به نام خدا

دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر

سیستم های هوشمند

پروژه پایانی

دی 99

فهرست مطالب

مقدمه و توضیحات کلی

شما در طول تمرین با مدل های متفاوتی برای مسئله های طبقه بندی، رگرسیون و خوشه بندی آشنا شده اید. در این تمرین قصد داریم تا اهمیت مدل های کلاسیک که در درس آموخته اید را در پژوهش و کاربرد های عملی بررسی کنید. از این رو تمرین شامل دو قسمت کلی است. در قسمت اول با مطالعه مقالاتی که در قسمت های بعدی مشخص می‌شود، با فرایند استفاده و بکارگیری مدل خاصی آشنا می‌شوید. در قسمت بعدی مسئله ساده شدی ای در همان موضوع را پیاده سازی خواهید کرد.

آشنایی با مقالات مربوط

در حل یک مسئله، مطالعه تحقیقات انجام شده در آن موضوع اهمیت بسیاری دارد. مطالعه مقاله های مربوط از چند جهت در پیشبرد تحقیق شما مفید خواهد بود. در درجه اول با اطلاع ار تحقیقات پیشین مطمئن خواهید شد که روش/ایده/مدل شما قبلا توسط محقق دیگری امتحان و بررسی نشده باشد. علاوه بر آن در انتهای مطالب علمی گاها نویسندگان مسیری برای ادامه و پیشبرد تحقیقات ارائه می‌دهند که این پیشنهادات می‌تواند در شکل دهی ایده های شما و مسیری که برای تحقیقات خود انتخاب می‌کنید بسیار مفید باشد.

در این قسمت از تمرین، از شما خواسته می‌شود تا مقالاتی که در اختیار شما قرار داده شده است را مطالعه کنید و خلاصه ای از آن تهیه کنید. در این خلاصه باید چند قسمت اصلی متداول را حتما ذکر کنید. لذا خلاصه شما باید پاسخگو و شامل قسمت های زیر باشد.

۱. خلاصه ای از مقدمه

۲. در تحقیق از چه مجموعه داده‌ای استفاده شده است؟

۳. آیا داده های توسط خود نویسندگان جمع آوری شده است یا خیر؟

۴. از چه روش هایی برای پیش‌پردازش داده ها و انتخاب ویژگی ها استفاده شده است؟

۵. از چه مدل هایی برای طبقه بندی/خوشه بندی/رگرسیون استفاده شده است؟

۶. عملکرد مدل ها با چه اطلاعاتی گزارش شده است؟ آیا این گزارش دقیق است یا خیر؟ در صورتی که پاسخ منفی است، شیوه بهتری برای گزارش عملکرد مدل پیشنهاد دهید.

۷. نتیجه گیری و دست‌آورد های پژوهش

در انتخاب مقالات تلاش شده است تا موضوعات تا جای ممکن به قسمت پیاده سازی تمرین شباهت داشته باشد تا پیش از پیاده سازی دید بهتری نسبت به مسئله داشته باشید. با این حال اگر تمایل به بررسی مقاله دیگری با موضوع مشابهی را دارید، می‌توانید درخواست خود را به دستیار آموزشی(به همراه مقاله مورد نظر) اطلاع دهید.

