به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس جستجو و بازیابی اطلاعات در وب

استاد ممتازی

تمرین سوم

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

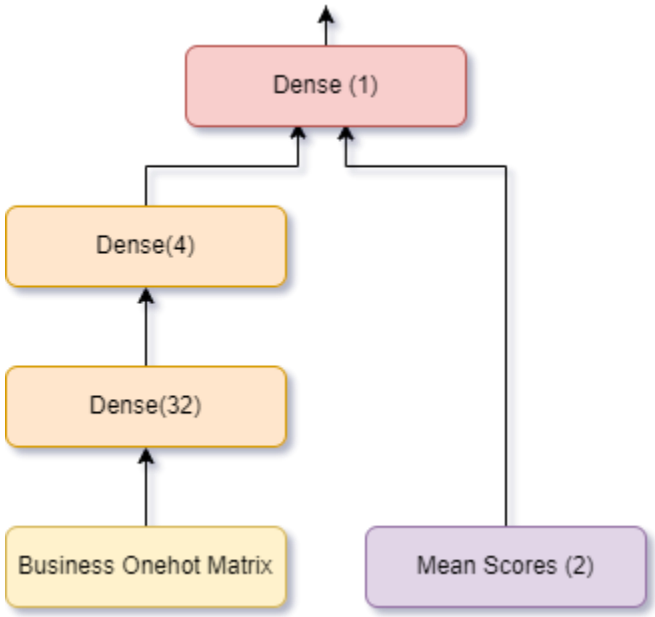
بخش دوم – سامانه توصیه‌گر مبتنی بر شبکه عصبی

مطبق با صورت تمرین برای این قسمت تنها مجاز به استفاده از امتیاز کاربر به رستوران هستیم. ما به شبکه‌ای نیاز داریم که تعبیه‌ای از یک کاربر و تعبیه از یک رستوران برای ما ایجاد کند.

مطابق با بررسی‌های انجام‌شده کاربران بخش آموزش، اعتبارسنجی و آزمون کاملا متفاوت هستند. بنابرین داشتن آیدی کاربر و تعبیه‌ای متناسب با آن بی‌معناست. در این شرایط تنها میانگین امتیازهای یک کاربر می‌تواند مفید باشد. در مورد رستوران به طور مشابه میانگین امتیازهای دریافتی آن یک ویژگی بسیار خوب است. به علاوه رستوران‌های سه بخش مجموعه‌داده دارای اشتراکاتی با یک دیگر هستند. بنابراین می‌توان برای هر رستوران یک بردار one-hot در نظر گرفت و آن را هم در شبکه دخیل کرد. بدین ترتیب برای رستوران‌های آموزش که در قسمت‌های دیگر وجود داشته باشد، اطلاعات مفیدی خواهیم داشت.

برای آموزش مدل از ده گام آموزش به همراه یک Early Stopping Callback بهره گرفتیم تا پیش از بیش‌برازش شدن آموزش مدل متوقف شود. شایان ذکر است که شبکه امتیاز یک کاربر به یک رستوران را پیش‌بینی می‌کند و حالت رگرسیونی دارد. بهینه‌ساز Adam، فعال‌سازهای ReLU و خطای MSE هم مورد استفاده بوده است.

در تصویر ۱ معماری شبکه پیشنهادی من برای این قسمت آورده شده است.

