باسمه تعالی

دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین
نیم‌سال اول سال ۹۵-۱۳۹۴

پروژه دوم **شجاعی، اوصیا**

نکات قابل توجه

- پاسخ‌های خود را به آدرس ml941.sharif@gmail.com ارسال کنید.
  نام فایل ارسالی تان به صورت STDNUM-P2.zip باشد.
- در انتخاب محیط و کتاب‌خانه‌ها برای پیاده‌سازی محدودیتی وجود ندارد.
- مشاهده تقلب در پروژه‌ی ارسالی پیامدهای جالبی نخواهد داشت.

۱ مقدمه

در فاز اول پروژه تعدادی از روش‌های یادگیری نظارتی را دیدیم. در این فاز قرار است تعدادی از روش‌های غیرنظارتی را بینید و مقایسه‌ای روی آن‌ها داشته باشید. بدین منظور دو مساله‌ی کاهش بعد و خوشه‌بندی در نظر گرفته شده است. همان‌طور که می‌دانید در مساله‌ی خوشه‌بندی داده‌ها برچسب نخورده‌اند و هدف ما این است که این برچسب‌ها را یاد بگیریم. تاکنون با تعدادی روش احتمالی و غیر احتمالی آشنا شده‌اید که باید از آن‌ها در این بخش استفاده کنید. همچنین در بسیاری موارد خوب است که با کاهش بود سعی کنیم نویز و همبستگی بین ابعاد را پایین بیاوریم تا به نتیجه‌ی بهتری برسیم. در قسمت اول مساله‌ی خوشه‌بندی مطرح می‌شود و روش‌های kmeans و مدل مخلوط با هم مقایسه می‌شوند. در قسمت دوم مساله‌ی کاهش بعد در دسته‌بندی چند کلاسه مطرح است. چالشی که در این‌جا وجود دارد این است که آیا روی داده‌های هر کلاس به طور مستقل کاهش بعد را داشته باشیم یا اینکه از راه دیگری استفاده کنیم. بدین منظور سه روش کاهش بعد روی داده‌ها مطرح شده است و می‌خواهیم مقایسه‌ای روی آن‌ها داشته باشیم.

۲ خوشه‌بندی

در این بخش زیرمجموعه‌ای از مجموعه دادگان MNIST بدون برچسب را در اختیار دارید و قرار است روی تصاویر آن خوشه‌بندی انجام دهید. این زیرمجموعه شامل ارقام صفر تا پنج است یعنی از ارقامی که شباهت زیادی با هم دارند تنها یکی‌شان در این مجموعه داده وجود دارد. در قسمت اول ابتدا با گذاشتن آستانه روی شدت روشنایی هر نقطه، تصاویر را دودویی کنید. ابتدا از روش kmeans برای خوشه‌بندی استفاده کنید. سپس فرض کنید نقاط تصاویر از یک مدل مخلوط

برنولی [۱] تولید شده اند.[۲] روش EM را خودتان پیاده‌سازی کنید و خوشه‌بندی را انجام دهید. حال آستانه را بردارید و تصاویر اولیه را در نظر بگیرید. این بار فرض کنید نقاط تصاویر از یک مدل مخلوط چندجمله [۳] تولید شده‌اند.[۴] روابط EM را بنویسید و خوشه‌بندی را انجام دهید.

پس از اتمام کار، بعد داده‌ها را با استفاده از PCA کاهش دهید و دوباره سه روش یاد شده را تست کنید و اثر کاهش بعد را گزارش کنید. کلیه‌ی روش‌های پیاده شده را با هم مقایسه کنید.

روش ارزیابی

برای مقایسه‌ی روش‌های که پیاده کرده‌اید برچسب داده‌های آموزش به شما داده می‌شود. دقت کنید که شما این برچسب‌ها را نمی‌توانید در هنگام خوشه‌بندی استفاده کنید و باید فرض کنید آن‌ها را در اختیار ندارید. از معیارهای RandIndex، Macro-F۱ و Micro-F۱ برای ارزیابی استفاده کنید.

۳ کاهش ابعاد

شرح مسئله

یک مسئله‌ی دسته‌بندی با چند کلاس را در نظر بگیرید که بعد داده‌های آن بسیار زیاد است و قبل از اینکه یک الگوریتم دسته‌بندی را بر آن‌ها اعمال کنیم، باید به طریقی بعد آن‌ها را کم کنیم. برای مثال فرض کنید ۱۰ کلاس وجود دارد و می‌خواهیم داده‌ها را به ۱۰۰ بعد کاهش دهیم. شاید اولین گزینه برای این کار PCA باشد. پس راه اول این این است که روی کل داده‌ها PCA انجام دهیم. ولی در این حالت از برچسب داده‌ها استفاده‌ای نشده است. یک پیشنهاد دیگر این است که این ۱۰۰ بعد این گونه انتخاب شوند که PCA با $k = ۱۰$ روی داده‌ها هر کدام از کلاس‌ها اعمال شوند تا ویژگی‌هایی که واریانس داده‌های هر کلاس در آن راستا بیشینه است، استخراج شوند. در نهایت این پایه‌های ۱۰تایی برای هر کلاس را تجمیع می‌کنیم و یک داده‌ی تست جدید را روی فضایی ۱۰۰ بعدی حاصل تصویر می‌کنیم.

ایرادی که ممکن است به راه‌حل بالا گرفته شود این است مسئله برای هر کلاس بدون توجه به پایه‌هایی که برای کلاس‌های دیگر پیدا شده حل می‌شود. چه بسا بخاطر شباهت ذاتی داده‌ها بعضی از پایه‌هایی که واریانس را بیشینه می‌کند بین کلاس‌ها مشابه یا حتی مشترک باشد [۵]، در این حالت تعدادی از بعدهایی که استفاده می‌کنیم در حقیقت «به هدر می‌روند» و مهمتر از آن، ویژگی‌های پیدا شده اصلا جداکننده [۶] نیستند.

