به نام خدا

پروژه‌ی اول

رگرسیون خطی و دسته‌بندی

یادگیری ماشین - بهار ۱۴۰۱

فهرست مطالب

۰ نکته‌ها

نکته‌ها
۱. پروژه به صورت گروهی و گروه‌های دو نفره خواهد بود. ۲. از طریق این لینک کافیست یکی از افراد تیم اعضای تیم را مشخص کند. ۳. برای پیاده‌سازی پروژه از زبان پایتون استفاده کنید. ۴. پروژه را در قالب دفترچه‌ی Jupyter پیاده‌سازی کنید تا در کنار کد بتوانید مستندات خود را نیز اضافه کنید. ۵. برای تحویل پاسخ علاوه بر دفترچه‌ی خود باید خروجی پایتون آن را نیز تحویل دهید.

۱ مقدمه

در این پروژه به مسئله‌ی Learning to Rank (LTR) خواهیم پرداخت. به طور کلی هدف آموزش مدلی است که بتوان با آن به مسئله‌ی رتبه‌بندی نتایج در جست‌وجو پرداخت. برای حل این مسئله رویکردهای مختلفی وجود دارد که در زیر به دو نمونه از آن‌ها اشاره شده است:

- **point-wise**: در این روش ورودی مدل دوتایی (q, d) خواهد بود که q درخواست[۱] کاربر و d سندی[۲] است که می‌خواهیم رتبه‌ی[۳] آن را در مرتبط بودن با درخواست مشخص کنیم. خروجی مدل می‌تواند به صورت عددی حقیقی بین ۰ و ۲ باشد که میزان ارتباط سند با درخواست را مشخص می‌کند و هر چه به ۲ نزدیک‌تر باشد سند مربوطه به درخواست مرتبط‌تر خواهد بود.
- **pair-wise**: در این رویکرد ورودی مدل به صورت سه‌تایی (q, d_1, d_2) خواهد بود و هدف تشخیص این موضوع است که آیا d_1 به q مرتبط‌تر است یا d_2؟ خروجی مدل می‌تواند به صورت گسسته و مقدارهای ۰ یا ۱ باشد که ۱ به معنای این است که d_1 حداقل به اندازه‌ی d_2 به q مرتبط بوده، و ۰ مخالف آن را نشان می‌دهد.

۲ آشنایی با دادگان[۴] مسئله

دادگان استفاده شده LETOR 4.0 خواهد بود که در سال ۲۰۰۹ توسط شرکت مایکروسافت جمع‌آوری و منتشر شده است. این مجموعه از دادگان مختلفی تشکیل شده که در ماموریت‌هایی که در ادامه آمده‌اند از دادگان مناسب استفاده خواهد شد. اما به طور کلی هر داده‌ی آن که یک جفت (q, d) است توسط یک بردار ۴۶ بعدی کد شده که شامل تمام اطلاعات مورد نیاز این جفت می‌شود. برچسب هر داده نیز میزان ارتباط سند مربوطه به درخواست را مشخص می‌کند. به عنوان مثال در ادامه دو نمونه از داده‌های این دادگان آمده است:

```
2 qid:10032 1:0.056537 2:0.000000 3:0.666667 4:1.000000 5:0.067138 …
45:0.000000 46:0.076923 #docid = GX029-35-5894638 inc = 0.0119881192468859
prob = 0.139842

0 qid:10032 1:0.130742 2:0.000000 3:0.333333 4:0.000000 5:0.134276 …
45:0.750000 46:1.000000 #docid = GX140-98-13566007 inc = 1
prob = 0.0701303
```

که ستون اول برچسب داده بوده و ستون بعدی شماره‌ی درخواست است. ۴۶ ستون بعدی ویژگی‌های داده را تشکیل داده و هرچه پس از # آمده نیز توضیحاتی اضافی در رابطه با این سطر است.

