**تمرین 6**

**1. روش‌های KernelPCA**:

این روش‌ها بر اساس استفاده از توابع هسته (kernel) در PCA (Principal Component Analysis) عمل می‌کنند. PCA به صورت خطی وابستگی‌های غیر خطی را از داده‌ها استخراج می‌کند. با استفاده از توابع هسته مانند تابع گاوسی یا تابع تنگنا، این روش‌ها می‌توانند روابط غیر خطی را نیز در داده‌ها کشف کنند و ابعاد آن‌ها را کاهش دهند.

تجزیه و تحلیل مؤلفه اصلی استاندارد (PCA) یک تکنیک کاهش بعد خطی است که جهت‌های بیشینه واریانس در داده‌های ورودی را پیدا می‌کند. PCA هسته‌ای (Kernel PCA) این رویکرد را گسترش می‌دهد تا بتواند با استفاده از توابع هسته، استخراج ویژگی غیرخطی را مدیریت کند.

**1. ترفند هسته:** در Kernel PCA، یک تابع هسته برای تبدیل فضای ویژگی اصلی به یک فضای بعد بالاتر به‌کار می‌رود. این تبدیل امکان ضبط روابط غیرخطی بین ویژگی‌ها را با تصویرسازی آن‌ها در یک فضای بعد بالاتر که ممکن است جداپذیری خطی داشته باشد، فراهم می‌کند.

**2. محاسبه مؤلفه‌های اصلی:** در فضای هسته، Kernel PCA مؤلفه‌های اصلی را با پیدا کردن بردارهای ویژه ماتریس هسته به جای ماتریس همبستگی استاندارد PCA محاسبه می‌کند. ماتریس هسته شامل اندازه‌گیری‌های همبستگی یا فاصله دو به دو بین بردارهای ویژگی تبدیل شده است.

**3. کاهش بعد:** پس از به‌دست آوردن بردارهای ویژه ماتریس هسته، Kernel PCA یک زیرمجموعه از بردارهای ویژه (مؤلفه‌های اصلی) را انتخاب می‌کند که اطلاعات مهم‌تری را در داده‌ها ضبط می‌کنند. تعداد مؤلفه‌های اصلی انتخاب شده برای کاهش بعد داده تعیین می‌کند.

**4. توابع هسته:** Kernel PCA بستگی به توابع هسته مانند توابع گوسی (RBF)، چندجمله‌ای یا سیگموئید دارد. این توابع اندازه‌گیری شباهت یا فاصله بین نقاط داده‌ها در فضای ویژگی تبدیل شده را انجام می‌دهند. انتخاب تابع هسته بستگی به خصوصیات داده و تبدیلات غیرخطی مورد نظر دارد.

**5. انتخاب پارامتر:** Kernel PCA شامل انتخاب پارامترهای مناسب است، مانند نوع تابع هسته و هایپرپارامترهای مرتبط با آن (مانند عرض تابع هسته گوسی). این پارامترها می‌توانند تأثیر قابل توجهی بر عملکرد و کارآیی روش داشته باشوند. از روش‌های اعتبارسنجی متقابل یا جستجوی شبکه می‌توان برای یافتن مقادیر بهینه پارامترها استفاده کرد.

**6. کاهش بعد:** Kernel PCA امکان کاهش بعد را با انتخاب یک زیرمجموعه از مؤلفه‌های اصلی فراهم می‌کند. تعداد مؤلفه‌های انتخاب شده تعیین‌کننده بعد داده‌های تبدیل شده است. انتخاب تعداد نامناسبی از مؤلفه‌ها ممکن است منجر به از دست دادن اطلاعات شود، در حالی که انتخاب تعداد زیادی ممکن است منجر به بیش‌برازش یا پیچیدگی محاسباتی زیاد شود.

**7. کاربردها:** Kernel PCA در حوزه‌های مختلفی از جمله پردازش تصویر، بینایی ماشین، بیوانفورماتیک و پردازش زبان طبیعی استفاده می‌شود. این روش می‌تواند برای وظایفی مانند تشخیص چهره، تحلیل بافت، تحلیل بیان ژن و خوشه‌بندی اسناد استفاده شود.

Kernel PCA می‌تواند برای استخراج ویژگی غیرخطی مفید باشد هنگامی که داده مبنایی با روابط غیرخطی پیچیده دارد که نمی‌تواند به خوبی در یک فضای بعد کمتر نمایش داده شود.

