Preferences in college applications

A non-parametric Bayesian analysis of top-10 rankings

Alnur Ali[1] Thomas Brendan Murphy[2] Marina Meilă[3]
Harr Chen[4]

[1]Microsoft

[2]University College Dublin

[3]University of Washington

[4]Massachusetts Institute of Technology

Outline

College Applications

- Irish college applicants apply through a central system administered by the College Applications Office (CAO).
- Applicants list up to ten degree courses in order of preference.
- Applicants are awarded points on the basis of their Leaving Certificate results; these determine course entry.

Students pay price of exam success as points jump

John Walshe
Education Editor

INSIDE:
Going to college: All the CAO offers in our 12-page supplement
Points analysis Page 12

High flyers' hopes dashed as points hit record highs

John Walshe
Education Editor

The Top Ten CAO Courses

1) Medicine (with Leaving Cert scholarship)	Royal College of Surgeons	590*
2) Law and French	Trinity College Dublin	590
3) Medicine	University College Dublin	570*
4) Law	Trinity College Dublin	570*
5) Medicine	Trinity College Dublin	570*
6) Medicine	NUI Galway	570*
7) Medicine	Royal College of Surgeons	570*
8) Medicine	University College Cork	565*
9) Physiotherapy	University of Limerick	565*
10) Psychology	Trinity College Dublin	565*

*Not all on these points get a place - random selection applies

Masterclass students set new record for grades

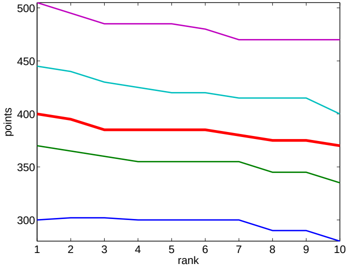
Minister insists school subjects are not being 'dumbed down'

John Walshe and Catherine Donnelly

Inside Results day dawns for students P8
Leaving home for a life in college P14

Goals

- It has been postulated that a number of factors influence course choices:
    - Institution & Location
    - Degree subject
    - Degree type (Specific vs. General)
    - Points Requirement
    - Gender



Do points requirements influence ranks?

Dataset

- We study the cohort of applicants to degree courses from the year 2000.
- The applications data has the following properties:
  - There were 55737 applicants;
  - They selected from a list of 533 courses;
  - Applicants selected up to 10 courses.

Data Coding

- The data coding $(s_1, s_2, \ldots, s_t)$ of $\pi|\sigma$ is defined by

$$s_j + 1 = \text{rank of } \pi^{-1}(j) \text{ in } \sigma \text{ after removing } \pi^{-1}(1:j-1).$$

Example, if $\sigma = [a\ b\ c\ d]$ and $\pi = [c\ a\ b\ d]$

$\pi^{-1}(1) = c$ $\quad s_1 = 2$

$\pi^{-1}(2) = a$ $\quad s_2 = 0$

$\pi^{-1}(3) = b$ $\quad s_3 = 0$

$\pi^{-1}(4) = d$ $\quad s_4 = 0$

σ

a	b	**c**	d
a	b	·	d
·	**b**	·	d
·	·	·	**d**

- Kendall's distance is $d_{\text{Kendall}}(\pi, \sigma) = \sum_{j=1}^{t-1} s_j$.

Generalized Mallow's models

- Mallow's model assumes that

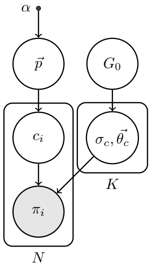
$$P(\pi|\sigma,\theta) = \frac{1}{\psi(\theta)} \exp\left(-\theta \sum_{j=1}^{t-1} s_j(\pi|\sigma)\right).$$

- Can extend Mallow's model to allow for varying precision in ranking

$$P(\pi|\sigma,\vec{\theta}) = \frac{1}{\psi(\vec{\theta})} \exp\left(-\sum_{j=1}^{t-1} \theta_j s_j(\pi|\sigma)\right).$$

- Location parameter σ, scale parameters $(\theta_1, \ldots, \theta_{\max t-1})$.
- $\psi(\vec{\theta})$ is a tractable normalization constant.

Dirichlet process mixture models



- $\vec{p} \sim Dirichlet(\alpha/K, \ldots, \alpha/K)$
- $c_i \sim Multinomial(p_1, \ldots, p_K)$
- $\sigma_c, \vec{\theta}_c \sim G_0 \propto P^0(\sigma, \vec{\theta}; \nu, \vec{r})$
- $\pi_i \sim GM(\pi_i|\sigma_c, \vec{\theta}_c)$

- Prior: conjugate to GM, informative w.r.t. $\vec{\theta}$.
- DPMM benefits: no need to specify K upfront, identifies both large and small clusters.

