به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس یادگیری ماشین

تمرین پنجم

محمد ناصری

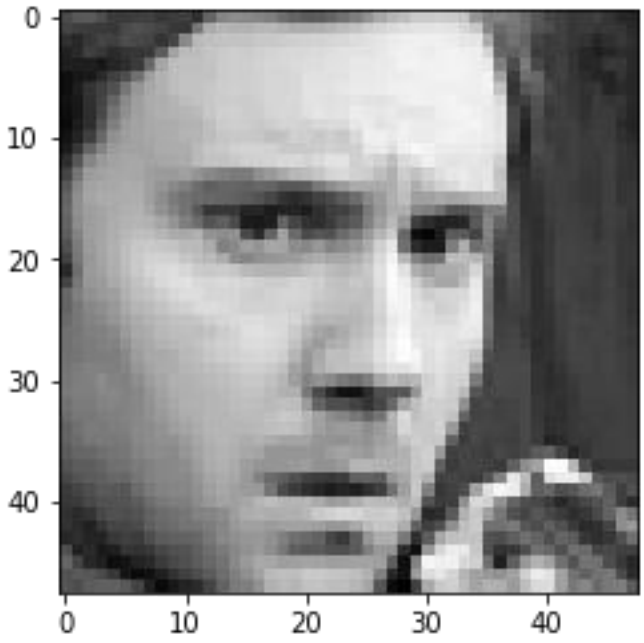
۸۱۰۱۰۰۴۸۶

خرداد ماه ۱۴۰۱

سوال ششم

کاهش ابعاد با استفاده از PCA تکنی ک متداولی برای فشرده کردن تصاویر است. تعداد کامپوننت های مورد استفاده بر نرخ فشردگی (rate compression) و کیفیت تصویر تاثیرگذار است. در این سوال از دیتاست FER2013 استفاده میکنید

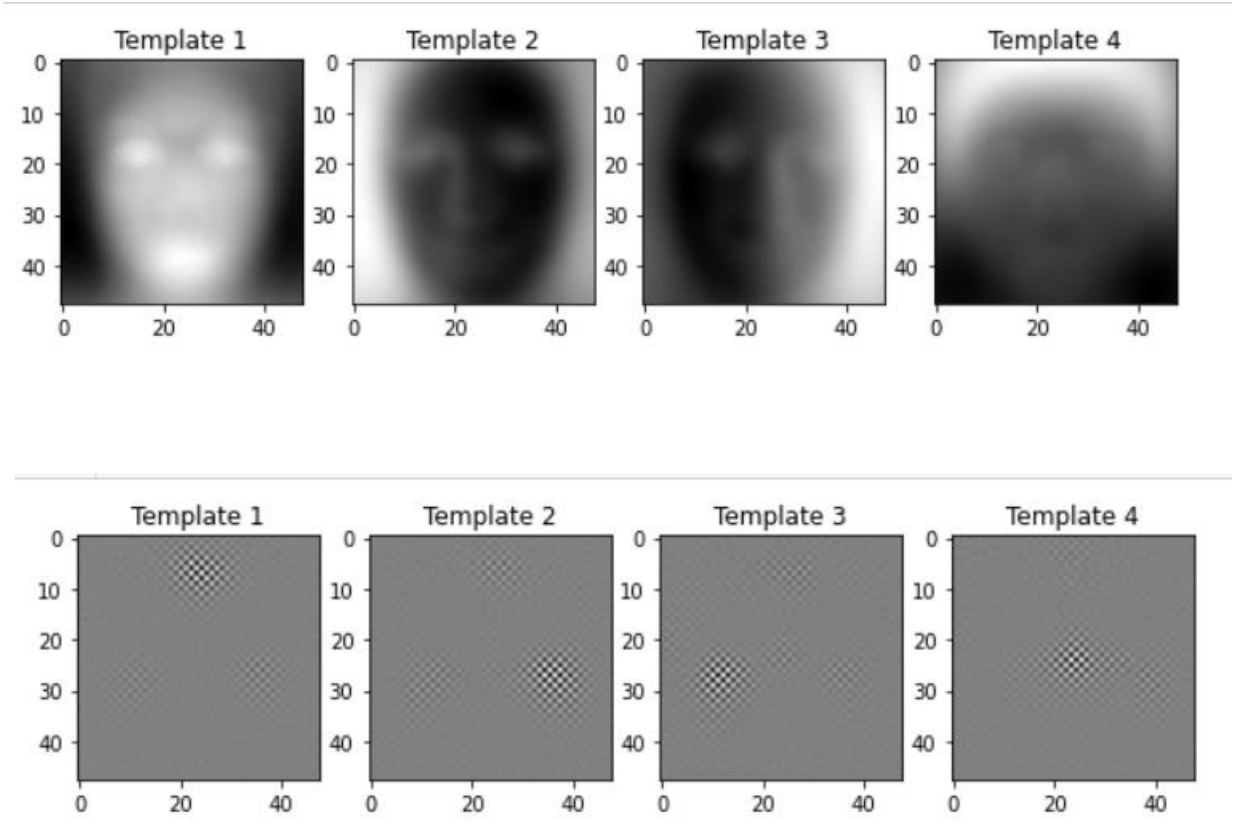
پس از ایمپورت کردن کتابخانه های لازم ، ابتدا داده‌های مورد نظر را خوانده و آنها را به شکل دلخواه خود نمایش میدهیم. داده اولیه شامل ستونی به صورت string میباشد که پس از split کردن به صورا ماتریسی با ابعاد(35887, 2304) درمی‌آید. سپس برای گرفتن دید از دیتاست یکی از تصاویر را تولید کرده و نمایش میدهیم.



