باسمه تعالی

جبرخطی کاربردی – تکلیف سری سوم

مهلت تحویل: سه‌شنبه ۳۰ مهر ۱۳۹۸

مجموعه مسائل این تکلیف در مورد آلگوریتم خوشه‌بندی k-means است.

برای اجرا و بررسی نتایج آلگوریتم، از یک سری داده تصویری مربوط به ارقام فارسی (از مجموعه ارقام دستنویس هدی) استفاده می‌شود. در شکل زیر نمونه‌ای از این تصاویر را ملاحظه می‌کنید.

هر تصویر یک ماتریس ۴۰ در ۳۰ باینری (صفر و یک) است. تصاویر به صورت سیاه و سفید هستند (صفر برای سیاه و ۱ برای سفید، یا برعکس). برای اینکه بتوان از آلگوریتم k-mean استفاده کرد، هر تصویر با قرار گرفتن پشت سر هم سطرهای آن به یک بردار سطری ۱ در ۱۲۰۰ تبدیل شده است (از عنصر ۱ تا ۳۰ در سطر اول، ۳۱ تا ۶۰ در سطر دوم و ...). داده‌های مربوط به ۲۰،۰۰۰ تصویر در قالب یک فایل متنی با نام TrainData.txt در اختیار شما قرار گرفته است. این فایل شامل ۲۰،۰۰۰ سطر است که هر سطر داده‌های یک تصویر را در بر دارد. داده‌ها با ویرگول از هم جدا شده‌اند.

برای شروع ابتدا سعی کنید فایل را در محیط برنامه‌نویسی خود فراخوانی کرده و داده‌ها را از آن استخراج کنید. سپس برای اینکه بتوانید داده‌ها و نتایج اجرای آلگوریتم را به صورت تصویری ببینید، برنامه کوتاهی بنویسید که هر بردار سطری ۱ در ۱۲۰۰ مربوط به یک تصویر را به ماتریس ۴۰ در ۳۰ آن تصویر تبدیل کرده و آن را نشان دهد. اکنون می‌توانید مراحل زیر را با استفاده از برنامه k-means که برای تکلیف اول نوشته‌اید بر روی این داده‌ها اجرا کنید.

۱- با فرض k=10 آلگوریتم k-means را روی داده‌ها پیاده‌سازی کنید. در هر مرحله بردارهای شاخص اولیه و بردارهای شاخص نهایی (بهینه) را در گزارش نشان دهید. دقت کنید که نمایش بردارهای شاخص به صورت تصویر (و نه بردار یا ماتریس) است. همچنین مقدار عددی تابع هزینه بهینه را ثبت کنید. علاوه بر این شاخص‌های نهایی را برای استفاده بعدی ذخیره کنید. برای اجرای آلگوریتم نیاز به یک دسته شاخص اولیه دارید. شاخص اولیه را با روش‌های زیر محاسبه کرده و برای هر شاخص اولیه نتایج را گزارش کنید:

a. ۱۰ بردار ۱ در ۱۲۰۰ تصادفی از صفر و یک تولید و به عنوان شاخص‌های اولیه استفاده کنید.
b. از داده‌ها ۱۰ بردار را به صورت تصادفی انتخاب کرده و به عنوان شاخص‌های اولیه استفاده کنید.
c. کل داده‌ها را به ۱۰ قسمت مساوی تقسیم کنید (مثلا ۲۰۰۰ داده اول یک قسمت، ۲۰۰۰ داده بعدی قسمت دوم و ...). سپس میانگین هر قسمت را محاسبه کرده و به عنوان شاخص‌های اولیه استفاده کنید.

d. تعدادی از تصاویر اولیه را ببینید. سپس برای هر رقم یک تصویر که به نظرتان مناسب‌تر است را انتخاب کنید. از این تصاویر به عنوان شاخص‌های اولیه استفاده کنید.

۲- قسمت (۱) را به ازای k=20 و برای شاخص‌های اولیه a، b و c تکرار کنید.

۳- فرض کنید k=13. برای شاخص‌های اولیه مانند مرحله d تصاویر را ببینید. برای رقم‌های ۴، ۵ و ۶ دو تصویر (یک تصویر برای هر شکل نوشتن رقم) و برای سایر ارقام یک تصویر مناسب انتخاب کنید. سپس آلگوریتم را برای این شاخص‌های اولیه اجرا و نتایج را گزارش کنید.

۴- فایل TestData.txtحاوی داده مربوط به ۱۳ تصویر است. برای نتایج به دست آمده از هر یک از خوشه‌بندی‌های قبلی (۴ خوشه‌بندی مربوط به قسمت (۱)، ۳ خوشه بندی مربوط به قسمت (۲) و یک خوشه‌بندی مربوط به قسمت (۳))، هر یک از این ۱۳ تصویر را در خوشه مناسب دسته بندی کنید. نتایج را با توجه به رقم مربوط به هر تصویر در یک جدول ثبت کرده و آن را تحلیل کنید.

تحویل تکلیف: برای تحویل تکلیف، لطفا فایل‌های زیر را در یک پوشه قرار دهید. سپس پوشه را فشرده کرده و با نام خود ذخیره کنید. این فایل فشرده را برای من با ایمیل یا تلگرام بفرستید.

۱- برنامه اصلی آلگوریتم k-means. این برنامه به صورت تابع با سه ورودی (داده‌ها، تعداد خوشه‌ها و شاخص‌های اولیه) و سه خروجی (شاخص‌های بهینه، بردار c و مقدار بهینه تابع هزینه) نوشته می‌شود. برای نوشتن برنامه می‌توانید به دلخواه از هر یک از سه زبان متلب، پایتون و R استفاده کنید.

۲- برنامه یا دستور مربوط به هر یک از قسمت‌های فوق. در صورتی که برای هر یک از این قسمت‌ها یا ترکیبی از آنها برنامه مجزایی می‌نویسید، لطفا برنامه را با نام مناسب ذخیره و ارسال کنید. اگر از مجموعه‌ای از دستورها استفاده می‌کنید آنها را در گزارش خود ذکر کنید.

۳- گزارش تایپ شده با فرمت pdf شامل نتایج انجام قسمت‌های فوق و تحلیل آنها.

موفق باشید - ایزدی