تمرین سری سوم درس یادگیری ماشین

تاریخ ارسال: ۲۷ دی ۹۶

تاریخ تحویل: ۵ بهمن ۹۶

نمره: ۴ نمره از ۲۰ نمره پایان ترم

فولدری به نام ML_Assignment03_name1_name2 بسازید که name1 و name2 نام افراد گروه هستند. در این فولدر دو فولدر به نام های ML_Problem31_name1_name2 و ML_Problem22_name1_ بسازید و پاسخ دو مسئله را به ترتیب در این فولدرها قرار دهید. دقت کنید در تمام فایل‌های کد و گزارش ها اسامی افراد تیم بصورت کامنت در بالای پاسخ‌ها عنوان شود. در نهایت فولدر اصلی را فشرده کنید و به آدرس taherian.khu@gmail.com ارسال کنید. فراموش نکنید که عبارت ML_Assignment03_name1_name2 (با جایگزینی نام افراد گروه) را در subject ایمیل بیاورید.

سوال ۱: در دسته بندی متن هدف این است که موضوع یک متن (خبر، مقاله، وبلاگ، ...) مشخص شود. سه فایل training, test, topics ضمیمه شده اند. در هر یک از دو فایل training, test متن‌هایی به فرمت زیر قرار دارند: (خط اول موضوع، سپس یک خط خالی، عنوان، خط خالی، محل و تاریخ، خط خالی، متن اصلی)

topic (classification)
blank line
title
blank line
location, date
blank line
text

موضوعات مختلفی که متن‌ها می توانند داشته باشند در فایل topics قرار دارد. هدف این است که موضوع هر یک از متن های داخل فایل test را با استفاده از الگوریتم k-nearest neighbor پیش بینی کنید. از توابع فاصله یا شباهت زیر و نمایش‌های نظیر آن‌ها استفاده کنید.

(a) فاصله Hamming: هر متن را با یک بردار دودویی نمایش دهید که هر بیت نشان دهنده این است که آیا کلمه نظیر در متن ظاهر شده یا خیر.

(b) فاصله اقلیدسی: هر متن با یک بردار عددی نمایش داده می‌شود که هر عدد نشان دهنده این است که کلمه نظیر چند بار در متن ظاهر شده (می تواند صفر باشد).

(c) شباهت کسینوسی: بردارهای عددی TF-IDF به روشی که توضیح داده می‌شود برای متن‌ها ساخته می‌شوند. سپس شباهت بین دو بردار بصورت کسینوس زاویه بین آن‌ها (ضرب داخلی دو بردار تقسیم بر حاصلضرب نرم آن‌ها) تعریف می‌شود.

فرض کنید w یک کلمه و d یک متن باشد؛ N(d,w) تعداد رخداد کلمه w در متن d باشد؛ W(d) تعداد کل کلمات متن d باشد؛ TF مخفف term frequency است و عبارتست از TF = N(d,w)/W(d).

فرض کنید D تعداد کل متن‌ها باشد و C(w) تعداد متن‌هایی باشد که شامل کلمه w هستند. IDF مخفف Inverted document frequency است و عبارتست از IDF(d,w) = log(D/C(w)) . وزن TF-IDF برای کلمه w در متن d بصورت مقابل تعریف می‌شود: TF(d,w)*IDF(d,w). این عددی است که در موقعیت هر کلمه w در بردار نظیر متن d باید بگذارید. (این یک معیار هیوریستیک است که سعی دارد محتوای اطلاعات هر کلمه را مشخص کند. بنابراین کلمه ای مثل the که در تمام متن ها ظاهر می‌شود IDF برابر با صفر دارد که نرخ زیاد حضور آن در متن را خنثی می‌کند. از طرف دیگر کلماتی که خاص همان متن هستند، تقویت می‌شوند.

برای هر یک از سه نمایش بالا k=1, 3, 5 را امتحان کنید. دقت کنید نزدیک ترین همسایه کسی است که فاصله اش کمتر و یا شباهتش بیشتر باشد.

