CSE 190 – Lecture 12

Data Mining and Predictive Analytics

Various stuff

HW 4!!!

- Is out
- There was a mistake – I forgot to include a regularizer which is needed to ensure that the problem is well-posed

Assignment 1

Task 1:

This leaderboard is calculated on approximately 50% of the test data.
The final results will be based on the other 50%, so the final standings may be different.

See someone using multiple accounts?
Let us know.

#	Δ5d	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	new	Andrew	0.85584	15	Mon, 11 May 2015 01:48:59
2	new	allen0139	0.85584	6	Tue, 12 May 2015 22:23:48
3	↓2	SamuelVange	0.85332	6	Mon, 04 May 2015 22:57:59
4	↑2	ArchitKhosla	0.83024	31	Tue, 12 May 2015 08:23:06 (-24.2h)
5	–	┬─┬ ノ(' - 'ノ)	0.80784	17	Wed, 13 May 2015 04:57:25
6	↑1	maddymanu	0.69716	43	Wed, 13 May 2015 03:54:42 (-25.6h)
7	new	HardyLou	0.69108	5	Wed, 13 May 2015 21:34:15 (-11.8h)
8	↓6	(ﾉ°□°)ﾉ ┻┻	0.68568	14	Wed, 13 May 2015 07:17:17 (-5.3d)
9	↓6	RichardTran93	0.67804	1	Fri, 01 May 2015 07:09:00
10	↓6	rantaoca	0.67804	1	Fri, 01 May 2015 08:39:03
11	new	Raymond Qiu	0.67804	2	Sun, 10 May 2015 04:09:04 (-0.3h)
12	new	guilly	0.67804	4	Wed, 13 May 2015 17:06:23 (-2.7d)
13	new	Jacob Tao	0.67804	3	Wed, 13 May 2015 00:33:42 (-30.2h)
14	new	✂	0.67804	4	Wed, 13 May 2015 04:52:02 (-24.5h)
15	new	Vnator	0.67804	1	Wed, 13 May 2015 21:19:59 (-22.4h)
16	new	Nazia Rizvi	0.67804	1	Wed, 13 May 2015 20:58:56
17	new	Margaret	0.67804	1	Wed, 13 May 2015 21:16:35
18	↓10	y2bd	0.58512	3	Thu, 07 May 2015 06:48:30 (-1.3h)

Assignment 1 – Task 1

- What are the limitations of the baseline?
- The baseline only considers *items*. Can you do better by also considering *users?*
- What other features might be useful here?

Assignment 1 – Task 2

- What are the limitations of the baseline?
- What do we know about the way Amazon surfaces helpful reviews to users?

HW: 3.2

Rating prediction

Let's start with the simplest possible model:

$$f(u, i) = \alpha$$

user item

$$\alpha = \frac{1}{N} \sum_{u,i \in \text{training data}} r_{u,i}$$

Here the RMSE is just equal to the **standard deviation** of the data

(and we cannot do any better with a 0th order predictor)

Rating prediction

What about the **2nd** simplest model?

$$f(u, i) = \alpha + \beta_u + \beta_i$$

user item

how much does this user tend to rate things above the mean?

does this item tend to receive higher ratings than others

e.g.

$\alpha = 4.2$

$\beta_{\text{pitch black}} = -0.1$

$\beta_{\text{julian}} = -0.2$

Rating prediction

The optimization problem becomes:

$$\arg\min_{\alpha,\beta} \underbrace{\sum_{u,i}(\alpha + \beta_u + \beta_i - R_{u,i})^2}_{\text{error}} + \underbrace{\lambda \left[\sum_u \beta_u^2 + \sum_i \beta_i^2\right]}_{\text{regularizer}}$$

Jointly convex in \beta_i, \beta_u. Can be solved by iteratively removing the mean and solving for beta

Rating prediction

Iterative procedure – repeat the following updates until convergence:

$$\alpha = \frac{\sum_{u,i \in \text{train}} (R_{u,i} - (\beta_u + \beta_i))}{N_{\text{train}}}$$

$$\beta_u = \frac{\sum_{i \in I_u} R_{u,i} - (\alpha + \beta_i)}{\lambda + |I_u|}$$

$$\beta_i = \frac{\sum_{u \in U_i} R_{u,i} - (\alpha + \beta_u)}{\lambda + |U_i|}$$

(exercise: write down derivatives and convince yourself of these update equations!)

HW: 3.2

- How else could we achieve the same result?
- Coordinate descent vs. gradient descent

Midterm

- Average mark was high (>25)
- Median mark was high (>25)
- Minimum mark was pretty high too

1. Interpretation of parameters

- Suppose we have a linear regression model to predict college GPA
- One of the features of this model encodes whether a student owns a car
- The fitted model looks like:

$$y = \ldots - 0.4[\text{owns a car}] + \ldots$$

Conclusion: "The GPA of the average student *who owns a car* is 0.4 lower than that of the average student"
Q: is this conclusion reasonable?

Difficult Qs

- 11c
- 12
- 15
- 17
- 19
- 2?