در این مرحله تمام داده را یکی در نظر گرفته و به کمک pca کاهش ابعاد انجام میدهیم. سعی داریم که تعداد ابعاد را کاهش دهیم به نحوی که ابعادی که بیشترین میزان واریانس در داده ها را نشان می دهند، نگه داریم. درنموداری که در تصویر زیر آمده است، به صورت کاهشی کامپوننت هایی که بیشترین میزان واریانس مورد نظر را دارند نشان داده ایم. به این ترتیب در ستون عمودی component per variance Explainedنشان داده شده است. در محور افقی نیز تعداد کامپوننت ها آورده شده است. همانطور که واضح است کامپوننت های اولیه اساسی تر بوده اند و میزان component per variance Explained برای آن ها بسیار بیشتر از سایر کامپوننت های دیگر بوده است. هرچه کامپوننت ها اضافه شده اند، تاثیری که در به دست آوردن این میزان واریانس گذاشته اند، کمتر و کمتر شده است. در تصویر نیز Cumulative Variance Explained نشان داده شده است. با استفاده از این نمودار می توان دریافت که حدودا با تعداد کامپوننت نزدیک به ۵۰۰ حدودا ۱۰۰ درصد واریانس داده ها به دست آمده است. بنابراین می توانیم به این میزان کاهش کامپوننت داشته باشیم. مقادیر ۴ بیشترین واریانس و کمترین واریانس در نوت‌بوک نشان داده شده است.

تصاویری که با استفاده از ۴ کامپوننتی که بیشتری میزان واریانس را کپچر کرده اند را نیز رسم کرده ایم. همانطور که در این تصاویر کاملا مشهود است، کامپوننتهای اولیه ای که تصاویر بر اساس آنها رسم شده اند دارای بیشترین میزان explainability می باشد. به بیان دیگر با استفاده از این کامپوننت ها تمام اجزایی که در تصاویر وجود دارند و این تصاویر را با معنا کرده اند، در این تصاویر اورده شده است. مثلا با نگاه کردن به این تصاویر کاملا مشخص است که این تصاویر، چهره هستند وساختار کلی چشم و ابرو و بینی و دهان واضح است.

تصاویری که با کامپوننت هایی که کمترین میزان واریانس را کپچر می کنند در شکل آورده شده است. همانطور که انتظار میرود این تصایر کاملا بی معنا هستند، زیرا دارای کمترین میزان وارانس بوده اند، درنتیجه عملا هیچ فایده ای در فهمیدن تصاویر نداشته اند. این نشان دهنده این نکته است که برخی کامپوننت ها عملا سودی در به دست اوردن اطلاعاتی که در تصویر است ندارند به همین ترتیب بسیار کارامد تر است درصورتیکه حذف شوند.