معرفی مسئله مورد بررسی

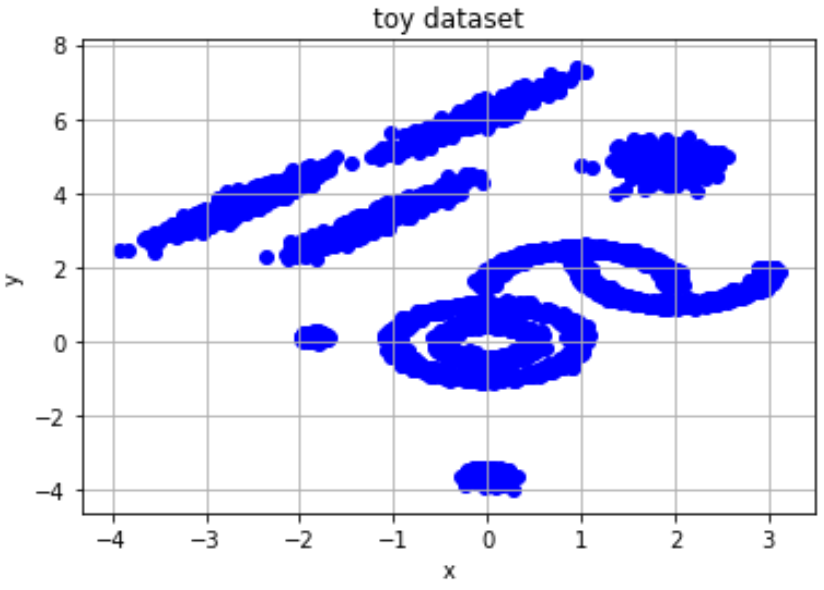
در حوزه یادگیری ماشین تحقیقات جامعی بر روی روش های supervised انجام شده است. نتیجه این تحقیقات در کاربرد های عملی همچون بینایی ماشین و پردازش زبان های طبیعی به وضوح مشخص است. با این حال این روش ها همواره نیاز به برچسب زنی تعداد زیادی داده دارند که مانع بزرگی در استفاده این روش ها برای کاربرد های خاص می باشد. بر خلاف روش های supervised ، روش های unsupervised نیازی به برچسب زنی ندارند به همین دلیل یافتن داده برای این روش ها چالش برانگیز نخواهد بود. در میان این طیف از روش های یادگیری روش های semi-supervised وجود دارند.

در این تمرین تعدادی روش خوشه‌بندی را بررسی می‌کنید و تفاوت های آن را مشاهده می‌کنید. همچنین این روش ها بر پایه های متفاوتی از جمله partitioning, distribution و fuzzy theorem استوار هستند، لذا با منطق های مختلف برای مسئله خوشه بندی بیشتر آشنا می‌شوید.

معرفی مجموعه داده

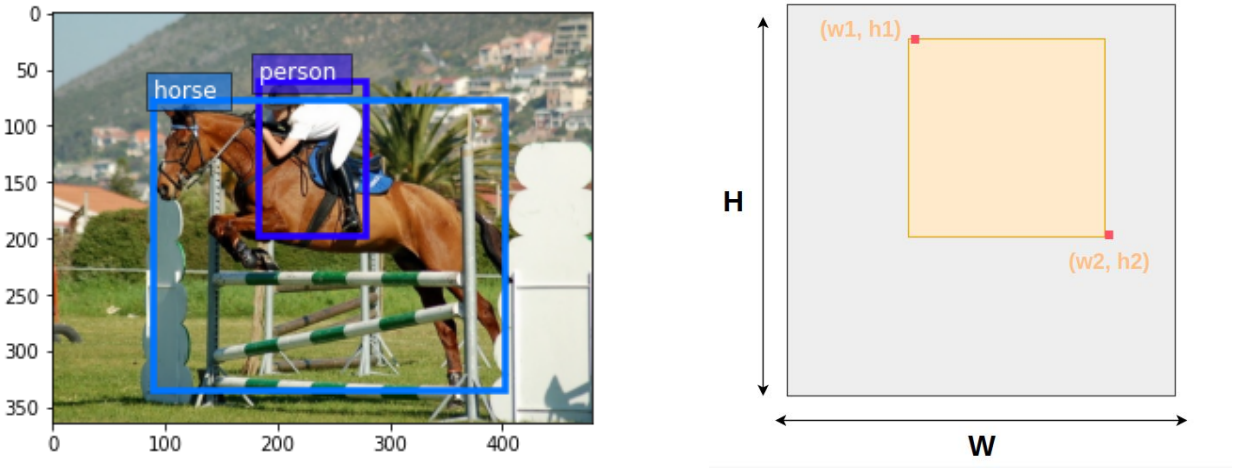
در این تمرین از دو مجموعه داده استفاده خواهید کرد. یکی از این مجموعه داده ها ساختگی است و دیگری کاربردی عملی دارد. بعد هر دو این مجموعه دادها برابر با ۲ است، از این جهت عملکرد الگوریتم های خوشه بندی را می‌توان به سادگی با استفاده از scatter plot و دیگر نمودار های مشابه مقایسه کرد.

