به نام خدا

دانشکده مهندسی برق و کامپیوتر دانشگاه تهران

هوش مصنوعی، ترم بهار۹۸-۹۷

پروژه Bayesian، مهلت: ۱ اردیبهشت

تشخیص پیام‌های هرز[1]

مقدمه

در این پروژه شما باید با استفاده از naive bayesian، سیستمی را طراحی کنید که بتواند با گرفتن یک پیام، مشخص کند که این پیام هرز است یا خیر.

توضیح مسئله

فایل train_test.csv که در اختیار شما قرار داده شده، یک فایل شامل حدود ۵۰۰۰ نمونه است و هر سطر شامل ۲ ستون است .ستون اول (type)نوع پیام ham) یا (spamرا مشخص می‌کند و ستون دوم (text)متن پیام را نشان می‌دهد.

شما باید در ابتدا، فایل ورودی را خوانده و متن پیام‌های مختلف را تا جایی که می‌توانید Normalize کنید. (برای مثال کارهایی که می‌توانید بکنید حذف حروف پرتکرار مانند this, that, theو ... از متن‌ها، تبدیل همه‌ی حروف بزرگ به کوچک، تبدیل همه‌ی فعل‌ها به مصدر و ... است.) هر چه بیشتر بتوانید متن‌ها را یکپارچه کنید، در انتها خروجی دقیق‌تری کسب خواهید کرد .برای این کار می‌توانید از کتابخانه‌هایی مانند nltk استفاده کنید.

در مرحله‌ی بعدی، باید ویژگی‌های (Feature) مختلفی را که به ذهنتان می‌رسد در متن‌ها بررسی کنید و برای هر کدام از ویژگی‌ها، پیام‌های spam و ham را از لحاظ داشتن یا نداشتن آن ویژگی (یا مقدار آن) مقایسه کنید و اگر بین دو گونه‌ی پیام تفاوت معناداری وجود داشت، از آن ویژگی برای دسته‌بندی پیام‌ها استفاده کنید، در غیر اینصورت از استفاده از آن صرف نظر کنید.

مثالی از این ویژگی‌ها، طول پیام‌ها، کلمات پراستفاده در هر کدام از انواع پیام‌ها، دنباله‌های پرتکرار در هر نوع و ... است.

شما باید حداقل دو ویژگی را مورد بررسی قرار دهید و در گزارش کار خود، با کشیدن نمودارهای مربوط، دلیل استفاده یا عدم استفاده از آن ویژگی را توضیح دهید.

برای مثال نمودار میانگین طول نمودار را برای هر یک از انواع پیام‌ها رسم کنید و مقادیر را مقایسه کنید.

همچنین استفاده از بیش از دو ویژگی و ایده‌های خلاقانه‌ی شما برای بررسی متن‌ها، نمره امتیازی خواهد داشت.

[1] Spam

Overfit

در حل این مسئله داده های خود را به درستی به train و test تقسیم بندی کنید و بررسی کنید که آیا overfit رخ داده است یا خیر. دقت کنید در قسمت گزارش کار فقط خطای مربوط به داده های تست را گزارش کنید.

معیار ارزیابی الگوریتم

Recall = CorrectDetectedSpams / Spams

Precision = CorrectDetectedSpams / DetectedSpams

Accuracy = CorrectDetected / Total

CorrectDetectedSpams: تعداد پیام‌هایی که به درستی به عنوان اسپم شناسایی شده‌اند.

DetectedSpams: تعداد پیام هایی که به عنوان اسپم شناخته شده اند.

CorrectDetected: تعداد پیام هایی که درس تشخیص داده شده اند.

Total: تعداد کل پیام‌ها

- توجه کنید که از داده های تست در فرایند یادگیری نباید استفاده کنید و در غیر این صورت خطای به دست آمده غیر واقعی خواهد بود.

ارزیابی برنامه

یک فایل evaluate.csv در میان فایل‌ها قرار دارد که ستون دوم آن (text) یک پیام است و ستون اول آن، id پیام را مشخص می‌کند. شما باید الگوریتم خود را در نهایت و پس از طی مراحل train,test روی این مجموعه داده اجرا کنید و مشخص کنید که هر پیام spam است یا ham. در نهایت خروجی آن را باید در قالب یک فایل csv که ستون اول آن id پیام است و ستون دوم آن نوع پیام است، آپلود کنید. نمونه فایل خروجی sample_output.csv نیز در میان فایل‌ها قرار داده شده است.

گزارش کار

در گزارش کار خود موارد زیر را بیان کنید.

۱. توضیح کلی داده ها و تعیین فیچرهای مورد استفاده
۲. توضیح کلیت الگوریتم پیاده سازی شده
۳. توضیح overfit و راهکار تشخیص آن
۴. بررسی وجود overfit در راه حل پیاده سازی شده
۵. دقت نهایی پروژه بر اساس معیارهای ارزیابی خطا (دقت روی داده های تست را ثبت کنید)
۶. دو ویژگی بررسی شده به همراه نمودارهای آن ها

* نهایتا باید گزارش کار خود را به همراه کدها و output.csv آپلود کنید.

-در صورتی که سوال داشتید در فروم درس مطرح کنید و اگر ابهامی داشتید هم می توانید حضوری یا از طریق ایمیل با من صحبت کنید.

Shahryar.Soltanpour@gmail.com

موفق باشید.

-شهریار-