بـــه نـــام خـــدا

دانشکده مهندسی کامپیوتر

هوش مصنوعی و سیستم های خبره

پروژه سری چهارم: ماشین بردار پشتیبان

دکتر آرش عبدی

بهار 1403

طراحان : مصطفی مشکینی و محمدمهدی بردال

- **در صورت وجود هر گونه ابهام در سوالات تنها به طراح آن سوال پیام دهید.**
- **با توجه به تنظیم شدن ددلاین تمارین توسط خود شما امکان تمدید وجود ندارد.**
- **خوانا و مرتب بنویسید.**
- **از مباحث این تمرین از شما کوییز نیز گرفته خواهد شد پس حتما تمامی سوالات را "خودتان" حل کنید تا به مشکل نخورید :)**

آیدی تلگرام طراحان :

@Ayatollah_Dark_Blue
@mmbardal

مقدمه

Support Vector Machine (SVM) یکی از الگوریتم‌های قدرتمند یادگیری ماشین برای دسته‌بندی و رگرسیون است. در این پروژه، چند سوال در اختیار شما قرار گرفته که از سوال اول تا آخر سوالات به ترتیب کمی سخت تر و سخت تر می شوند. هر سوال شامل توضیحات مراحل و نحوه بارگذاری دیتاست است ولی پیشنهاد اکید من به شما این هستش که به توضیحاتی که در هر سوال داده شده زیاد تکیه نکنید و خودتان دنبال راه های متفاوت بگردید و اگر در سوالی یا مسئله‌ای گیر کردید، بعدش بیایید و از توضیحات کمک بگیرید. دلیل وجود توضیحات توی این سری از تمرینات این است که با قابلیت ها، کتابخانه ها، دیتاست ها و معیار های ارزشیابی جدیدی آشنا بشوید. (مانند TF-IDF، ماتریس سردرگمی، F1-score و غیره) اما این مورد را هم در نظر داشته باشید که در گرفتن نمره بهتر و کامل تر میزان جستجو و آزمون و خطای نقش بالایی دارد. برای مثال در سوالاتی که یک کرنل پیشنهادی معرفی شده از شما توقع می رود باقی کرنل ها را هم امتحان کرده باشید و بدانید که چرا کرنلی که انتخاب کردید (چه کرنل پیشنهادی باشد و چه نباشد) بهتر است.

در نهایت درباره‌ی دیتاست ها هم باید ذکر شود که ما برای هر سوال یک لینک دسترسی به دیتاست سوال را قرار دادیم ولی اگر با لینک های ما به مشکلی خوردید نگران نباشید، دیتاست های معروفی انتخاب شده‌اند که بتوانید با سرچ ساده منابع دیگری نیز پیدا کنید. همچنین بعضی دیتاست ها بخاطر جامع بودن دارای چند فایل و بخش هستند، مانند دیتاست سوال آخر. در این بار شما در انتخاب اینکه از چند تا از اون بخش ها و کدام بخش ها استفاده کنید اختیار دارید ولی مشروط به اینکه در نهایت دقت مدل شما از حد متناسب با سوال و توقع ما پایین تر نیاید.

موفق باشید

سوالات :

1) تشخیص اسپم ایمیل

هدف

آشنایی با نحوه استفاده از SVM برای پردازش متن و تشخیص اسپم.

توضیحات پروژه

1. دیتاست ایمیل‌های اسپم و غیر اسپم را بارگذاری کنید.
2. داده‌ها را پیش‌پردازش کنید (مثلاً تبدیل متن به ویژگی‌ها با استفاده از TF-IDF).
3. یک مدل SVM با کرنل خطی آموزش دهید.
4. مدل را ارزیابی کنید و دقت آن را گزارش دهید.
5. نمودار ROC رسم کنید.