۱. مجموعه داده Toy Dataset که در شکل ۱ نشان داده شده است، از آن جهت اهمیت دارد که شامل انواع مختلفی خوشه می‌باشد(شامل خوشه ها کروی، پوسته ای، گوسی و ..). از همین جهت تشخیص درست خوشه های برای برخی الگوریتم های می‌تواند دشوار باشد.



شکل ۱. مجموعه Toy Dataset

۲. یکی از کاربرد های عملی الگوریتم های خوشه بندی در بینایی ماشین(به صورت خاص Object Localization) است. برای مثال در شبکه های تشخیص اجسام معروف YOLO از فریم های مشخصی به عنوان Boundary Box استفاده می‌شود. مجموعه داده دوم مربوط به Boundary Box ها مجموعه داده PASCAL Visual Object Classes یا همان VOC می‌باشد. این مجموعه داده به صورت خاص برای مسئله Object Localization و Object Detection جمع آوری شده است، با این حال Boundary Box های آن به صورت متداول برای خوشه بندی استفاده می‌شوند. در این مجموعه داده هر باکس به صورت (w1, h1, w2, h2) نشان داده می‌شود. برای آن که تاثیر ابعاد تصویر از خوشه بندی حذف شود ای ویژگی ها را به صورت نرمالیزه شده ((w1-w2)/W, (h1-h2)/H) نشان می‌دهیم.



شکل ۲. مجموعه داده VOC

روش ها

روش های خوشه بندی همواره دارای هایپر پارامتری برای تعیین وابستگی نمونه های در فضای ویژگی ها هستند. در روش های کلاسیک این هایپر پارامتر ها شامل متریک استفاده شده برای تعیین فاصله/شباهت نمونه ها است، در حالی که در روش های دیگر پارامتر های دیگر از جنس فاصله تعلق دو نمونه به یکدیگر یا خوشه های دیگر را تعیین می‌کنند.

۱. روش K-Means

۲. روش Agglomerative Clustering

۳. روش Mean Shift

۴. روش Affinity Propagation

۵. روش DBSCAN

۶. روش OPTICS

۷. روش Spectral Clustering

۸. روش Guassian Mixture Model

در صورتی که تمایل دارید در مورد هر یک از این روش ها بیشتر مطالعه کنید، مقاله های مربوط به هر روش همراه با صورت پروژه در اختیار شما قرار گرفته شده است.

برای مقایسه کیفیت خوشه بندی باید متریک عددی تعریف شود و روش های مختلف با استفاده از آن متریک نسبت به هم آزموده شوند. در صورتی که بعد بردار ویژگی کمتر از ۴ باشد می‌توان کیفیت خوشه بندی را با معیار های نمایشی همچون scatter plot نیز بررسی کرد. با این حال زمانی که بعد داده ها افزایش می‌یابد تنها راه عملی برای مقایسه دو روش خوشه بندی متریک های عددی است. از متریک عددی زیر برای نمونه می‌توان استفاده کرد.