تحت تاثیر این انتقاد روش سومی پیشنهاد می‌شود. برای شرح این روش فرض کنید که داده‌ها با x و برچسب‌ها با y نشان داده می‌شوند. همچنین $C_i = \mathbb{E}[xx^T|y = i]$ ماتریس گشتاور دوم داده‌های کلاس iم را نشان می‌دهد. در این

[۱] Bernouli Mixture Model

[۲] 9.3.3 from Bishop

[۳] multinomial Mixture Model

[۴] Example 9.19 from Bishop

[۵] این مسئله می‌تواند مانع استقلال خطی ۱۰۰ بردار پیدا شده بعد از تجمیع مولفه‌های اصلی هر کلاس شود. شما هنگام پیاده‌سازی باید تدبیری برای این حالت اندیشیده باشید.

[۶] discriminative

روش بجای روش دوم که به دنبال یافتن پایه‌هایی مثل v بودیم که $v^T C_i v$ را بیشینه کند، دنبال بیشنه کردن نسبت این کمیت برای دو کلاس هستیم یعنی

$$R_{ij}(v) = \frac{\mathbb{E}[(v^T x)^2 | y = i]}{\mathbb{E}[(v^T x)^2 | y = j]} = \frac{v^T C_i v}{v^T C_j v}$$

نشان‌دهید vهایی که بیشینه محلی این تابع هستند،بردار ویژه‌های تعمیم یافته خواهند بود. یعنی vهایی که $C_i v = \lambda C_j v$. دقت کنید که در روش دوم اگر ۱۰ کلاس داشته باشیم، ۱۰ سری مولفه اصلی [۷] پیدا می‌شود که از هر سری همه یا تعدادی که اصلی‌تر هستند انتخاب می‌شوند. اما در روش سوم به ازای هر i و j یک سری جواب خواهیم داشت یعنی $\binom{10}{2}$ مجموعه از v که از هر کدام تعدادی را انتخاب می‌کنیم تا فضای کاهش بعد را تشکیل دهند.

در ادامه فرض کنید بردار ویژه‌ها طوری نرمال شده اند که $v^T C_j v = 1$ و $v^T C_i v = \lambda$. همچنین یک جمله منظم‌سازی به C_j اضافه می‌کنیم تا ماتریس در مخرج همیشه از مرتبه کامل باشد؛ به این ترتیب که

$$R_{ij}^{\beta} = \frac{v^T C_i v}{v^T (C_j + \beta I) v}$$

پیاده‌سازی

شما باید سه روش بالا را پیاده‌سازی کنید تا روشی که نتیجه‌ی بهتری می‌دهد پیدا شود. نتیجه بهتر یعنی خطای کمتر در دسته‌بندی چند کلاسه. برای مقایسه، روی خروجی هر روش یک دسته‌بند svm با هسته گاوسی واریانس ۱۰ و هزینه ۱ یاد بگیرید. (اگر از libsvm استفاده می‌کنید پارامترهای بیان شده معادل svm-train -c 1 -t 2 -g 0.1 خواهد بود)

تعیین تعداد ابعادی که در کاهش بعد استفاده می‌کنید به عهده‌ی شماست. همچنین در پیاده‌سازی روش سوم یک آستانه α در نظر بگیرید و برداروِیژه‌هایی که مقدار ویژه متناظر آنها کوچکتر از آستانه است یعنی $\lambda < \alpha$ را دور بریزید. مقدار این آستانه و مقدار ضریب منظم‌سازی یعنی β باید با cross validation تعیین شوند. فرض کنید β ضریبی از میانگین مقدارویژه‌های C_j است و در cross validation مقدار آن ضریب را تعیین کنید.

علاوه بر این، روش سوم را روی داده‌های MNIST اجرا کرده و چهار بردار اول برای بهینه‌سازی‌های $R_{۰.۸}$ ، $R_{۶۸}$، $R_{۱۷}$ و $R_{۴۹}$ را به صورت تصویر نشان دهید. در مورد کیفیت و توانایی جداسازی این بردارها برای کلاس‌های مربوطه‌شان بحث کنید.

مجموعه داده

داده‌ای که در این قسمت استفاده می‌کنید، زیر مجموعه‌ای از 20 news groups است. این مجموعه داده متشکل از متون ۲۰۰۰۰ خبر است که تقریبا به طور مساوی زیر عنوان یکی از ۲۰ موضوع خبری است. در این فاز شما با یک زیرمجموعه از این داده‌ها کار خواهید که داده‌های مربوط به ۵ موضوع از ۲۰ موضوع کل داده‌هاست. هم‌چنین متن‌ها پیش‌پردازش شده‌اند و ویژگی tf-idf از آن‌ها استخراج شده اند. در نهایت ۲۹۱۹ بردار ویژگی ۴۶۳۹۷ بعدی به همراه برچسب هر کدام تحویل شما داده شده است.

[۷] Principal Component

۴ مستندات

مستند ارسالی حداقل باید شامل موارد زیر باشد:

- توضیحاتی در مورد نحوه‌ی پیاده‌سازی هر قسمت و ابزارهای مورد استفاده.
- جداول نتایج به دست آمده در هر قسمت. به همراه مقادیر استفاده شده در cross vlaidation و جواب گرفته شده برای هر کدام.
- محاسبات یا اثبات‌های لازم برای بدست آوردن روابط مورد استفاده.

موفق باشید.