[1] Query
[2] Document
[3] Rank
[4] Dataset

۳ شرح پروژه

۱.۳ ماموریت اول (روش **point-wise** و رگرسیون خطی)

در رویکردهای point-wise تابع هزینه به یک سند واحد در واحد زمان نگاه می‌کند. در این روش یک دسته‌بند / رگرسیون را بر روی یک سند آموزش می‌دهند تا میزان مرتبط بودن آن با درخواست فعلی را پیش‌بینی کنند. رتبه‌بندی نهایی صرفاً با مرتب‌سازی لیست نتایج بر اساس این امتیازهای اسناد به دست می‌آید. برای رویکردهای point-wise، امتیاز برای هر سند مستقل از سایر اسنادی است که در لیست نتایج جستجو هستند.

درواقع هزینه کل به عنوان مجموع هزینه‌های هر سند d_i که فاصله بین امتیاز پیش‌بینی شده (s_i) و امتیاز واقعی (y_i) برای $i = 1, 2, .., n$ است، محاسبه می‌شود. با انجام این کار، مسئله خود را به یک مسئله رگرسیون تبدیل می‌کنیم، جایی که مدلی را برای پیش‌بینی y آموزش می‌دهیم.

ما در این پروژه می‌خواهیم که شما با استفاده از روش point-wise یادگیری رتبه را انجام دهید. با این هدف مسئله را به شکل یک مسئله رگرسیون خطی فرمول‌بندی می‌کنیم که در آن بردار ورودی x را به عدد اسکالر y نگاشت می‌کنیم. شما این کار را به دو صورت زیر انجام خواهید داد:

آ) آموزش یک مدل رگرسیون خطی بر روی دادگان با استفاده از فرمول بسته[۵]. (این روش امتیازی است)

ب) آموزش یک مدل رگرسیون خطی بر روی دادگان با استفاده از روش کاهش گرادیان[۶].

دادگان این پروژه از جفت مقادیر ورودی x و مقادیر هدف t تشکیل شده است. مقادیر ورودی شامل ۴۶ ستون دادگان است و مقادیر هدف اسکالرها (برچسب‌های مرتبط) هستند که یکی از سه مقدار ۰، ۱ و ۲ را می‌گیرند؛ به طوری که هر چه برچسب مربوط بزرگ‌تر باشد، تطابق بین درخواست و سند بالاتر است.

ورودی: برای یک درخواست q تعداد n سند $D = \{d_1, d_2, ...d_n\}$ داریم که می‌خواهیم بر اساس برچسبِ مرتبط بودن رتبه‌بندی شوند. زوج مرتب‌های $x_i = (q, d_i)$ ورودی مدل خواهند بود.

خروجی: به ازای هر زوج مرتب $x_i = (q, d_i)$ یک برچسب امتیاز (مرتبط بودن) y_i وجود دارد که مدل آن را پیش‌بینی می‌کند و این برچسب خروجی مدل خواهد بود.

۲.۳ ماموریت دوم (روش **pair-wise** و دسته‌بندی)

در این روش برخلاف روش قبلی دو سند داریم. می‌خواهیم پیش‌بینی کنیم که کدام یک از این دو سند مرتبط‌تر است. یعنی در واقع مسئله‌ی ما تبدیل به یک مسئله‌ی دسته‌بندی دوتایی[۷] می‌شود و از روی این موضوع که امتیاز کدام سند بیش‌تر است خروجی صفر یا یک مشخص می‌شود. در این قسمت باید مسئله را به عنوان یک مسئله logistic regression فرمول بندی کنید. دقت کنید که در این قسمت ممکن است لازم باشد ابتدا داده را به فرمت مناسب تبدیل کنید. یعنی سه تایی‌های (q, d_i, d_j) را تولید کنید.

[5] Closed-Form

[6] Gradient Descent

[7] Binary

ورودی: ورودی مدل به صورت (q, d_i, d_j) است که منظور از q درخواست و d_i, d_j سند هستند.

خروجی: اگر امتیاز سند d_i بیش‌تر بود مدل باید یک و در غیر این صورت صفر را خروجی دهد.

برای اطلاعات بیش‌تر و دقیق‌تر در مورد روش‌های بالا اکیدا پیشنهاد می‌شود این لینک را مشاهده کنید.