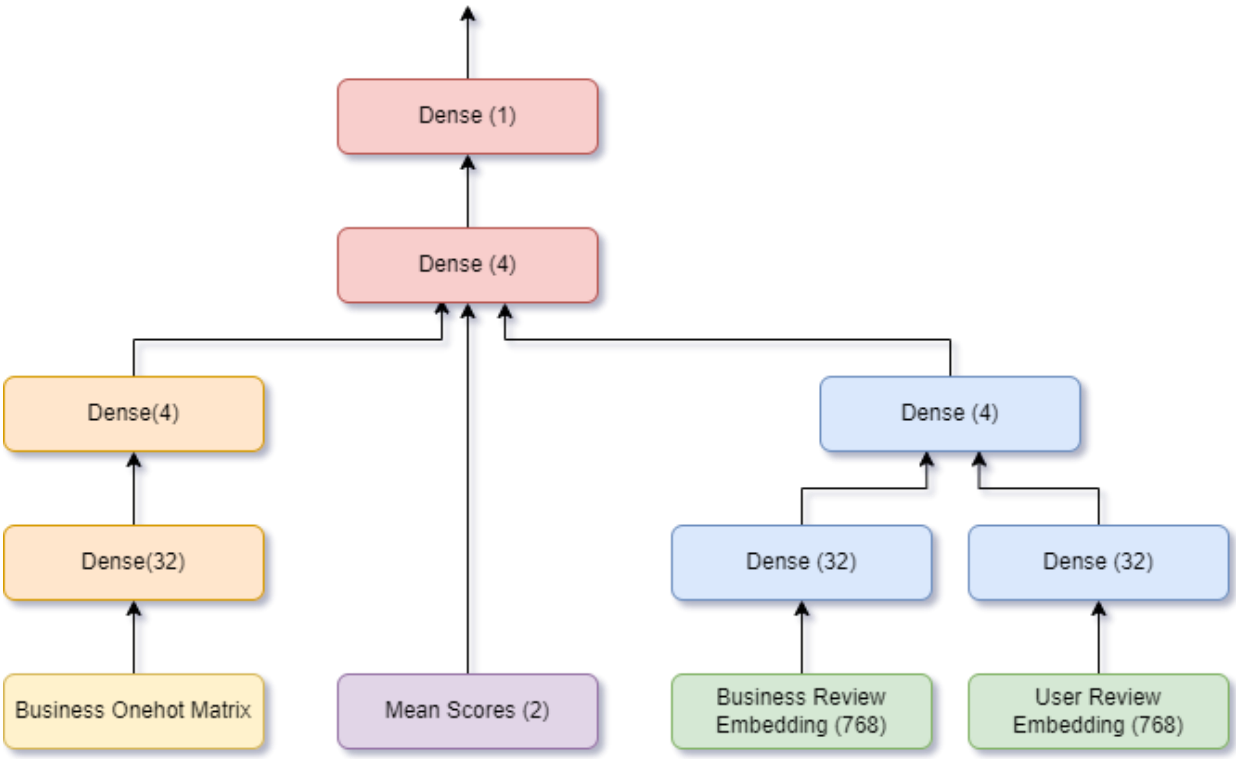


تصویر ۱ – معماری مدل مبتنی بر امتیاز

بخش سوم – سامانه توصیه‌گر با استفاده از اطلاعات جانبی

برای این بخش علاوه بر میانگین امتیازهای کاربر، میانگین امتیازهای رستوران و بردار تعبیه رستوان، می‌توان از دو دسته ویژگی‌های متنی استفاده کرد. به بیان دقیق‌تر می‌توان برای هر کاربران لیست نظرات کاربر به تمام رستوران‌ها به غیر از رستوان مورد آزمون و برای هر رستوان تمام نظرات کاربران به آن رستواران به غیر از کاربر مورد آزمون را در نظر گرفت. برای استفاده از این اطلاعات متنی از یک مدل BERT پیش آموزش‌یافته استفاده کردم. بنا به سادگی برای هر دسته نظر میانگین بردار آن‌ها را به عنوان بردار نظرات آن در نظر گرفتم.

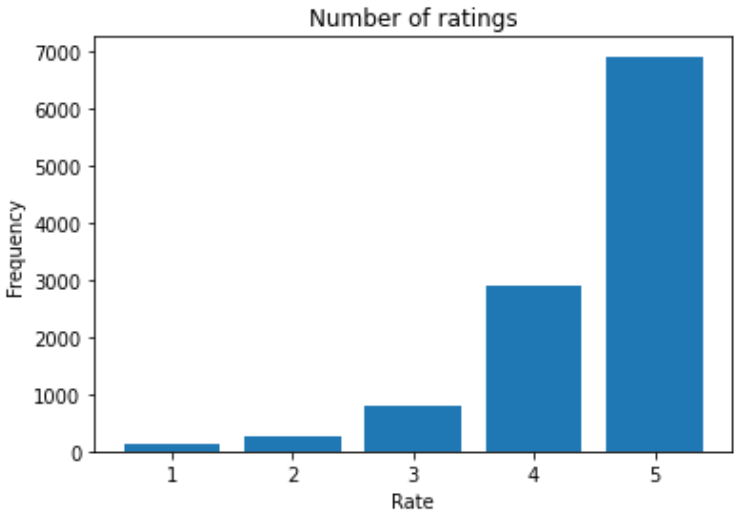
سایر نکات مربوط به آموزش مدل مشابه قسمت پیشین است. در تصویر ۲ معماری شبکه پیشنهادی من برای این قسمت آورده شده است.