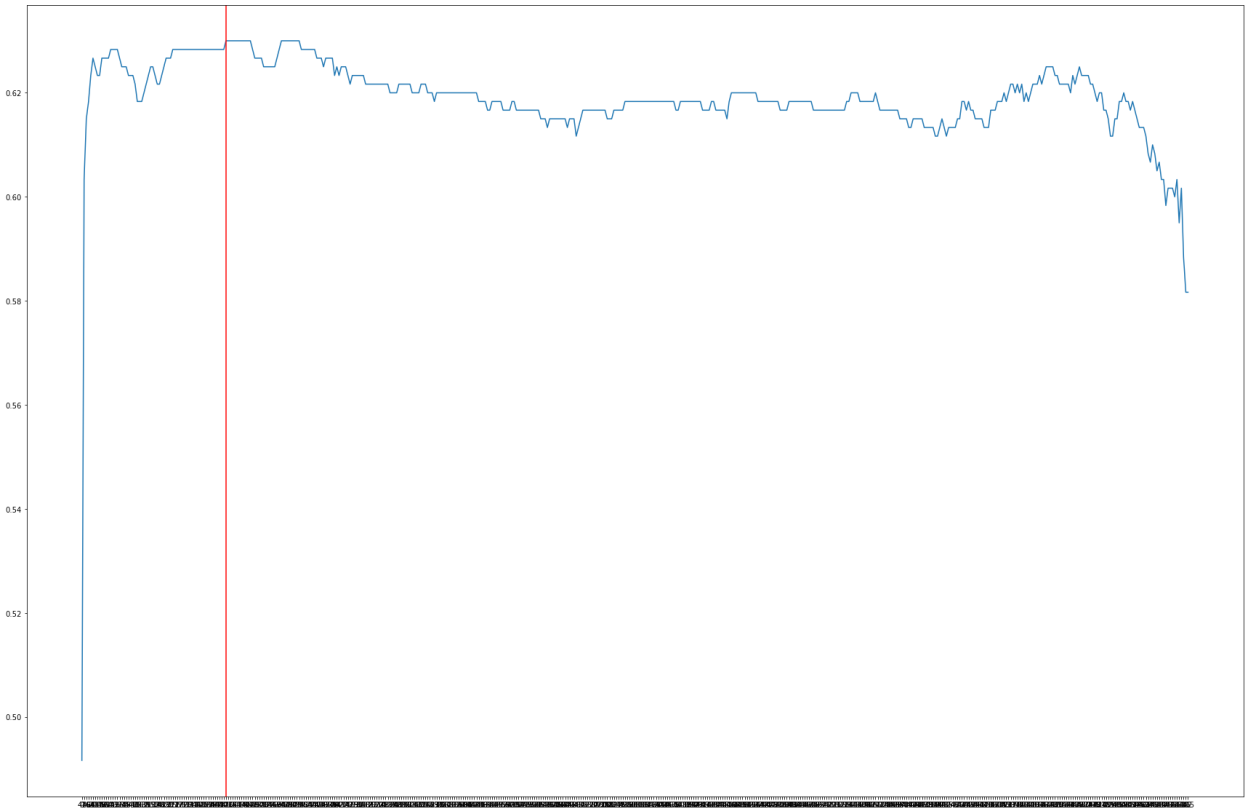
در ادامه به کمک طبقه‌بند KNN طبقه بندی بر روی داده‌های خام و کاهش‌یافته انجام شده و مقایسه شده اند. در این مقایسه که در آن دقت ها به کمک روش cross-validation بدست آمده اند، از نتایج درمی‌یابیم که تفاوت زیادی بین کاهش بعد انجام شده و داده خام وجود ندارد با این حال داده کاهش یافته اندکی دقت بهتری دارد.

