تشخیص اسپم ایمیل

محمدرضا تاجیک

400521198

هدف

هدف این پروژه توسعه یک سیستم تشخیص اسپم با استفاده از ماشین‌های بردار پشتیبان (SVM)با هسته خطی است SVM .ها مدل‌های یادگیری ماشین قدرتمندی هستند که در وظایف طبقه‌بندی متنی مانند تشخیص اسپم عالی عمل می‌کنند.

توضیحات پروژه

این پروژه شامل مراحل زیر می‌باشد:

1. بارگذاری دیتاست :بارگذاری یک دیتاست شامل ایمیل‌های دارای برچسب، که میان ایمیل‌های اسپم و غیر اسپم (هم) تفکیک شده‌اند.
2. پیش‌پردازش داده :تبدیل داده‌های متنی به ویژگی‌های عددی با استفاده از تکنیک TF-IDF(ترم فرکانس-معکوس فرکانس سند)، این تبدیل اجازه می‌دهد که مدل SVMبه خوبی پیام‌های متنی را درک و دسته‌بندی کند.

3. آموزش مدل :SVM آموزش یک مدل SVM با هسته خطی روی داده‌های پیش‌پردازش شده. هسته خطی به دلیل کارایی بالا در وظایف طبقه‌بندی متنی انتخاب شده است.
4. ارزیابی مدل :ارزیابی عملکرد مدل آموزش داده‌شده SVM برای انجام طبقه‌بندی اسپم و هم. معیار ارزیابی اصلی استفاده شده، دقت مدل است که درصد دقیقه ایمیل‌های درست طبقه‌بندی شده را اندازه‌گیری می‌کند.
5. نمودار :ROC رسم نمودار ROC برای مشاهده عملکرد مدل .SVM نمودار ROC نشان می‌دهد که چطور مدل در تفکیک درست بین نرخ مثبت واقعی (حساسیت) و نرخ مثبت غلط (1 - اختصاص) عمل می‌کند.

دیتاست

برای این پروژه از دیتاست "SMS Spam Collection" استفاده می‌شود. این دیتاست شامل مجموعه‌ای از پیام‌های SMS با برچسب‌های spam و ham است.

مراحل دقیق

1. بارگذاری دیتاست :دیتاست به یک DataFrame Pandas بارگذاری می‌شود، جایی که هر پیام به عنوان اسپم یا هم شناسایی می‌شود.
2. پیش‌پردازش داده :داده‌های متنی با تبدیل به ویژگی‌های TF-IDF پیش‌پردازش می‌شوند TF-IDF .یک آمار عددی است که اهمیت یک کلمه را در یک سند نسبت به یک مجموعه اسناد نشان می‌دهد. این کمک می‌کند تا معنای هر کلمه در زمینه طبقه‌بندی اسپم مشخص شود.

3. آموزش مدل :SVM یک مدل SVM با هسته خطی روی ویژگی‌های TF-IDF آموزش داده می‌شود .SVM ها به دلیل توانایی آنها در ایجاد یک هایپرپلین بهتر برای جداسازی کلاس‌ها در یک فضای بعد بالا، برای وظایف طبقه‌بندی متنی مناسب هستند.
4. ارزیابی مدل :مدل SVM آموزش داده‌شده با استفاده از مجموعه آزمون ارزیابی می‌شود تا عملکرد آن در