Code Sample: LaTeX

Sara Gong

October 25, 2020

In this LaTeX code sample, I have included (1) an excerpt of a data description appendix I wrote as a research assistant at USC, and (2) a homework problem from a machine learning class. Note that the table in (1) is highly condensed: because the survey is still unpublished, I've only included the metadata.

1 Data Appendix

The data for this study was collected through the Accelerating Commercialization of Collegiate Engineering and Science Survey (ACCESS), a project supported by the National Science Foundation's I-Corps Program to study early-stage entrepreneurs and the evolution of nascent technology ventures.

The ACCESS project spanned six iterations of a survey that was implemented in Qualtrics. Each version of the survey contained around 60 questions (encoded in approximately 200 variables) on topics that included the demographic characteristics of entrepreneurs, their areas of knowledge and experience, the activities and goals of their companies, their relationships with their advisors, and the scope and development of their business models. The six versions differed only by the inclusion or exclusion of a few questions, and every shared question was identical in wording.

From 2014 to 2019, over 500 responses were collected from entrepreneurs participating in programs held by the Innovation Node-Los Angeles. In total, these responses represented 326 entrepreneurs and 273 companies; this included entrepreneurs who were involved with more than one company, as well as companies from which multiple founders responded. We also tracked entrepreneurs and companies over time, allowing for the construction of two-wave panel datasets with units of either individual entrepreneurs or distinct companies. In the panel with units of individual entrepreneurs, the sample size was 135 and the average number of days between responses was 359. In the panel with units of distinct companies, the sample size was 120 and the average number of days between responses was 354.

To construct our datasets for analysis, we consolidated, harmonized, and validated responses across the six versions of the ACCESS survey. Our methods for data cleaning were primarily programmatic. While some parts of this process–for instance, assigning unique IDs to entrepreneurs and companies who gave nonidentical identifying information–required manual methods, we did modify any variables in the raw data. Instead, we created new variables to aid in other operations executed by scripts in R. Besides the IDs for entrepreneurs and companies, these variables included indicator variables that facilitate the construction of the cross-sectional and panel datasets, depending on the desired unit of analysis. We also generated additional variables that were useful to our analysis, as well as clean versions of variables whose values are analyzed as numbers or dates. From these steps, we obtained a master dataset containing 507 observations of 214 variables.

The following table details the variables in the master dataset and, if applicable, the method of their construction:

Variable	**Description or Question Text**	**Type**	**Values**
Survey_Progress	Qualtrics: Survey Progress	Text	
Survey_RecordedDate	Qualtrics: Survey Recorded Date	Text	
Survey_StartDate	Qualtrics: Survey Start Date	Text	
Survey_EndDate	Qualtrics: Survey End Date	Text	
Survey_ResponseType	Qualtrics: Survey Response Type	Text	
Survey_IP	Qualtrics: Survey IP Address	Text	
Survey_Duration	Qualtrics: Survey Duration	Text	
Survey_Finished	Qualtrics: Survey Finished	Text	
Survey_ID	Qualtrics: Survey ID	Text	
Survey_LastName	Qualtrics: Survey Last Name	Text	
Survey_FirstName	Qualtrics: Survey First Name	Text	
Survey_Email	Qualtrics: Survey Email	Text	
Consent	Consent Form	Binary	Yes, No
Consent_Media	I agree to be audio/video recorded/photographed.	Binary	Yes, No
Version	Source of response among various Qualtrics survey versions	Categorical	1.1, 1.2, 1.3, 2.1, 2.2, 2.3
Wave_Unique_I	Identifier to assign responses as Wave 1 or Wave 2 responses, in the dataset with units of unique individuals	Categorical	1, 2, NA
Wave_Unique_C	Identifier to assign responses as Wave 1 or Wave 2 responses, in the dataset with units of unique companies	Categorical	1, 2, NA
Wave_Unique_IC	Identifier to assign responses as Wave 1 or Wave 2 responses, in the dataset with units of one-to-one individual-company pairs	Categorical	1, 2, NA
RespondentID	Respondent identifier	Text	1-399
CompanyID	Company identifier	Text	1-265
INLA_TeamName	Team name (if relevant)	Text	
Name	Please enter your: - Name	Text	
Email	Please enter your: - Email	Text	
CompanyName	Does your company have a name?	Binary	Yes, No

2 Homework Problem

Recall Corollary 4.6: Let H be a finite hypothesis class, let Z be a domain, and let $l : H \times Z \rightarrow [0,1]$ be a loss function. Then, H has the uniform convergence property with sample complexity

$$m_H^{UC}(\epsilon, \delta) \leq \frac{log(2|H|/\delta)}{2\epsilon^2}$$

and is agnostic PAC learnable by ERM with sample complexity

$$m_H(\epsilon, \delta) \leq m_H^{UC}(\epsilon/2, \delta) \leq \frac{2log(2|H|/\delta)}{\epsilon^2}.$$

Now, if the range of the loss function is instead $[a,b]$, we can repeat the original argument that finite classes have the uniform convergence property.

Take H and fix ϵ, δ. We need to find a sample size m that guarantees for any distribution D, the probability over samples of size at least m that $|L_S(h) - L_D(h)| \leq \epsilon$ for all $h \in H$ is at least $1 - \delta$. This event is the same as the event that the probability over samples of size at least m that the complement, $|L_S(h) - L_D(h)| > \epsilon$ for some $h \in H$, is less than or equal to δ.

Note that:

$$\begin{aligned} &P(\{S : \exists h \in H : |L_S(h) - L_D(h)| > \epsilon\}) \\ &\qquad = P(\cup_{h \in H}\{S : |L_S(h) - L_D(h)| > \epsilon\}) \\ &\qquad \leq \textstyle\sum_{h \in H} P(\{S : |L_S(h) - L_D(h)| > \epsilon\}) \text{ by the union bound.} \end{aligned}$$

Also, by Hoeffding's Inequality, $P(S : |L_S(h) - L_D(h)| > \epsilon) \leq 2\exp\left(-2m\epsilon^2/(b-a)^2\right)$ since $L_S(h)$ is the sample average of m random variables with expectation $L_D(h)$. Then:

$$\begin{aligned} &P(\{S : \exists h \in H : |L_S(h) - L_D(h)| > \epsilon\}) \\ &\qquad \leq \textstyle\sum_{h \in H} 2\exp\left(-2m\epsilon^2/(b-a)^2\right) \\ &\qquad \leq |H| 2\exp\left(-2m\epsilon^2/(b-a)^2\right) \end{aligned}$$

So, choosing $m \geq \frac{log(2|H|/\delta)(b-a)^2}{2\epsilon^2}$ will ensure that $P(\{S : \exists h \in H : |L_S(h) - L_D(h)| > \epsilon\}) \leq \delta$. This gives an upper bound for the sample complexity for the uniform convergence property.

Finally, to find an upper bound for the sample complexity for agnostic PAC learnability, I just use Corollary 4.4 to show that $m_H(\epsilon, \delta) \leq m_H^{UC}(\epsilon/2, \delta) \leq \frac{2log(2|H|/\delta)(b-a)^2}{\epsilon^2}$.