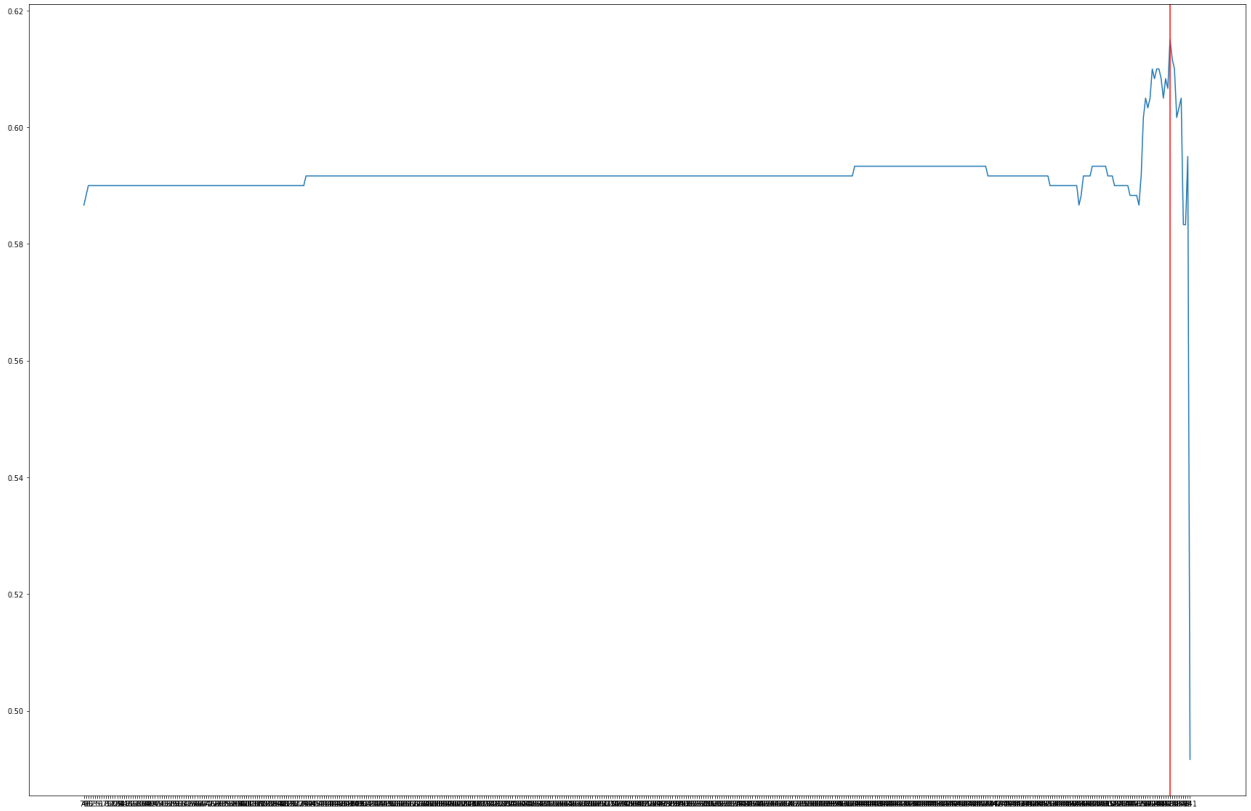
سوال هفتم

در این سوال از طبقه بند بیز به عنوان معیاری برای مقایسه ی خطای هر مرحله از مجموع مربعات خطا استفاده کنید و قسمت الف و ب را با توجه به داده های دیتاست Madelon انجام دهید.

به صورت کلی و جدای این آزمایش مزیت روش backward زمان اجرای این الگوریتم است که بسیار سریعتر از forward میباشد، همچنین این روش چون با تمام ویژگی ها کار خود را شروع می کند می تواند وابستگی بین ویژگی های مختلف را در نظر بگیرد. برای زیرمجموعه های بزرگ الگوریتم backward نتایج بهتری نسبت به forward خواهد داشت اما برای مجموعه های با ویژگی های کم الگوریتم forward می تواند با سرعت باالایی به جواب مناسب برسد. همچنین به دلیل اینکه روش backward وابستگی بین ویژگی ها را در نظر میگیرد در صورتی که یکی از این ویژگی ها مناسب حذف باشد به دلیل همبستگی این الگوریتم آن را حذف نخواهد کرد اما این مورد در الگوریتم forward وجود ندارد .ضعف بزرگ هر دو الگوریتم نیز این است که در صورتی که یک ویژگی حذف شد دیگر قابل برگشت نیست.

اما در آزمایش انجام شده، برخلاف حالت کلی، سرعت رسیدن به threshold در forward بیشتر از backward بود و زمان اجرای کمتری داشت. که میتواند ناشی از تعداد ابعاد نسبتا پایین (در مقایسه با دیتاست های با ابعاد خیلی بزرگ) باشد.