**2. AutoEncoder:**

AutoEncoder یک نوع شبکه عصبی است که برای یادگیری نمایشی از داده‌ها بدون نظارت استفاده می‌شود. این شبکه داده‌ها را به یک فضای مخفی کاهش می‌دهد و سپس با استفاده از این فضای مخفی، داده‌های اصلی را با تلاش برای بازسازی تقریبی مجدداً بازسازی می‌کند. این فرایند باعث می‌شود که ویژگی‌های مهم و نماینده داده‌ها در فضای مخفی استخراج شوند، که ممکن است ویژگی‌های غیرخطی را نیز شامل شوند.

**1. طراحی شبکه:** اتوانکودرها از مدل‌های شبکه عصبی تشکیل شده‌اند که شامل یک کدگذار (Encoder) و یک کدگشا (Decoder) هستند. کدگذار ورودی را به یک نمایش با بعد کمتر (کدگذاری) فشرده می‌کند، در حالی که کدگشا ورودی اصلی را از کدگذاری بازسازی می‌کند.

**2. یادگیری بدون ناظر:** اتوانکودرها به صورت بدون ناظر آموزش می‌بینند، به این معنی که نیازی به داده‌های برچسب‌دار ندارند. هدف کمینه کردن خطا در بازسازی بین ورودی و خروجی است، که باعث می‌شود مدل ویژگی‌های مهمتر در داده را یاد بگیرد.

**3. تبدیلات غیرخطی:** اتوانکودرها با استفاده از توابع فعال‌سازی غیرخطی (مانند سیگموید، ReLU) در لایه‌های شبکه عصبی، تبدیلات غیرخطی را مدیریت می‌کنند. استفاده از این توابع فعال‌سازی غیرخطی، به مدل امکان یادگیری توابع نگاشت پیچیده بین ورودی و فضای کدگذاری را می‌دهد.

**4. کاهش بعد:** با محدود کردن اندازه لایه کدگذار، اتوانکودرها قادر به انجام کاهش بعد هستند. بخش کدگذار شبکه یک نمایش فشرده از داده ورودی را یاد می‌گیرد و از این طریق بعد آن را کاهش می‌دهد. بعد فضای کدگذاری می‌تواند بر اساس نیازهای مسئله تنظیم شود.

**5. نسخه‌ها:** اتوانکودرها در انواع مختلفی وجود دارند، از جمله اتوانکودر ساده، اتوانکودر پراکنده، اتوانکودر پاکسازی نویز و اتوانکودر واریابل (VAE). هر نسخه تغییرات خاصی را در معماری یا فرآیند آموزش به کار می‌برد تا عملکرد را بهبود بخشد یا هدف‌های اضافی را دست یابد.

**6. لایه نگه‌دارنده:** لایه نگه‌دارنده در معماری اتوانکودر نمایانگر نمایش کدگذاری یا فشرده شده از داده ورودی است. این لایه ویژگی‌های مهم و نمایان را از ورودی در فضای با بعد کمتر ثبت می‌کند. بعد لایه نگه‌دارنده مقدار فشرده‌سازی و کاهش بعد را تعیین می‌کند.

**7. پیش‌آموزش و تنظیم نهایی:** اتوانکودرها، به خصوص اتوانکودرهای عمیق با لایه‌های چندگانه، می‌توانند به صورت پیش‌آموزش لایه به لایه به صورت بدون ناظر آموزش داده شوند. این پیش‌آموزش به مدل کمک می‌کند تا وزن‌های شبکه را مقدماتی کند و می‌تواند فرآیند تنظیم نهایی با استفاده از داده‌های برچسب‌دار برای وظایف خاص مانند طبقه‌بندی یا رگرسیون را بهبود ببخشد.

**8. مدل‌های تولیدی:** اتوانکودرهای واریابل (VAE) نوعی خاص از اتوانکودرها هستند که قادر به تولید نمونه‌های جدیدی از داده هستند. VAEها توزیع احتمال پشت سر داده ورودی را مدل می‌کنند و به این ترتیب با نمونه‌برداری از توزیع یادگرفته شده در فضای نهان، نمونه‌های جدیدی را تولید می‌کنند.

**9. کاربردها:** اتوانکودرها کاربردهای گسترده‌ای دارند که شامل پردازش تصویر و ویدئو، پردازش زبان طبیعی، شناسایی ناهنجاری و فشرده‌سازی داده می‌شود. آنها می‌توانند برای وظایفی مانند حذف نویز تصویر، بازسازی تصویر، تولید متن و یادگیری نمایش بدون نظارت استفاده شوند.

**3. t-SNE:**

t-SNE (t-Distributed Stochastic Neighbor Embedding) یک روش غیر خطی برای کاهش ابعاد و تجسم داده‌ها است. این روش بر اساس احتمال توزیع همسایگی داده‌ها در فضای بالا (اصلی) و فضای کاهش یافته عمل می‌کند. با استفاده از توزیع‌های احتمال مربوطه، t-SNE داده‌ها را در فضای کاهش یافته تجسم می‌دهد، به طوری که داده‌های مشابه در فضای اصلی به عنوان همسایگان در نزدیکی هم قرار بگیرند. این روش معمولاً برای تجسم داده‌های پیچیده و روابط غیر خطی میان آن‌ها استفاده می‌شود.

