993 earned income tax credit expansion on the labor supply of unmarried women, *Public Finance Review*, **38**, 11–40.

Statement of Compliance

Please copy and sign the following statement. You may do it on paper (and include the image file or drop the sheet in Ott's mailbox in MGH 370), or add the following text with your name and date in your final document.

I affirm that I have had no conversation regarding this exam with any persons other than the instructor or the teaching assistant. Further, I certify that the attached work represents my own thinking. Any information, concepts, or words that originate from other sources are cited in accordance with University of Washington guidelines as published in the Academic Code (available on the course website). I am aware of the serious consequences that result from improper discussions with others or from the improper citation of work that is not my own.

(signature)

(date)