سوال هشتم

در این سوال میخواهیم از model mixture Gaussian به عنوان یک مدل generative استفاده کنیم. برای این کاربرد نیاز به کار با دیتاست اعداد دست نوشته داریم.

در این بخش نگاهی به مدل‌های (GMMs) خواهیم داشت، که می‌تواند به عنوان بسط ایده‌های پشت k-means در نظر گرفته شود. همانطور که مشخص است، در این شیوه خوشه‌بندی، فرض بر این است که هر خوشه از داده‌هایی با توزیع نرمال (گوسی) تشکیل شده و در حالت کلی نیز داده‌ها نمونه‌ای از توزیع آمیخته نرمال هستند. هدف از خوشه‌بندی مدل آمیخته گوسی یا نرمال، برآورد پارامترهای توزیع هر یک از خوشه‌ها و تعیین برچسب برای مشاهدات است. به این ترتیب مشخص می‌شود که هر مشاهده به کدام خوشه تعلق دارد. چنین روشی را در یادگیری ماشین، خوشه‌بندی برمبنای مدل می‌نامند. در شیوه خوشه‌بندی با مدل آمیخته گاوسی، برآورد پارامترهای توزیع آمیخته از یک «متغیر پنهان (Latent Variable) «استفاده می‌شود. به این ترتیب به کمک الگوریتم EM (Expectation Maximization) می‌توان برآورد مناسب برای پارامترها و مقدار متغیر پنهان به دست آورد.

دیتاست روبروی ما دیتایی از اعداد با دست نوشته شده است که لازم است تا با کمک ماشین تشخیص داده شوند. در این دیتاست ما ۱۷۹۷ نمونه که هرکدام ۶۴ فیچر دارند را مورد بررسی قرار میدهیم.

در مرحله اول روی داده خام GMM اجرا کرده و نمودار AIC رسم میکنیم و نتایج را مشاهده میکنیم و میبینیم که این الگوریتم converge کرده و متوقف میشود.

معیار اطلاعاتی آکائیکه (به انگلیسی: Akaike information criterion، یا به طور مخفف AIC) معیاری برای سنجش نیکویی برازش است. این معیار بر اساس مفهوم انتروپی بنا شده‌است و نشان می‌دهد که استفاده از یک مدل آماری به چه میزان باعث از دست رفتن اطلاعات می‌شود. به عبارت دیگر، این معیار تعادلی میان دقت مدل و پیچیدگی آن برقرار می‌کند. این معیار توسط هیروتسوگو آکائیکه برای انتخاب بهترین مدل آماری پیشنهاد شد.\با توجه به داده‌ها، چند مدل رقیب ممکن است با توجه به مقدار AIC رتبه بندی شوند و مدل دارای کمترین AIC بهترین است. از مقدار AIC می‌توان استنباط نمود که به عنوان مثال سه مدل بهتر وضعیت نسبتاً یکسانی دارند و بقیه مدل‌ها به مراتب بدتر هستند، اما معیاری برای انتخاب مقدار آستانه‌ای برای AIC که بتوان مدلی را به واسطه داشتن AIC بزرگتر از این مقدار رد کرد وجود ندارد

در حالت کلی ، AIC برابر است با $AIC = 2k - 2\ln(L)$ که k تعداد پارامترهای مدل آماری است و L مقدار حداکثر تابع درستنمایی برای مدل برآورد شده است.

در مرحله بعد همین اعمل را بر روی داده کاهش بعد یافته انجا میدهیم و نتایج را خروجی میگیریم و به کمک معیار rand_score مقایسه انجام میدهیم. با توجه به دانش تخصیص کلاس labels_true و تخصیص الگوریتم خوشه‌بندی ما از همان نمونه‌ها labels_pred، شاخص رند (تعدیل‌شده یا تنظیم‌نشده) تابعی است که شباهت دو تخصیص را اندازه‌گیری می‌کند و جایگشت‌ها را نادیده می‌گیرد.

به صورت کلی انتظار داریم که GMM قادر به converge برای داده با ابعاد بالا نباشد و برای همین از PCA استفاده میکنیم. اما در آزمایش صورت گرفته در این سوال علاوه بر اینکه GMM روی داده خام Converge میکند بلکه نتایج اندکی بهتر از داده کاهش یافته از خود نشان میدهد.