تکنیک کاهش بعد t-SNE، تاکید بر حفظ ساختار محلی داده‌ها در فضای کم‌بعد است. این تکنیک به خصوص برای تصویرسازی داده‌های بعد بالا و کشف خوشه‌ها یا الگوها بسیار موثر است.

**1. محاسبه شباهت محلی:** t-SNE شباهت بین نقاط داده‌ها در فضای بعد بالا را محاسبه می‌کند. این الگوریتم شباهت دو به دو بین نقاط را بر اساس احتمالی که دو نقطه یکدیگر را به عنوان همسایه انتخاب کنند، با در نظر گرفتن فاصله آن‌ها در فضای اصلی، محاسبه می‌کند.

**2. نگاشت به بعد کمتر:** در ابتدا، t-SNE نقاط داده‌ها را به صورت تصادفی در یک فضای کم‌بعد (مانند ۲ یا ۳ بعدی) قرار می‌دهد. سپس یک توزیع احتمالی مشابه در فضای کم‌بعد ساخته می‌شود که به هدف حفظ ساختار محلی داده‌ها می‌پردازد.

**3. بهینه‌سازی با کاهش گرادیان:** t-SNE همگرایی بین توزیع احتمالی فضای بعد بالا و کم‌بعد را کمینه می‌کند. با استفاده از بهینه‌سازی کاهش گرادیان، مکان نقاط داده‌ها در فضای کم‌بعد به صورت مکرر تغییر می‌کند. الگوریتم سعی می‌کند تا بهترین ترتیبی را پیدا کند که نقاط مشابه توسط نقاط نزدیک نمایش داده شوند و نقاط نامشابه در فاصله‌ای دور از هم قرار بگیرند.

**4. پرپلکسیتی:** t-SNE یک پارامتر به نام پرپلکسیتی معرفی می‌کند که تعادل بین ساختار محلی و ساختار کلی را تنظیم می‌کند. پرپلکسیتی به طور تقریبی تعداد همسایگان نزدیکی است که برای هر نقطه داده در حین کاهش بعد در نظر گرفته می‌شود. مقدار بالاتر پرپلکسیتی بیشتر روی ساختار کلی تأکید می‌کند در حالی که مقدار کمتر بیشتر روی ساختار محلی تمرکز می‌کند. مقدار پرپلکسیتی باید با توجه به خصوصیات داده‌ها به دقت انتخاب شود.

**5. مقداردهی تصادفی:** t-SNE قبل از بهینه‌سازیموقعیت نقاط داده‌ها را به صورت تصادفی در فضای کم‌بعد مقداردهی اولیه می‌کند. مقداردهی‌های مختلف می‌تواند به نتایج متفاوت منجر شود، بنابراین اجرای چندین بار الگوریتم و انتخاب بهترین نتیجه معمولاً توصیه می‌شود.

**6. پیچیدگی محاسباتی:** t-SNE پیچیدگی محاسباتی بالاتری نسبت به روش‌های خطی کاهش بعد مانند PCA دارد. این امر اصلیاً به دلیل فرآیند بهینه‌سازی تکراری و نیاز به محاسبه شباهت‌های دو به دو بین نقاط داده است. برای مجموعه‌داده‌های بزرگ، می‌توان از تقریب‌ها و تکنیک‌های بهینه‌سازی استفاده کرد تا محاسبات را تسریع کند.

**7. تصویرسازی:** t-SNE به طور گسترده‌ای برای تصویرسازی داده‌های بعد بالا استفاده می‌شود. این روش می‌تواند خوشه‌ها، الگوها و روابط محلی بین نقاط داده را در فضای کم‌بعد نمایش دهد. با این حال، مهم است توجه داشت که t-SNE فواصل سراسری یا ساختارهای خطی را حفظ نمی‌کند، بنابراین در تفسیر فواصل بین نقاط در فضای کم‌بعد، باید احتیاط کاملی داشته باشیم.

**8. کاربردها:** t-SNE در حوزه‌های مختلفی از جمله تحلیل تصویر، پردازش زبان طبیعی، ژنومیک و تحلیل شبکه‌های اجتماعی با موفقیت استفاده شده است. این روش به طور عمومی برای وظایفی مانند تصویرسازی تعبیه‌های بعد بالا، کاوش مجموعه‌داده‌ها و شناسایی خوشه‌ها یا گروه‌ها استفاده می‌شود.

