مباحث ویژه در سیستم‌های دیجیتال

نیم‌سال اول ۱۴۰۰-۱۴۰۱

مدرس: دکتر ایمان غلام‌پور

دانشکده مهندسی برق

توضیحات پروژه

نام و نام‌خانوادگی: امین کشیری

شماره دانشجویی: ۹۷۱۰۱۰۲۶

۱ مقدمه

در این پروژه، با استفاده از ایده‌ها و روش‌های مختلفی که در طول ترم آموختیم، اطلاعاتی را از داده‌ها بیرون کشیدم. بخش‌های مختلف این پروژه را در فایل‌های jupyter جداگانه قرار داده‌ام، و توضیحات هر بخش را نیز جداگانه در این فایل نوشته‌ام.

۱.۱ توضیحات کلی

۱. در ابتدای تمامی کدها تنظیمات اولیه اسپارک را انجام دادم، و سپس فایل csv داده شده را load کردم.

۲. در بعضی از بخش‌ها، روز ۸م را از داده‌ها حذف کردم. دلیل این کار این بود از تمامی روزها به اندازه‌ی متناسب با هم داده داشته باشیم. در غیر این صورت تعداد داده‌ها از روز سه شنبه دو برابر باقی روزها می‌شد. کارهای دیگری نیز می‌توانست انجام بگیرد. مثلا می‌شود داده‌های روز سه شنبه را میانگین بگیریم (یعنی در تمام قسمت‌هایی که تعداد متغییری را شمرده‌ایم، برای روز سه شنبه این تعداد را تقسیم بر ۲ کنیم). در بعضی از قسمت‌ها اما این زیادتر بودن داده‌های روز سه شنبه مشکلی ایجاد نمی‌کرد. اما دقت کنید که در بعضی از قسمت‌های دیگر می‌تواند تحلیل ما را دچار انحراف کند (مثلا ممکن است به اشتباه نتیجه بگیریم که روز سه شنبه روز پر تردد تری است، یا دوربین‌هایی که در روز سه شنبه دیده می‌شوند را به اشتباه مهم تر در نظر بگیریم).

۳. توضیحات کد و روند اجرا را در فایل‌های jupyter نوشته‌ام. سعی کرده‌ام که توضیحات منطق پشت کدها را در این مستند بنویسم (و نه در خود کدها). بنابرین توضیحات تکنیکال خود کد در اینجا کمتر نوشته شده است.