در ادامه و انتهای این سوال با توجه به نمونه‌ای از ارقام دست‌نویس، توزیع آن داده‌ها را به‌گونه‌ای مدل‌سازی کرده‌ایم که بتوانیم نمونه‌های کاملاً جدیدی از ارقام را از داده‌ها تولید کنیم. از آنجایی که GMM یک روش Generativeاست با کمک داده ماهش بعد داده شده نیز میتوان به داده‌های جدیدی در کلاس‌های داده‌های اصلی رسید. اینها ارقام دست‌نویسی هستند که به‌صورت جداگانه در مجموعه داده اصلی ظاهر نمی‌شوند.

سوال نهم

داخل نوت‌بوک

سوال دهم

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

EM method

$$\rightarrow Q(\theta, \theta^{g}) = \sum_{i=1}^{m}\sum_{j=1}^{n} \log\left(\alpha_i P_i(x_j|\theta_i)\right) p(i|x_j, \theta^{g})$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n} \log(\alpha_i)\, p(i|x_i, \theta^{g}) + \sum_{i=1}^{m}\sum_{j=1}^{n} \log\left(P_j(x_j|\theta_i)\right) p(i|x_j, \theta^{g})$$

لاگرانژ

$$\rightarrow \hat{\alpha}_i = \frac{\sum_{j=1}^{n} p(i|x_j, \theta^{g})}{n}$$

تقریع پواسون $\rightarrow$

$$Q(\theta, \theta^{g}) = \sum_{i=1}^{m}\sum_{j=1}^{n} \log\left(\frac{\theta_i^{x_j} e^{-\theta_i}}{x_j!}\right) p(i|x_j, \theta^{g})$$

$$= x_j \log\theta_i - \log x! - \theta_i$$

$$\frac{\partial}{\partial\theta_i} \sum_{j=1}^{n} \log\left(\frac{\theta_i^{x_j} e^{-\theta_i}}{x_j!}\right) p(i|x_j, \theta^{g}) = 0$$

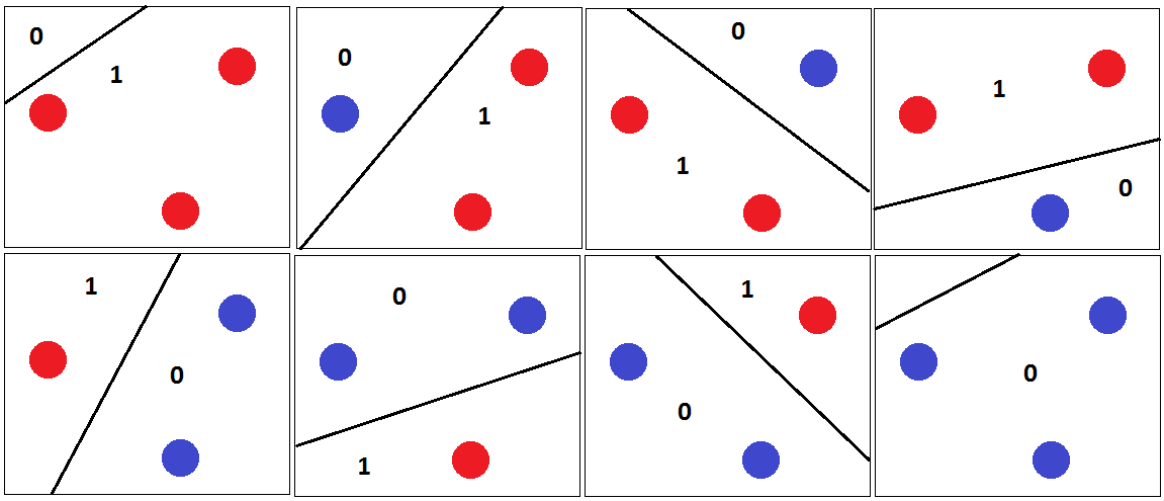
$$\rightarrow \hat{\theta}_i = \frac{\sum_{j=1}^{n} x_j \cdot p(i|x_j, \theta^{g})}{\sum_{j=1}^{n} p(i|x_j, \theta^{g})}$$

سوال یازدهم

الف و ب و ج) بعد VC یک طبقه‌بندی توسط Vapnik و Chervonenkis به‌عنوان کاردینالیته (اندازه) بزرگترین مجموعه نقاطی است که الگوریتم طبقه‌بندی می‌تواند آن‌ها را بشکند.

