Project Proposal

Siyu Chen

December 17, 2023

Abstract

reference link https://www.kaggle.com/code/sumithbhongale/notebook4186f97a6b
https://www.kaggle.com/code/kerneler/starter-college-admissions-c79081de-c
most relevant https://www.kaggle.com/code/apollostar/which-college-is-best-for-you#student-case-studies
The IPEDS dataset https://www.kaggle.com/code/yatin9045/us-university-selection-analysis/input
Illustration https://storage.googleapis.com/kagglesdsdata/datasets/11/6609/FullDataDocumentation.pdf?X-Goog-Algorithm=GOOG4-RSA-SHA256&X-Goog-Credential=gcp-kaggle-com%40kaggle-161607.iam.gserviceaccount.com%2F20231217%2Fauto%2Fstorage%2Fgoog4_request&X-Goog-Date=20231217T170158Z&X-Goog-Expires=259200&X-Goog-SignedHeaders=host&X-Goog-Signature=dbdc1da801cf76fb6b477286d5eec1b9d

1 Modeling the Admission Rate

I model the acceptance likelihood as a function of the test score (including both the SAT and ACT scores, but without the high school GPA) and the student's gender. We use a product model for the likelihood as the following:

$$\ell(s, r, g, S, C) = \ell_{score}(s) \cdot \ell_{gender}(g)$$

where s is the test score vector, g represents the gender. Since there is only information on the enrolled students' (posterior) rather than the applicants' (prior) features such as percentage of each race, I do not model other factors' influence on the admission likelihood. I illustrate the model in the following.

Test Score Likelihood Model. We model the test score prior as $s \sim p_{score} = \mathcal{N}(\mu, \Sigma)$, where we assume Σ to be diagonal (This assumption is a little bit strong. But as our dataset only contains the 25 and 75 quantile of each individual score, it is hard to extract the correlation between these scores in the variance. Note that we still allow their means to be related.) The distribution of the test score among the admitted students depends on two factors: the school's admission likelihood $f_{school}(s)$ based on the test scores and the student's likelihood of accepting this school's offer $f_{student}(s)$ given his/her test scores. Note that the school's admission likelihood $f_{school}(s)$ should be monotonically increasing with respect to the test scores while the student's preference $f_{student}(s)$ could be monotonically decreasing with respect to the test scores. The rationality behind this is that schools tend to admit students with higher test scores while students with high test scores seldom consider schools with low rankings, which is often associated with low admission standards.

Therefore, we can model enrollment likelihood as $f_{enroll}(s) = f_{school}(s)f_{student}(s)$. By the previous discussion, we can model $f(s)$ as another Gaussian distribution $\mathcal{N}(\mu_1, \Sigma_1)$, where Σ_1 is also diagonal. Therefore, the posterior distribution of scores upon acceptance is just

$$q(s \mid a = 1) \propto f(s)\ell_{score}(s) \sim \mathcal{N}\left(\underbrace{(\Sigma^{-1} + \Sigma_1^{-1})^{-1}(\Sigma^{-1}\mu + \Sigma_1^{-1}\mu_1)}_{\mu_2}, \quad \underbrace{(\Sigma^{-1} + \Sigma_1^{-1})^{-1}}_{\Sigma_2}\right),$$

where $a \in \{0, 1\}$ is the indicator for admission. Entrywise, we have

$$\mu_2^i = \frac{(\sigma_1^{(i)})^2\mu^i + (\sigma^{(i)})^2\mu_1^i}{(\sigma_1^{(i)})^2 + (\sigma^{(i)})^2}, \quad (\sigma_2^{(i)})^2 = \frac{(\sigma_1^{(i)})^2 \cdot (\sigma^{(i)})^2}{(\sigma_1^{(i)})^2 + (\sigma^{(i)})^2}. \tag{1.1}$$ eq: posterior e

Therefore, we can estimate the posterior Gaussian, and solve for the admission model (μ_1, Σ_1). The remaining question is how to obtain the mean and variance of the prior. Note that the mean and variance are biased in the data since the test scores are collected only for enrolled students here. However, this websites https://blog.prepscholar.com/sat-historical-percentiles-for-2014-2013-2012- and https://www.act.org/content/dam/act/unsecured/documents/Natl-Scores-2013-National2013.pdf comes to our rescue as it collectes summary for SAT scores in 2013. We summarize the results on the following table

	mean	standard deviation
SAT Critical Reading	496	115
SAT Mathematics	514	118
SAT Writing	488	114
ACT Composite	20.9	5.4

