گزارش تمرین پیاده سازی دوم درس هوش محاسباتی

بخش اجباری (FCM.py)

شرح مختصری راجع به نحوه ی پیاده سازی الگوریتم و برخی موارد جزئی تر مانند تعیین شرط خاتمه و یا مقداردهی های اولیه، ارائه دهید.

پس از آنکه فایل سمپل تست مورد نظر از کاربر دریافت شد، برنامه ابتدا نقاط را پلات می‌کنیم (اگر داده‌ها دوبعدی باشند) و سپس به ازای c=2 تا c=10، الگوریتم را اجرا می‌کنیم.

روند اجرای الگوریتم بدین صورت بوده است که ابتدا مراکز خوشه‌ها (V_i ها) را به صورت رندوم جنریت می‌کنیم.

سپس در یک حلقه، توابع تعلق (u_{ik} ها) را حساب کرده و مجدد مراکز خوشه‌ها را از روی توابع تعلق حساب می‌کنیم.

در نهایت و پس از خاتمه حلقه (حلقه 10 دفعه اجرا می‌شود. این عدد تجربی است و برای دسته‌بندی داده‌ها مناسب بوده است)، آرایه توابع تعلق مربوط به آن c را به یک آرایه به نام $c\_memberships$ (آرایهِ تمام آرایه‌های توابع تعلق) می‌افزاییم.

در ادامه، سراغ تعیین بهترین c می‌رویم. از فرمول زیر برای محاسبه آنتروپی هر c به خصوص استفاده می‌کنیم:

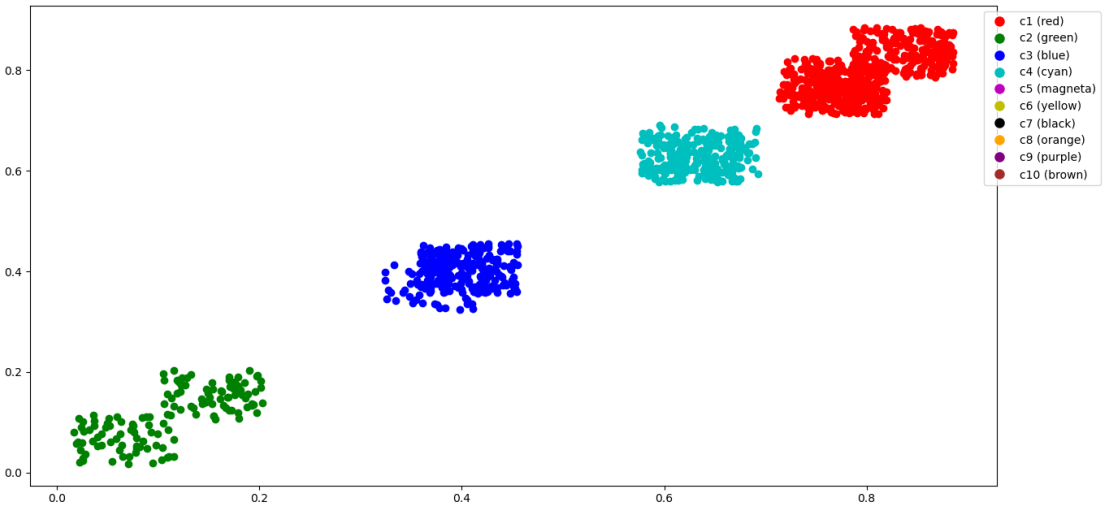
$$Entropy_c = -\sum_i \sum_k \frac{u_{ik} \ln(u_{ik})}{\ln(c)}$$

پس از اینکه بهترین c تعیین شد، نقاط را دسته بندی می‌کنیم. دسته بندی بدین صورت است که هر داده به خوشه‌ای نگاشت می‌شود که تابع تعلق آن ماکزیمم باشد. نهایتا، در کیس داده‌های دوبعدی، این داده‌ها رنگ خوشه مورد نظر خود را می‌گیرند و پلات می‌شوند.

پارامتر m را چه مقداری در نظر گرفتید ؟ مقدار این پارامتر چه تأثیری بر نتایج دارد؟

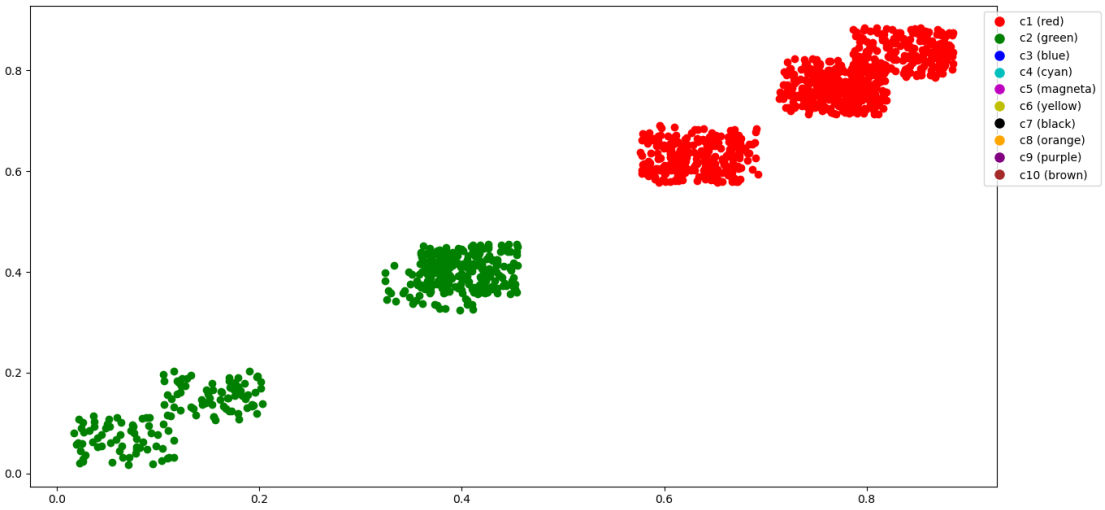
مقدار 3 در نظر گرفته شده است. این مقدار به صورت تجربی به دست آمده است. با افزایش m، شاهد کم شدن تعداد خوشه ها می‌شویم. به عنوان مثال:

به ازای $m = 3$ داریم:



sample2. m = 3. c = 4

به ازای $m = 4$ داریم:



sample2. m = 4. c = 2

رابطه تابع هزینه در الگوریتم FCM را بنویسید و ارتباط آن را با پارامتر c شرح دهید

تابع هزینه در الگوریتم *FCM* به صورت زیر می‌باشد:

$$\sum_i \sum_k u_{ik}^m ||x_k - c_i||^2$$

به طور کلی، این تابع هزینه فاصله داده‌ها از مراکز خوشه‌ها را در u_{ik}^m ضرب کرده و از حاصل جمع می‌گیرد.

ارتباط این پارامتر با c بدین شکل است که با افزایش مکرر c، چون احتمال نزدیک بودن یک داده به یکی از مراکز خوشه بیشتر می‌شود، در نتیجه این سیگما نیز مقدار کمتری پیدا می‌کند. در حالتی که تعداد خوشه ها بسیار زیاد باشد (مثلا برابر تعداد داده ها باشد) آنگاه هر داده مرکز خوشه تکی خود می‌شود؛ در نتیجه این هزینه برابر صفر می‌شود.

شایان ذکر است که برای ما افزایش تعداد خوشه ها لزوما امری مثبت نیست؛ و بنابراین از معیارهای دیگر مانند آنتروپی داده‌ها برای تعیین بهترین c استفاده می‌کنیم.

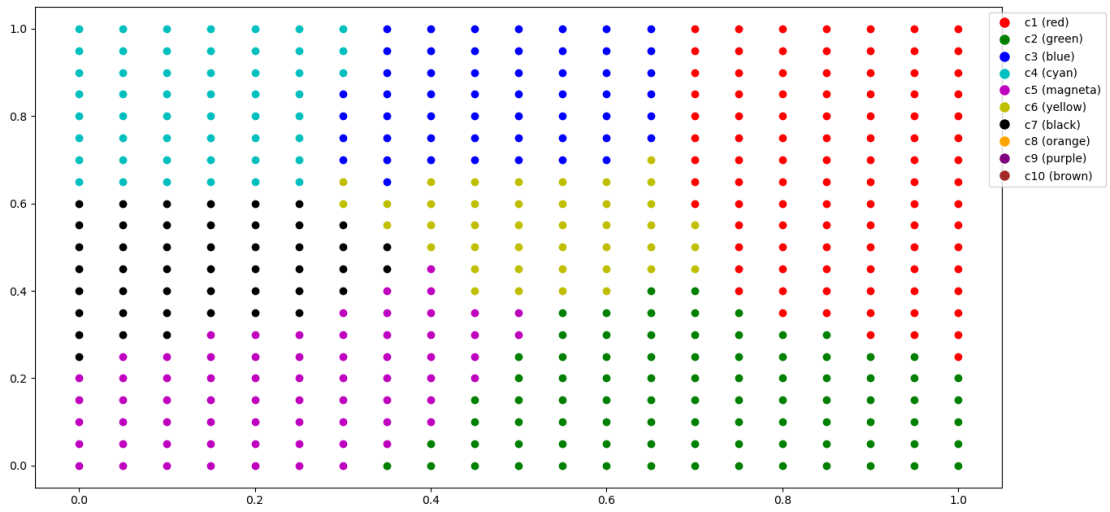
بخش امتیازی (FkNN.py)

در این بخش نیز، پس از آنکه از بخش قبل بهترین c را پیدا کرده ایم، می‌آییم و تعدادی داده تست (test data) جنریت می‌کنیم. (این داده ها به صورت evenly spaced و منظم، کل فضا 1*1 مورد نظر را می‌پوشانند)

سپس می آییم و برای هر کدام از داده‌های تست، نزدیک ترین k نقطه داده‌های اصلی مان به این داده را پیدا می‌کنیم. (این کار با k بار پیمایش روی داده‌های اصلی صورت می‌گیرد)

پس از آنکه برای هر داده تست، نزدیک ترین k داده به آن را پیدا کردیم، می‌آییم و تابع تعلق داده تست را نسبت به خوشه‌هایمان (که در بخش قبل یافتیم) محاسبه می‌کنیم.

در آخر، بسته به اینکه هر داده به کدام خوشه از همه بیشتر تعلق دارد، رنگ آن را به رنگ خوشه متناظرش در می‌آوریم و به ازای همه داده‌های تست، پلات می‌کنیم. نمونه ای از این مرزبندی آورده شده است:



sample3. c = 7