دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین تابستان ۱۴۰۰

پروژه عملی - تحلیل احساسات (فاز دوم)

مدرس: دکتر محمّدحسین رهبان زمان تحویل: ۳۰ تیر

۱ مسئله واکاوی خوشه‌ها[1] (یادگیری بدون نظارت)

در این بخش با استفاده از روش‌های بدون نظارت به خوشه‌بندی می‌پردازیم و تلاش می‌کنیم ارتباط بین مستندهای مشابه را پیدا کنیم. با استفاده از روش‌های k-means ، Gaussian Mixture Model و یک روش دلخواه دیگر (بهتر است این روش به نتیجه بهتری نسبت به دو روش دیگر برسد) به خوشه‌بندی داده‌ها بپردازید. هر روش را به ازای چندین حالت (دست‌کم دو حالت) مختلف برای تعداد خوشه‌ها آموزش دهید. ورودی هریک از این روش‌ها را بردارهای بدست آمده در فاز اول یا ویژگی[2]های لایه آخر MLP (که ممکن است در فاز اول پیاده‌سازی کرده باشید) در نظر بگیرید. مواردی که در این بخش باید بررسی کنید به شرح زیرند:

آ) به ازای هریک از روش‌های خوشه‌بندی و تعداد خوشه‌ها، یک روش دلخواه کاهش بعد مانند PCA را استفاده کرده و داده‌ها را طوری در فضای دو بعدی نمایش دهید که به‌راحتی بتوان داده‌های خوشه‌های مختلف را مشاهده کرد.

ب) روش‌های خوشه‌بندی را به ازای دو خوشه اجرا کرده و خوشه‌های بدست آمده را با استفاده از چند معیار دلخواه با برچسب مستندها مقایسه کنید. نتیجه‌های بدست آمده را تحلیل کنید.

پ) به ازای یکی از روش‌های خوشه‌بندی و یکی از حالت‌های تعداد خوشه‌ها (که مخالف ۲ است)، مستندهایی که در خوشه یکسانی قرار می‌گیرند را مقایسه کنید. آیا شباهتی بین مستندهای یک خوشه وجود دارد؟ در این مورد نیازی به استفاده از روش‌های خودکار نیست و کافیست تحلیل کلی درستی بر مبنای مشاهده‌هایتان بیان کنید.

۲ تنظیم دقیق پارامترها

در بسیاری از مسئله‌های دنیای واقعی، دسترسی ما به داده‌های دارای برچسب و قابل تحلیل کم است. به طور کلی گردآوری مجموعه دادگانی که به طور مستقیم در آموزش مدل‌ها به کار روند هزینه زیادی دارد. به همین خاطر در خیلی از مسئله‌ها از شبکه‌هایی استفاده

[1] Cluster Analysis

[2] Feature

می‌شود که با یک مجموعه داده مشابه برای مسئله‌ای دیگر به کار رفته‌اند؛ به این شکل که ابتدا مدل روی یک مجموعه داده غنی (که بیشتر نمونه‌های آن دارای برچسب هستند) آموزش داده می‌شود و سپس از این مدل آموزش داده شده برای آموزش دوباره روی مجموعه دادگان جدید استفاده می‌شود. به این فرایند، تنظیم دقیق[1] گفته می‌شود.

تا این مرحله مجموعه دادگان مورد استفاده در پروژه شامل ۴۵۰۰۰ نظر بوده است. در کنار این داده‌ها، مجموعه دادگان بسیار کوچک‌تری در دسترس است که از نظرهای کاربران یک سامانه فروشگاهی برخط[2] بدست آمده. مجموعه دادگان اولیه که مربوط به نظرهای کاربران در یک شبکه اجتماعی دیگر بود، دارای موضوع‌های متفاوتی با مجموعه دادگان اخیر است. این مجموعه دادگان جدید را می‌توانید از اینجا دریافت کنید. این مجموعه تنها شامل ۵۰۰ نظر دارای برچسب از نظرهای کاربران در این فروشگاه برخط است.

آ) با استفاده از الگوریتم MLP، مدل‌سازی روی مجموعه دادگان را با پیش‌پردازش‌ها و روش‌های استخراج ویژگی مناسب انجام دهید. بهترین دقتی را که روی یک مجموعه اعتبارسنجی[3] بدست می‌آورید، گزارش کنید.

ب) حال بهترین مدلی که در فاز اول بدست آورید را دوباره روی مجموعه دادگان جدید آموزش دهید. نتیجه‌های بدست آمده را با قسمت قبل مقایسه کنید.

نتیجه: نتیجه‌های بدست آمده در هر دو بخش **آ** و **ب** را همانند فاز اول با استفاده از تابع analysis(. , .) گزارش دهید.

پاینده باشید

[1] Fine-Tuning
[2] Online
[3] Validation