٢ General

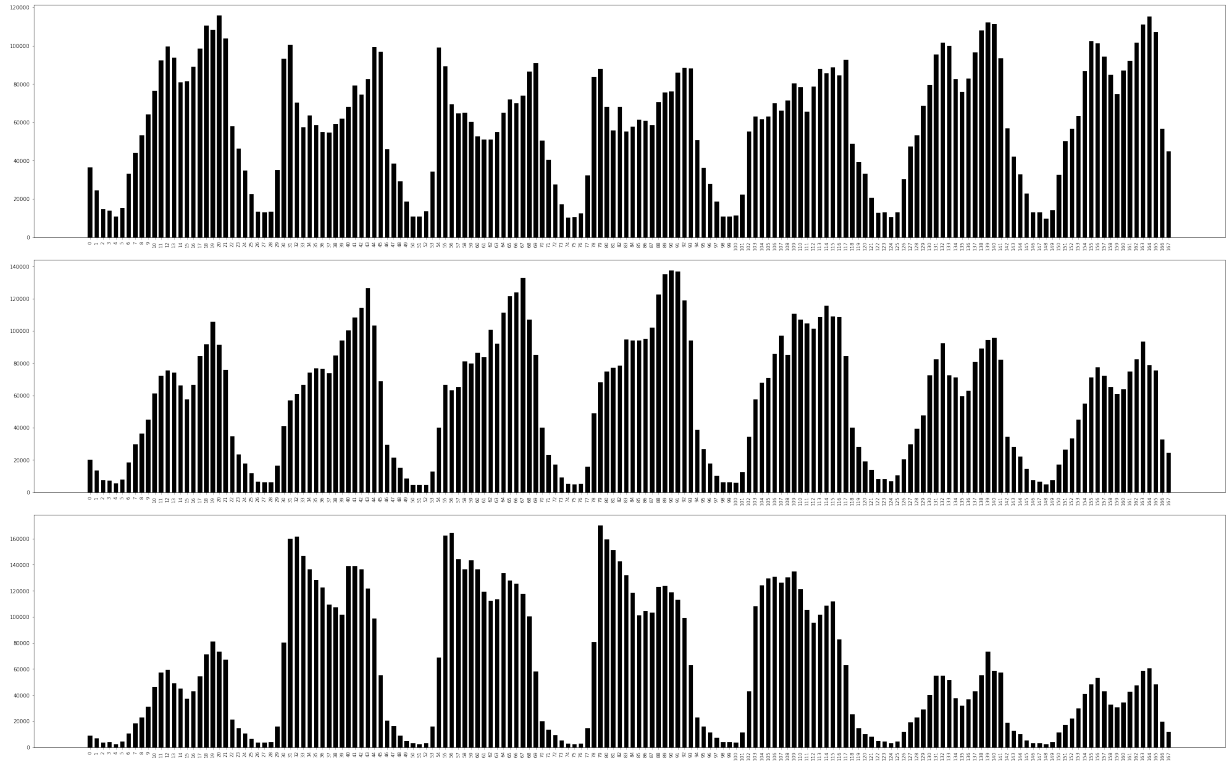
۳ Clustering

۱.۳ توضیحات کلی

با استفاده از خوشه سازی، سعی کردم دوربین‌ها را به دسته‌های مختلفی تقسیم کنم، و برای هر دسته مفهومی بیابم. نماینده هر دوربین در این روش، یک بردار با اندازه‌ی 7×24 است، که در هر خانه‌ی آن تعداد تردد در آن ساعت از روزهفته قرار گرفته‌است. ۲۴ ساعت اول برابر با یک‌شنبه است، ۲۴ ساعت بعدی برای دو شنبه و الی آخر. پس بردار متناظر هر دوربین، تعداد ترددها در هر ساعت از یک هفته را برای آن دوربین مشخص می‌کند.

برای خوشه‌سازی از الگوریتم LDA یا Latent Dirichlet Allocation استفاده شده است، که همانطوری که در درس دیدیم استفاده اولیه آن پیدا کردن توزیع topic های مختلف و کلمات آن‌ها برای هر مقاله است. با استفاده از این الگوریتم برای داده‌های تردد ماشین‌ها نیز می‌توانیم دقیقا به چنین توزیعی برسیم.

یکی از متغییرهای بسیار مهم در این بخش، cluster_center است، که تعداد کلاسترهای نهایی را مشخص می‌کند . با تغییر این متغییر می‌توانیم تعابیر متفاوتی از داده داشته باشیم. اما یکی از واضح‌ترین نتیجه‌ها برای $\text{cluster\_center} = 3$ به دست می‌آید که آن را در شکل‌های زیر می‌بینید:



۲.۳ تحلیل نتایج

مهم‌ترین نکته‌ی این سه تصویر، روند تغییر ترددها در هر روز است. در دسته‌ی اول، تردد در ساعات اولیه روز افزایش می‌یابد، در هنگام ظهر کاهش پیدا می‌کند، سپس دوباره در شب افزایش پیدا می‌کند. در دسته‌ی دوم، تردد در صبح کم است، اما کم کم افزایش می‌یابد و در شب به اوج خود می‌رسد. دسته‌ی سوم روندی دقیقا عکس دسته‌ی دوم دارد، و بیشترین تردد را در صبح دارند و سپس کاهش می‌یابد.

سه دسته‌ی بالا را می‌توانیم به این صورت تفسیر کنیم. دسته‌ی اول نقاط پر تردد شهر هستند، که هم در روز و هم در شب تردد بالایی دارند. این نقاط احتمال مکان‌هایی وسط شهر هستند که تمام طول روز تردد دارند (البته طبیعتا تردد در ظهر کاهش می‌یابد). دسته‌ی دوم، احتمالا مکان‌های دیدنی و تفریحی و یا بازارهای شبانه هستند که در طول روز تردد زیادی ندارند (به دلیل این که مردم مشغول کار و مدرسه و ... هستند). دسته‌ی سوم نیز احتمالا مکان‌هایی هستند که در طول روز تردد بالایی دارند، مانند مکان‌های اداری، مسیر مدارس و ادارات، یا دوربین‌های نزدیک به مثلا نانوایی‌ها.

یک نکته‌ی بسیار جالب دیگری که در این تصاویر دیده می‌شود، این است که که تردد در روزهای جمعه، شنبه و یک شنبه به طرز جالبی پایین است. با چک کردن این ۳ روز روی تقویم، فهمیدم که این روزها تعطیل رسمی بوده‌اند (قیام ۱۵ خرداد و شهادت امام جعفر صادق (ع)). بسیار جالب است که این کاهش ترددها، فقط در دسته‌ی سوم رخ داده‌است، که

دقیقا با شهود ما همخوانی دارد، که دسته‌ی سوم مکان‌هایی مانند مدارس و ادارات هستند. همچنین، الگوی کاهشی تردد در این سه روز از بین رفته است که باز هم مطابق با الگوی پیدا شده است.

Pixie ۴