یک پیکربندی از N نقطه در صفحه فقط هر قرار دادن N نقطه است. برای داشتن یک بعد VC حداقل N، یک طبقه‌بندی‌کننده باید بتواند یک پیکربندی واحد از N نقطه را بشکند. برای از بین بردن پیکربندی نقاط، طبقه‌بندی‌کننده باید بتواند برای هر تخصیص مثبت و منفی برای نقاط، صفحه را به‌طور کامل تقسیم کند که نقاط مثبت از نقاط منفی جدا شوند. برای پیکربندی N نقطه، N^2 تخصیص مثبت یا منفی وجود دارد، بنابراین طبقه‌بندی‌کننده باید بتواند نقاط هر یک از آنها را به درستی جدا کند.

در مثال زیر، نشان می‌دهیم که بعد VC برای یک طبقه‌بندی‌کننده خطی حداقل 3 است، زیرا می‌تواند این پیکربندی 3 نقطه را از بین ببرد. در هر یک از $2^3 = 8$ انتساب ممکن مثبت و منفی، طبقه بندی کننده قادر است دو کلاس را کاملاً از هم جدا کند.



تعریف بعد VC این است: اگر مجموعه‌ای از n نقطه وجود داشته باشد که توسط طبقه‌بندی‌کننده شکسته شود و هیچ مجموعه‌ای از n+1 نقطه وجود نداشته باشد که توسط طبقه‌بندی‌کننده شکسته شود، بعد VC طبقه‌بندی‌کننده n است.

تعریف نمی گوید: اگر هر مجموعه ای از n نقطه را بتوان توسط طبقه بندی کننده شکست...

اگر بعد VC یک طبقه‌بندی کننده 3 باشد، لازم نیست تمام ترتیبات ممکن از 3 نقطه را بشکند.

اگر از بین تمام آرایش‌های 3 نقطه، حداقل یکی از این ترتیبات را پیدا کنید که توسط طبقه‌بندی‌کننده شکسته شود، و نتوانید 4 نقطه را پیدا کنید که می‌تواند شکسته شود، بعد VC برابر ۳ است.

درک مفاهیم عملی آن نیز مهم است. در بیشتر موارد، بعد VC طبقه‌بندی‌کننده چندان مهم نیست. بلکه بیشتر برای طبقه بندی انواع مختلف الگوریتم ها بر اساس پیچیدگی هایشان استفاده می شود. برای مثال، کلاس طبقه‌بندی‌کننده‌های ساده می‌تواند شامل اشکال پایه‌ای مانند خطوط، دایره یا مستطیل باشد، در حالی که یک دسته از طبقه‌بندی‌کننده‌های پیچیده می‌تواند شامل طبقه‌بندی‌کننده‌هایی مانند پرسپترون‌های چندلایه، درختان تقویت‌شده یا دیگر طبقه‌بندی‌کننده‌های غیرخطی باشد. پیچیدگی یک الگوریتم طبقه بندی، که به طور مستقیم با بعد VC آن مرتبط است، به مبادله بین بایاس و واریانس مربوط می شود.

یک مدل با پیچیدگی کم دارای سوگیری بالا و واریانس کم خواهد بود. در حالی که قدرت بیان پایینی دارد که منجر به سوگیری زیاد می شود، همچنین بسیار ساده است، بنابراین عملکرد بسیار قابل پیش بینی دارد که منجر به واریانس کم می شود. برعکس، یک مدل پیچیده از آنجایی که بیان بیشتری دارد، سوگیری کمتری خواهد داشت، اما واریانس بالاتری خواهد داشت زیرا پارامترهای بیشتری برای تنظیم بر اساس داده های آموزشی نمونه وجود دارد. به طور کلی، یک مدل با بعد VC بالاتر به داده های آموزشی بیشتری برای آموزش صحیح نیاز دارد، اما می تواند روابط پیچیده تری را در داده ها شناسایی کند.