Note that the estimated standard deviation $\sigma_2^{(i)}$ could potentially be larger than the prior, which leads to no reasonable solution according to (1.1). This could happen when our model does not capture what is happening in the real world, for instance, when the acceptance likelihood is far from a Gaussian distribution. However, as indicated by the experimental results, such an event is very rare in the data (8 in 468 schools). In that case, I still estimate the variance according to (1.1) but set the "nan" and negative values to infinity. The reason is that when a school' admission rate on a spefic score violates the model, a natural thing to do is ignoring the dependency of the admission rate on the violated score, which is achieved by using an extremely flat normal.

We then model the likelihood by the following function "truncated dnorm"

$$\ell_{score}(s) = \prod_i \exp\left(-\frac{(\mathrm{ReLU}(\mu_1^{(i)} - s^{(i)}))^2}{2(\sigma_1^i)^2}\right), \tag{1.2}$$ eq: truncated d

where $\mathrm{ReLU}(x) = \max\{0, x\}$ is the rectified linear unit function, and we allow $\sigma_1^{(i)}$ to be $+\infty$. The intuition for (1.2) is that following what I have assumed $\ell_{score}(s) = f_{school}(s)f_{student}(s)$ and the fact schools always tend to choose students with higher scores, where students tend to choose better schools (with higher admission standards), we can model $f_{school}(s)$ and $f_{student}(s)$ as two "half"

Gaussian, i.e.

$$f_{school}(s) = \exp\left(-\frac{(\mathrm{ReLU}(\mu_1^{(i)} - s^{(i)}))^2}{2(\sigma_1^i)^2}\right),$$
$$f_{student}(s) = \exp\left(-\frac{(\mathrm{ReLU}(s^{(i)} - \mu_1^{(i)}))^2}{2(\sigma_1^i)^2}\right).$$

The product of these two functions then gives us the Gaussian likelihood $f_{enroll}(s)$ of enrollment.

Likelihood Model for Gender. We next characterize the influence of gender. We assume a priori that the likelihood of a applicant being male or female is equally likely, i.e., the gender prior is 0.5 for both male and female applications (which is close to the truth but may not be accurate in reality). Suppose that gender is weakly coupled with academy performance. We model the bias between gender as a likelihood ratio, i.e.,

$$\ell_{gender}(g) = \frac{\text{percentage}(g)}{\max\{\text{percentage}(g), 1 - \text{percentage}(g)\}},$$

Note that when there is no bias, the gender does not influence the admission rate. When bias exists, ℓ_{gender} models the "disadvantage" of a specific gender in the admission process.

2 Bayesian Factor

In the analysis, we take an example by using the student's family income to illustrate the computation of Bayes Factors. Let $E = \{\text{student's family-income} > x\}$ be the event of interest. Let H denote the school of interest and Ω denote all the schools. Let's define $p(H)$ as the prior probability of enrollment at school H, namely the proportion of students enrolled in school H among all the students enrolled in Ω. The Bayesian Factor $\mathrm{BF}(E, H)$ is then defined as the likelihood ratio as the following:

$$\mathrm{BF}(E, H) = \log \frac{p(E \mid H)}{p(E \mid \Omega)}.$$

A simple application of the Bayesian rule gives

$$\mathrm{BF}(E, H) = \log \frac{p(E \mid H)}{p(E \mid \Omega)} = \log \frac{p(H \mid E)}{p(H)} = \log \frac{p(H, E)}{p(H)p(E)}.$$

Note that the Bayes Factor reflects the "advantage" of school H over the other schools in terms of the event E. I will include the Bayes Factor in the analysis with some manually selected coefficients to determine the score of each school.

3 Data Cleaning

References

https://www.kaggle.com/code/sumithbhongale/notebook4186f97a6b
https://collegescorecard.ed.gov/
https://nces.ed.gov/ipeds/use-the-data