نحوه بارگذاری دیتاست

برای این پروژه می‌توان از دیتاست "SMS Spam Collection" استفاده کرد که شامل پیامک‌های اسپم و غیر اسپم است. این دیتاست را می‌توانید از لینک زیر دانلود کنید:

SMS Spam Collection Dataset

2) تشخیص بیماری دیابت با استفاده از دیتاست Pima Indians Diabetes

هدف

آشنایی با نحوه استفاده از SVM برای دسته‌بندی داده‌های پزشکی و تشخیص بیماری دیابت.

توضیحات پروژه

1. دیتاست Pima Indians Diabetes را بارگذاری کنید.
2. داده‌ها را به داده‌های آموزشی و تست تقسیم کنید.
3. داده‌ها را استانداردسازی کنید.
4. یک مدل SVM با کرنل RBF آموزش دهید.
5. مدل را ارزیابی کنید و معیارهایی مانند دقت، فراخوانی و F1-Score را گزارش دهید.
6. نمودار ماتریس سردرگمی و ROC را رسم کنید.

نحوه بارگذاری دیتاست

دیتاست Pima Indians Diabetes شامل اطلاعات پزشکی بیماران و تشخیص بیماری دیابت است که می‌توان آن را از لینک زیر دانلود کرد:

Pima Indians Diabetes Dataset

3) پیش‌بینی قیمت مسکن با استفاده از دیتاست Boston Housing

هدف

آشنایی با نحوه استفاده از SVM برای مسائل رگرسیون.

توضیحات پروژه

1. دیتاست Boston Housing را بارگذاری کنید.
2. داده‌ها را به داده‌های آموزشی و تست تقسیم کنید.
3. داده‌ها را استانداردسازی کنید.
4. یک مدل SVR (Support Vector Regression) با کرنل RBF آموزش دهید.
5. مدل را ارزیابی کنید و خطای میانگین مربعات (MSE) را گزارش دهید.
6. نمودار مقایسه‌ای بین قیمت‌های واقعی و پیش‌بینی شده رسم کنید.

نحوه بارگذاری دیتاست

دیتاست Boston Housing شامل ویژگی‌های مختلف مربوط به خانه‌ها و قیمت آنها است و می‌توان آن را از لینک زیر بارگذاری کرد:

Boston Housing Dataset

4) تشخیص نویسنده با استفاده از دیتاست Newsgroups 20

هدف

آشنایی با نحوه استفاده از SVM برای دسته‌بندی متون و تشخیص نویسنده.

توضیحات پروژه

1. دیتاست Newsgroups 20 را بارگذاری کنید.
2. داده‌ها را پیش‌پردازش کنید (تبدیل متن به ویژگی‌ها با استفاده از TF-IDF).
3. یک مدل SVM با کرنل خطی آموزش دهید.
4. مدل را ارزیابی کنید و دقت آن را گزارش دهید.
5. نمودار دسته‌بندی را رسم کنید.

نحوه بارگذاری دیتاست

دیتاست Newsgroups 20 شامل مجموعه‌ای از مقالات خبری از 20 گروه خبری مختلف است و می‌توان آن را از لینک زیر بارگذاری کرد:

20 Newsgroups Dataset

5) پیش‌بینی بیماری‌های مزمن با استفاده از دیتاست Health Examination Survey

هدف

آشنایی با نحوه استفاده از SVM برای پیش‌بینی بیماری‌های مزمن با استفاده از داده‌های بررسی سلامت.

توضیحات پروژه

1. دیتاست Health Examination Survey را بارگذاری کنید.
2. داده‌ها را به داده‌های آموزشی و تست تقسیم کنید.
3. داده‌ها را استانداردسازی و نرمال‌سازی کنید.
4. یک مدل SVM با کرنل RBF آموزش دهید.
5. مدل را ارزیابی کنید و معیارهایی مانند دقت، فراخوانی و F1-Score را گزارش دهید.
6. نمودار ماتریس سردرگمی و ROC را رسم کنید.

نحوه بارگذاری دیتاست

دیتاست Health Examination Survey شامل داده‌های مربوط به بررسی‌های سلامت افراد است که می‌توان آن را از لینک زیر دانلود کرد:

Health Examination Survey