محمدامین شفیعی

سوال اول

۱) پس مطالعه ی تسک من تصور می کردم که برای هر کوئری حداقل یک سری پست لود شدن و ممکن هست یک سری کلیک هم داشته باشه اما در عمل برخی از کوئری ها وجود داشتند که کلیک داشتند ولی لود نداشتند.(مثلا کوئری با آی دی ۰۰۱۴۲c۵۹-۷۴۵c-۴۰۰۴-a۹۵۵-۶۹۸ddcf۱faa۶)

۲) برخلاف گفته ی صورت تسک لیست tokens همواره ۲۴ تایی نیست (len طول لیست tokens است)(احتمالا مربوط به قسمت هایی است که آگهی ها تمام شده اند.)

```
queries.loc[queries.len != 24].head()
```

	action	created_at	source_event_id	device_id	post_page_offset	tokens	post_index_in_post_list	post_token	len
6	load_post_page	1.609550e+12	576278e1-9af0-4ed9-95c8-09a6b6c2e958	TVDfsdqOS1eqxuUm2EHe6g	0.0	[wXr_DuyQ, wXpr43e5, wXl_rC8t, wXkLrzBv, wXV7M...	NaN	NaN	10
32	load_post_page	1.609550e+12	c38f238c-83dc-4b18-a928-728fe8947608	VszFqQDyS4WmwuBF6pYU7A	0.0	[wXrLw3_f, wXnf6zpr, wXk3doU-, wXkjcLjv, wXTLC...	NaN	NaN	8
33	load_post_page	1.609540e+12	ea658b82-c5cf-4b2a-957a-3f2fb6b91fab	oYxF1k4UT2yggpprUyUTlg	0.0	[wX4tAmXm, wX1Jmy5h, wXeSEm3B, wXVD7SNe, wXrO3...	NaN	NaN	7
57	load_post_page	1.609550e+12	e3ad3505-c716-41cd-8df0-b7559a7501c7	JD6SPf7DRAysn8PRCvQT3g	1.0	[wXdjs38e, wXcX25yp, wXWDQayT, wXXnhvgk, wXUrX...	NaN	NaN	7
67	load_post_page	1.609550e+12	0a951231-3f7d-47c2-bd1f-8f6899007584	fUlvCCkJTXSGg4wgh89mVA	27.0	[wXKvDwfH, wXK7iQH8, wXKnBWIR, wXBXh1G9, wXKbw...	NaN	NaN	21

۳) در فایل CSV داده شده مقادیر ستون created_at دقت خود را از دست داده بودند که با قرار دادن فایل جدید این مشکل رفع شد.

سوال سوم

هر کدام از معیار های گفته شده فوایدی دارند؛ به طور مثال متریک اول در واقع تعداد را وارد می کند و نشان دهنده این است که از نتایج نمایش داده شده چند مورد مطلوب کاربر بوده است؛ اما جایگاه نتایج مطلوب در این متریک تاثیر ندارد. در متریک دوم جایگاه تاثیر دارد اما تعداد موثر نیست (و البته میتوان گفت تقریبا پوشش دهنده متریک چهارم هم هست).

در مجموع به نظر من حالت ایده ای استفاده ی همزمان از دو متریک ابتدایی است که نکات مهم گفته شده را می توانند پوشش دهند. اما اگر قرار به انتخاب فقط یک متریک باشد؛ متریک اول گزینه جذاب تری است چرا در هر صورت نشان می دهد از بین نتایج چه قدر توانسته ایم کاربر را راضی کنیم.

سوال چهارم

همانطور که میدانیم تابع توزیع برنولی مطابق زیر است:

$$p(y|\theta) = \theta^y(1-\theta)^{1-y}$$

برای حل این سوال از logistic regression استفاده می کنیم.رگرسیون لجستیک مناسب زمان هایی است که نتیجه از بین دو کلاس انتخاب میشود. (مثلا {۰و ۱}). رگرسیون لجستیک که شبیه رگرسیون خطی هست با این تفاوت که رگرسیون لجستیک از توزیع برنولی استفاده می کند. اگر نتیجه ی این مدل را احتمال یک بودن(کلیک کردن کاربر) در نظر بگیریم روند کلی به صورت زیر خواهد بود (که از تابع Sigmoid استفاده کرده ایم):

$$p = \Pr(y_i = 1|\overrightarrow{x_i};\vec{\beta}) = \frac{e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}{1+e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i})}}$$