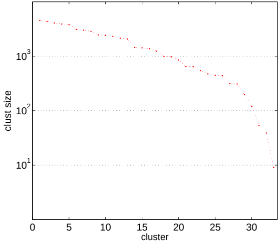
Gibbs sampler

1. Resample cluster assignments:
    1.1 Draw existing cluster w.p. $\propto \frac{N_c-1}{N+\alpha-1} GM(\pi|\sigma_c, \vec{\theta_c})$ or *Beta* function approximation.
    1.2 Draw new cluster w.p. $\propto \frac{\alpha}{N+\alpha-1} \frac{(n-t)!}{n!}$.
2. Resample cluster parameters:
    2.1 Draw $\vec{\theta_c}$ by *slice sampling* or a *Beta* distribution approx.
    2.2 Draw σ_c "stage-wise" or by a *Beta* function approx.

Beta approx. based sampler (Beta-Gibbs) faster than slice based sampler (Slice-Gibbs) (per iteration & overall time to convergence).

General properties of the clusterings

- The DPMM found 164 clusters.
- Thirty three of these clusters had nine or more members.

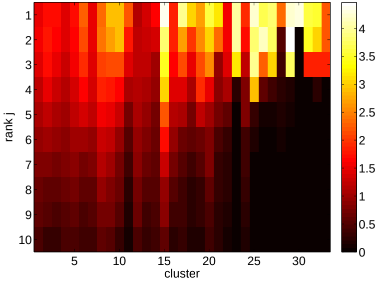


- The clusters were characterized by a number of features.

Cluster	Size	Description	Male (%)	Points Average (SD)
1	4536	CS & Engineering	77.2	369 (41)
2	4340	Applied Business	48.5	366 (40)
3	4077	Arts & Social Science	13.1	384 (42)
4	3898	Engineering (Ex-Dublin)	85.2	374 (39)
5	3814	Business (Ex-Dublin)	41.8	394 (32)
6	3106	Cork Based	48.9	397 (33)
...	...	...	...	...
33	9	Teaching (Home Economics)	0.0	417 (4)

Precision

- The precision parameters (θ_j) were very high for top rankings.



- The θ_j values tended to decrease with j.
- In many cases, the θ_j values dropped suddenly after a particular point.
- The central ranking σ for each cluster is of length 533; the θ_j values suggested a point to truncate the ranking.

Overall trends

- Subject
  - Subject matter is a key determinant of course choice.
  - The courses chosen are similar in subject area.
  - Some opt for general degrees (eg. Science) and others opt for specific (eg. Chemical Engineering).
- Gender
  - There is quite a difference in the percentage male/female applicants in some clusters.
  - Males tend to dominate CS/Engineering clusters.
  - Females tend to dominate social science/education clusters.
- Geography
  - There is evidence of the college location influencing choice.
  - The sixth largest cluster is dominated by courses from colleges in Cork (CIT and UCC).
  - There is evidence of a mix of subject matter and geography having a joint effect; the fourth largest cluster is dominated by engineering courses outside Dublin.

Overall trends

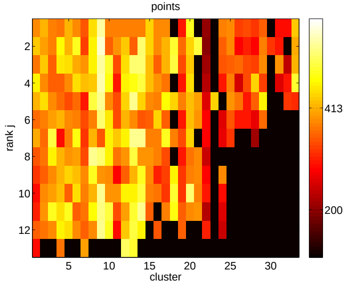
- Subject
  - Subject matter is a key determinant of course choice.
  - The courses chosen are similar in subject area.
  - Some opt for general degrees (eg. Science) and others opt for specific (eg. Chemical Engineering).
- Gender
  - There is quite a difference in the percentage male/female applicants in some clusters.
  - Males tend to dominate CS/Engineering clusters.
  - Females tend to dominate social science/education clusters.
- Geography
  - There is evidence of the college location influencing choice.
  - The sixth largest cluster is dominated by courses from colleges in Cork (CIT and UCC).
  - There is evidence of a mix of subject matter and geography having a joint effect; the fourth largest cluster is dominated by engineering courses outside Dublin.

Overall trends

- Subject
  - Subject matter is a key determinant of course choice.
  - The courses chosen are similar in subject area.
  - Some opt for general degrees (eg. Science) and others opt for specific (eg. Chemical Engineering).
- Gender
  - There is quite a difference in the percentage male/female applicants in some clusters.
  - Males tend to dominate CS/Engineering clusters.
  - Females tend to dominate social science/education clusters.
- Geography
  - There is evidence of the college location influencing choice.
  - The sixth largest cluster is dominated by courses from colleges in Cork (CIT and UCC).
  - There is evidence of a mix of subject matter and geography having a joint effect; the fourth largest cluster is dominated by engineering courses outside Dublin.

Points

- The points requirements for the courses in the truncated central rankings were not monotonically decreasing in any cluster.



- This suggests that points requirements are not important when students are ranking courses.

Conclusions & Lessons Learned

- The CAO system appears to be working more effectively than many suggest.
- The clusters revealed in this analysis tend to be cohesive in subject matter.
- The focus of possible improvements to the CAO system might be directed at how points are scored.
- The Generalized Mallows DPMM facilitated discovering small clusters that were missed in previous analyses.
- The model also allowed for the study of precision in rankings within clusters.

Questions?

Thanks!