گزارش خود از نتایج تمرین را با جداول و نمودارهایی نمایش دهید.

سوال ۲: دو فایل پیوست شده‌اند. یکی digit.txt که ۱۰۰۰ رکورد داده در این فایل وجود دارد. هر رکورد اطلاعات مربوط به ۱۵۷ پیکسل یک تصویر سیاه و سفید از یک رقم دست نوشته را در بر دارد. فایل دوم labels.txt، برچسب درست رکوردها را در بر دارد.

مجموع مربعات خطاهای درون خوشه ای:

هدف خوشه بندی می تواند می نیمم کردن پراکندگی درون خوشه ها باشد. فرض کنید C_k خوشه k اُم باشد و μ_k مرکز این خوشه باشد. در این صورت خطای درون خوشه k اُم بصورت زیر تعریف می شود:

$$SS(k) = \sum_{x_i \in C_k} | x_i - \mu_{C_k} |^2$$

توجه کنید که $|x_i - \mu_{C_k}| = \sqrt{\sum_{j=1}^{d} (x_{ij} - \mu_{C_{kj}})^2}$ همان فاصله اقلیدسی است که d بُعد را نشان می دهد.اگر K خوشه وجود داشته باشد خطای درون خوشه ای کل خوشه بندی مجموع تمام SS(k) ها روی تمام خوشه ها خواهد بود.

نرخ خطا:

از آن جا که برچسب واقعی تصاویر را داریم، با استفاده از آن ها می توانیم درستی خوشه بندی را تحلیل کنیم. برچسب نظیر خوشه C_k را برابر با رأی اکثریت رکوردهای آن خوشه قرار دهید. در صورت تساوی آرای دو برچسب، رقم کوچکتر را به آن خوشه نظیر کنید. برای مثال اگر خوشه ای با ۲۷۱ عضو دارید که ۱۰۰ عضو برچسب ۳، ۱۰۰ عضو برچسب ۸، ۵۰ عضو برچسب ۲و ۱۲ عضو برچسب ۹ و ۹ عضو برچسب ۰ دارند، برچسب ۳ را به خوشه نظیر کنید. در این حالت تعداد خطاهای این خوشه ۱۷۱ = ۱۰۰ – ۲۷۱ خواهد بود.

برای بدست آوردن نرخ خطا، تعداد خطاهای کل خوشه ها را با هم جمع کنید و بر تعداد کل رکوردها تقسیم کنید.

الف- الگوریتم را با k=2, k = 4, k=6 اجرا کنید. در هر اجرا، k رکورد اول موجود در فایل را بعنوان مراکز اولیه خوشه ها در نظر بگیرد. شرط توقف الگوریتم، عدم تغییر مراکز در دو تکرار متوالی و یا تعداد ۲۰ تکرار هست (یعنی اگر بعد از ۲۰ بار بروز کردن کلیه مراکز، هنوز همگرا نشده بود، الگوریتم را متوقف می کنیم.) تعداد تکرارهای الگوریتم، مجموع خطاهای درون خوشه ای کل خوشه ها و هم چنین نرخ خطای مربوط به اجراها را در یک جدول ارائه دهید.
ب- الگوریتم را برای k = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 اجرا کنید. دراین سوال در هر اجرا مراکز اولیه خوشه ها را به تصادف از بین رکوردها انتخاب کنید. نمودار مجموع خطاهای درون خوشه ای کل خوشه ها بر حسب k را رسم کنید. هم چنین نمودار نرخ خطا بر حسب k را رسم کنید.
ج- قسمت ب را تکرار کنید. با این تفاوت که برای هر k، الگوریتم را ۱۰ بار با مراکز اولیه تصادفی متفاوت اجرا کنید و بهترین جواب را نگه دارید. سپس نمودارها را برای این بهترین جواب ها رسم کنید.