در اینجا باید ویژگی هایی از داده ورودی که میخواهیم در مدل استفاده کنیم را مشخص کنیم. در مجموعه نمونه های ما مواردی که action کلیک دارند؛ نمونه های کلاس یک و پست های که لود شده اند اما کلیک نگرفته اند نمونه های کلاس صفر هستند. برای ویژگی ها میتوانیم **توکن پست، رتبه ی پست در لیست ها، آی دی دستگاه ، آی دی کوئری** و **حتی زمان درخواست** را در نظر بگیریم. همچنین باید ضریب این ویژگی ها (β) را بهینه کنیم که نتیجه ی تابع بالا بیشترین شباهت را به مقدار اصلی در نمونه های ورودی داشته باشد. برای این کار از روش Maximum Likelihood Estimation استفاده می کنیم. با توجه به این که در p از sigmoid استفاده کرده ایم و تابع نمایی($e^{-\beta X}$) وجود دارد لگاریتم Likelihood را بیشینه میکنیم(بیشینه کردن خود تابع یا لگاریتم آن تاثیری در نتیجه ی مطلوب ندارد.)(n تعداد نمونه های ورودی است)

$$L(D,\vec{\beta}) = \log\left(\prod_{i=1}^{n} Pr(y_i = 1|\overrightarrow{x_i};\vec{\beta})^{y_i} \times Pr(y_i = 0|\overrightarrow{x_i};\vec{\beta})^{1-y_i}\right)$$

در تابع بالا بخش $Pr(y_i = 1|\overrightarrow{x_i};\vec{\beta})^{y_i} \times Pr(y_i = 0|\overrightarrow{x_i};\vec{\beta})^{1-y_i}$ در واقع همان تابع احتمال توزیع برنولی است. با بسط دادن تابع L به تابع زیر عبارت زیر می رسیم:

$$\sum_{i=1}^{n} y_i \times \log Pr(y_i = 1|\overrightarrow{x_i};\vec{\beta}) + (1-y_i)\log Pr(y_i = 0|\overrightarrow{x_i};\vec{\beta})$$

نکته ای که در مورد تابع بالا وجود دارد این است که این تابع نسبت به β مقعر و همگرا و بنابراین یک بیشنیه مطلق دارد. برای رسیدن به این بیشینه پس از مقداردهی رندومِ اولیه هر بار مقدار مشتق تابع L را برای یک نمونه ورودی محاسبه و در آن جهت حرکت میکنیم تا در نهایت مشتق به صفر میل کند. در واقع تا زمانی که مشتق به صفر میل کند هموار روند زیر را انجام می دهیم:

$$\begin{aligned}
&\text{for } \ i=0 \ \text{ to } \ n: \\
&\quad \text{for } \ j=0 \ \text{ to } \ m: \\
&\qquad \overrightarrow{\beta_j^{new}} = \overrightarrow{\beta_j^{old}} + \alpha\left(y_i - \frac{1}{1+e^{-(\beta_0^{old}+\beta_1^{old}x_{1,i}+\cdots+\beta_k^{old}x_{k,i})}}\right)\overrightarrow{x_{i,j}} \\
&\quad \beta^{old} = \beta^{new}
\end{aligned}$$

با استفاده از این مدل می توان مقدار متریک اول را به صورت غیر مستقیم به دست آورد به گونه که مدل را بر روی هر پست اجرا کرده و در نهایت تعداد موارد کلیک شده را به کل نتایج تقسیم کرد(همین روند را میتوان برای متریک های دوم و سوم هم اجرا کرد).(البته الگوریتم های دیگر این روند را ساده تر میتوانند انجام دهند.)

با توجه به اینکه این مدل برای classification است و متریک چهارم هم از همین دسته است برای متریک چهارم هم مناسب است. در صورتی که متریک دوم را داشته باشیم میتوان متریک پایانی را هم محاسبه کرد. اما به طور کلی با داشتن یکی از متریک ها؛ محاسبه ی همه ی متریک دیگر ممکن نیست.