تمرین PCA

محمد صالح علی اکبری

تاریخ تحویل: ۲ خرداد

سؤال ۱

مراحل محاسبات PCA را با روابط هر مرحله تشریح کنید.

پاسخ

مراحل اصلی تحلیل مؤلفه‌های اصلی (PCA) به شرح زیر است:

۱. مرکزسازی داده‌ها: از هر ویژگی میانگین آن را کم می‌کنیم تا داده‌ها حول مبدأ قرار گیرند:

$$x_i^{\text{centered}} = x_i - \mu$$

که در آن μ میانگین هر ویژگی است.

۲. محاسبه ماتریس کوواریانس: با استفاده از داده‌های مرکزسازی‌شده، ماتریس کوواریانس را محاسبه می‌کنیم:

$$\Sigma = \frac{1}{n} X^T X$$

که در آن X ماتریس داده‌های مرکزسازی‌شده و n تعداد نمونه‌ها است.

۳. محاسبه مقادیر و بردارهای ویژه: مقادیر ویژه و بردارهای ویژه ماتریس کوواریانس را محاسبه می‌کنیم:

$$\Sigma \phi_i = \lambda_i \phi_i$$

که در آن λ_i مقدار ویژه و ϕ_i بردار ویژه متناظر است.

۴. مرتب‌سازی بردارهای ویژه: بردارهای ویژه را بر اساس مقادیر ویژه متناظر به‌صورت نزولی مرتب می‌کنیم.

۵. انتخاب مؤلفه‌های اصلی: m بردار ویژه اول را انتخاب می‌کنیم تا فضای ویژگی کاهش‌یافته را تشکیل دهند.

۶. پروژه‌سازی داده‌ها: داده‌های اصلی را روی فضای جدید پروژه می‌کنیم:

$$\hat{x} = \sum_{i=1}^{m} y_i \phi_i$$

که در آن $y_i = \phi_i^T x$ است.

سؤال ۲

projected_data را تعریف کرده و چرا این اقدام را انجام می‌دهیم؟

پاسخ

projected_data یا داده‌های پروژه‌شده، نمایش داده‌ها در فضای ویژگی کاهش‌یافته است که توسط مؤلفه‌های اصلی تعریف می‌شود. این داده‌ها با پروژه‌سازی داده‌های اصلی روی بردارهای ویژه انتخاب‌شده به‌دست می‌آیند:

$$Y = XW$$

که در آن:

- X ماتریس داده‌های مرکزسازی‌شده با ابعاد $n \times d$ است.
- W ماتریس بردارهای ویژه انتخاب‌شده با ابعاد $d \times m$ است.
- Y ماتریس داده‌های پروژه‌شده با ابعاد $n \times m$ است.

هدف از این پروژه‌سازی، کاهش ابعاد داده‌ها با حفظ بیشترین واریانس ممکن است. این کار باعث ساده‌سازی تحلیل داده‌ها، کاهش نویز و بهبود عملکرد الگوریتم‌های یادگیری ماشین می‌شود.