تصویر۲ – معماری مدل مبتنی بر اطلاعات جانبی

بخش چهارم – کار با روش‌های ارزیابی و گزارش نتایج

برای ارزیابی مدل با استفاده از دو معیار P@n و R@n نیاز است که امتیاز کاربران به رستوران‌ها را به امتیازهای باینری تبدیل کنیم. برای این تبدیل نیاز به یک حدآستانه برای امتیاز است که بیشتر از آن را مرتبط و کمتر از آن را نامرتبط تلقی کنیم. در تصویر ۱ هیستوگرام امتیازهای کاربران بخش آزمون آورده شده است. بر طبق این نمودار به نظر می‌رسد که بهتر است تنها امتیاز ۵ را به عنوان کلاس مثبت در نظر بگیریم و سایر کلاس‌ها را به عنوان امتیاز منفی لحاظ کنیم.



تصویر ۳ – هیستوگرام فراوانی امتیاز کاربران

پیش از هر چیز باید دانست که برای یک n خاص تنها می‌توان کاربرانی را از مجموعه‌داده آزمون در نظر گرفت که بیشتر از n نظر داده باشند. چراکه اگر n و یا تعداد کمتری نظر ثبت کرده باشند، مدل همواره خروجی یکسانی خواهند داشت و ارزیابی صحیح نخواهد بود. سه مقدار ۳، ۴ و ۵ برای n لحاظ شده است. در جدول ۱ مشخص شده است که به ازای هر کدام از این سه عدد چه تعداد کاربر از تمام کاربرهای بخش آزمون باقی می‌ماند.

جدول ۱ – تعداد کاربران تست برای هر n

n	تعداد داده
۵	۲۶۵
۴	۴۴۲
۳	۷۵۳

دو معیار P@n و R@n به ازای یک کاربر خاص قابل تعریف است. برای آنکه این معیار به کل مجموعه‌داده آزمون تعمیم یابد، از میانگین این اعداد استفاده خواهد شد. به علاوه برای معیارهای P@n و R@n به صورت کلی ممکن است که به ۱۰۰٪ نتوان رسید. برای دید بهتر حد بالای این اعداد هم گزارش شده است. نهایتا مقدار MSE هم به

عنوان یک شاخص دقیق‌تر آورده شده است. در جدول ۲ شاخص‌های مختلف دقت مذکور برای دو مدل ارائه شده و حالت ایده‌آل ارائه شده است.

جدول ۲ – دقت مدل‌های مختلف به ازای معیارهای مختلف

	مدل مبتنی بر امتیاز	مدل اطلاعات جانبی	مدل ایده‌آل
MSE	۰.۳۴۶	۰.۲۷۲	۰
P@5	۶۴.۶۸٪	۶۶.۸۷٪	۷۷.۷۴٪
R@5	۷۰.۹۸٪	۷۳.۸۳٪	۸۵.۴۲٪
P@4	۶۴.۳۷٪	۶۸.۰۴٪	۷۸.۳۴٪
R@4	۶۶.۷۰٪	۷۱.۴۴٪	۸۱.۹۷٪
P@3	۶۶.۰۰٪	۷۰.۰۸٪	۷۹.۶۸٪
R@3	۶۳.۵۶٪	۶۸.۰۹٪	۷۷.۵۵٪

مطابق با نتایج می‌توان دید همواره استفاده از اطلاعات جانبی باعث بهبود دقت مدل شده است ولی فاصله تا مدل ایده‌آل کم نیست. به صورت حدودی می‌توان گفت استفاده از اطلاعات جانبی باعث می‌شود تا یک سوم فاصله تا بهترین مدل کاسته شود.