۱. متریک Separation Index

ایده پایه این متریک در آن است که خوشه ها باید تا حد ممکن از یکدیگر دور باشند(خوشه ها نامشابه باشند) و درون هر خوشه داده ها باید تا حد همکان به یکدیگر نزدیک(مشابه) باشند. از این جهت نیاز است که تعریفی از فاصله برون خوشه‌ای (میان خوشه ای) و تعریفی از فاصله درون خوشه‌ای (ماین نمونه های هر خوشه) تعریف شود.

$$SI = \min_{j} \left\{ \min_{i(i \neq j)} \left\{ \frac{d(S_i, S_j)}{\max_l d(S_l, S_l)} \right\} \right\}$$

$$d(S_i, S_j) = \min_{x,y} \{d(x, y | x \epsilon S_i, y \epsilon S_j)\}$$

$$d(S_l, S_l) = \max_{x,y} \{d(x, y | x, y \epsilon S_l)\}$$

۲. متریک Fisher's Discrimination Index

ایده پایه این متریک استفاده از ماتریس واریانس برای مدل سازی فاصله درون خوشه ای و برون خوشه ای است. برای بدست آوردن معیاری عددی از انسجام درون خوشه ای از ترکیب ماتریس های واریانس استفاده می‌شود و برای بدست آوردن معیاری عددی از فاصله برون کلاسی از ماتریس کواریانس نشان دسته های استفاده خواهد شد. در رابطه زیر نشان دسته ها همان میانگین هر خوشه هستند.

$$FDI = trace\left(S_W^{-1} S_B\right)$$

$$S_W = \sum_{i=1}^{k} S_i \qquad S_B = \sum_{i=1}^{k} Q_i \left(\hat{\mu}_i - \hat{\mu}\right)\left(\hat{\mu}_i - \hat{\mu}\right)^T \qquad S_i = \sum_{q=1}^{Q_i} \left(x^q - \hat{\mu}_i\right)\left(x^q - \hat{\mu}_i\right)^T$$

$$\hat{\mu}_i = \sum_{x^q \epsilon S_i} x^q \qquad \mu = \sum_{q=1}^{Q} x^q$$

خواسته ها

در این تمرین از شما خواسته می‌شود که روش های خوشه بندی ای که در قسمت های پیشین معرفی شد را بر روی دو مجموعه داده مسئله اعمال کرده و عملکرد آن ها را با استفاده از Scatter Plot و شاخص های Cluster Validity که پیش از این مطرح شد، با یکدیگر مقایسه کنید.

۱. لازم نیست که تمامی روش های بالا را پیاده سازی کنید تنها کافی است که روش های Agglomerative Clustering, Affinity Propagation, DBSCAN و GMM را پیاده سازی کنید. برای اجرای باقی روش های می‌توانید از کتاب‌خانه ها استفاده کنید.

۲. الگوریتم های معرفی شده را بر روی هر دو مجموعه داده اعمال کنید.

۳. برای هر الگوریتم به ازای هر مجموعه داده scatterplot خوشه ها را رسم کنید.

۴. برای هر الگوریتم و به ازای هر مجموعه داده دو شاخص cluster validity معرفی شده را محاسبه کنید.

۵. برای هر یک از الگوریتم های خوشه بندی خلاصه ای از نحوه عملکرد و یادگیری آن تهیه کنید.

نکات پایانی

- پروژه را باید در گروه های مشخص شده انجام دهید.
- گزارش شما در فرآیند تصحیح از اهمیت ویژهاي برخوردار است. لطفاً تمامی نکات و فرض هایی که برای پیاده سازی ها و محاسبات خود در نظر میگیرید را در گزارش ذکر کنید.
- در گزارش خود برای تصاویر زیرنویس و برای جداول هم بالانویس اضافه کنید.
- الزامی به ارائه توضیح جزئیات پیادهسازی در گزارش نیست. اما باید نتایج بدست آمده را گزارش و تحلیل کنید.
- برای انجام پروژه استفاده از کتابخانه ها منعی وجود ندارد.
- لطفاً گزارش، فایل کدها و سایر ضمائم مورد نیاز را با الگو PROJECT_[StudentNumber].zip زیر در سامانه مدیریت دروس بارگذاری نمایید.
- در صورت وجود هرگونه ابهام یا مشکل میتوانید از طریق رایانامه‌ی زیر با دستیار آموزشی مربوطه سجاد پاکدامن در ارتباط باشید.

sj.pakdaman@ut.ac.ir

-- موفق باشید