**4. MDS (Multi-Dimensional Scaling):**

MDS نیز یک روش کاهش ابعاد است که بر اساس مقادیر فاصله بین داده‌ها عمل می‌کند. MDS تلاش می‌کند داده‌ها را در یک فضای کاهش یافته به طوری قرار دهد که فاصله‌های بین داده‌ها در فضای اصلی با فاصله‌های متناظر در فضای کاهش یافته تا حد ممکن مطابقت داشته باشند. این روش می‌تواند تمام روش‌های ذکر شده برای استخراج ویژگی‌های غیرخطی از داده‌ها را استفاده کند.

**۱. ماتریس شباهت یا عدم شباهت:** MDS یک تکنیک است که هدف آن نمایش داده‌های بعد بالا را در یک فضای کم‌بعدی با حفظ شباهت‌ها یا عدم شباهت‌های دو به دو بین نقاط داده است. برای این منظور، نیازمند یک ماتریس شباهت یا عدم شباهت است که روابط بین نقاط داده را بر اساس ویژگی‌های آنها اندازه‌گیری کند.

**۲. کمینه کردن تنش:** MDS تابعی را کمینه می‌کند که اختلاف بین فواصل دو به دو در فضای بعد بالا و فواصل در فضای کم‌بعدی را اندازه‌گیری می‌کند. این تابع تنش، میزان انطباق نمایش کم‌بعدی را با روابط دو به دو حفظ شده در فضای بعد بالا اندازه‌گیری می‌کند.

**۳. تبدیلات غیرخطی:** MDS با استفاده از توابع هسته‌ای برای محاسبه شباهت یا عدم شباهت بین نقاط داده، قادر به استخراج ویژگی‌های غیرخطی است. MDS هسته‌ای از یک ماتریس هسته برای ثبت روابط غیرخطی بین نقاط داده استفاده می‌کند، مشابه PCA هسته‌ای.

**۴. تصویرسازی:** MDS به طور معمول برای تصویرسازی داده‌های بعد بالا در دو یا سه بعد استفاده می‌شود. هدف آن حفظ فواصل نسبی بین نقاط داده است که امکان تصویرسازی خوشه‌ها، الگوها و شباهت‌ها در فضای کم‌بعدی را فراهم می‌کند.

**۵. MDS متریک و غیرمتریک:** MDS را می‌توان به دو نوع اصلی تقسیم کرد: MDS متریک و MDS غیرمتریک. MDS متریک فواصل واقعی بین نقاط داده را در فضای کم‌بعدی حفظ می‌کند، در حالی که MDS غیرمتریک بر حفظ ترتیب رتبه فواصل یا شباهت‌ها تمرکز دارد.

**۶. تابع تنش:** تابع تنش مورد استفاده در MDS اختلاف بین فواصل دو به دو در فضای بعد بالا و فواصل در فضای کم‌بعدی را اندازه‌گیری می‌کند. توابع تنش معمول شامل تنش کروسکال و تنش نگاشت سامون هستند. کمینه کردن تابع تنش فرآیند بهینه‌سیا به صورت تکراری است.

**۷. MDS کلاسیک و MDS هسته‌ای:** MDS کلاسیک فرض می‌کند که فواصل در فضای کم‌بعدی به صورت جمعی از مولفه‌ها هستند، در حالی که MDS هسته‌ای از توابع هسته برای ثبت روابط غیرخطی استفاده می‌کند. MDS هسته‌ای قادر به کار با داده‌های غیرخطی است و تبدیلاتی انعطاف‌پذیرتر ارائه می‌دهد.

**۸. بعد و تصویرسازی:** MDS قادر است داده‌ها را در فضاهای کم‌بعد مختلف، مانند فضای دوبعدی یا سه‌بعدی، نمایش دهد، بسته به تصویرسازی مورد نظر. باید با دقت بعد فضای هدف را انتخاب کرد تا تعادلی بین حفظ روابط و قابل فهمی تصویرسازی برقرار شود.

**۹. کاربردها:** MDS در زمینه‌های مختلف از جمله روانشناسی، بازاریابی، ژنومیک و تحلیل تصویر استفاده می‌شود. از آن برای تصویرسازی داده‌های بعد بالا، خوشه‌بندی بر اساس شباهت و بررسی ساختار داده استفاده می‌شود. MDS کاربردهای متنوعی در زمینه‌های مختلف از جمله روانشناسی، بازاریابی، علوم اجتماعی و تحلیل داده دارد. از آن برای تصویرسازی داده‌های نظرسنجی، بررسی ترجیحات مشتریان، تحلیل داده‌های ژنتیکی و درک الگوهای داده در سیستم‌های پیچیده استفاده می‌شود.