درس یادگیری ماشین

مدرس: دکتر سامان هراتی‌زاده

نیم‌سال اول ۹۶-۱۳۹۵

تمرین شماره‌ی دو

مهلت ارسال تمرین : ۱۳۹۵/۸/۲۲

الگوریتم Naïve Bayes

تاریخ تحویل حضوری : ۱۳۹۵/۸/۲۳

پیاده‌سازی سیستم تشخیص هرزنامه (۱۰۰نمره)

۱-۱) معرفی

الگوریتم Naïve Bayes یکی از روش‌های کارآمد در پالایش اسناد متنی به ویژه پالایش نامه‌های الکترونیک و پیامک‌ها محسوب می‌شود. این روش غالبا با استفاده از روش خورجین کلمات[۱] به منظور تفکیک هرزنامه‌ها[۲] از متون مشروع[۳] مورد استفاده قرار می‌گیرند. هدف از این تمرین پیاده‌سازی یک سیستم دسته بندی متون[۴] با استفاده از الگوریتم Naïve Bayes و با بهره گیری از مدلسازی متون به روش خورجین کلمات است.

۱-۲) داده‌ها

مجموعه داده‌ی پیوست شده(sms_spam.csv) شامل ۵۵۷۴ پیامک انگلیسی است که هر یک به مربوط به یکی از کلاس‌های spam (متن هرز) و ham (متن مشروع) هستند. اطلاعات مفصل‌تر در مورد این مجموعه‌ی داده را می‌توانید در این آدرس[۵] مطالعه کنید.

۱-۳) توضیحات

به منظور پیاده‌سازی سیستم مذکور به موارد زیر توجه کنید:

۱. برای ساخت بردار ویژگی از روش خورجین کلمات استفاده کنید. نحوه‌ی تبدیل هر یک از متن‌ها به یک بردار ویژگی در یک فایل متنی پیوست شده است.

---

[۱] Bag of Words
[۲] spams
[۳] legitimate
[۴] Text classification
[۵] http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/

۲. در زمان پیش‌پردازش متون حروف توقف[۶] انگلیسی، اعداد و علائم نگارشی را حذف کنید. این حروف در یک فایل به نام stopwords.csv پیوست شده‌است.

۳. از ۸۰ درصد دادگان موجود به عنوان دادگان آموزش و ۲۰ درصد به عنوان دادگان آزمون استفاده کنید.

۴. ماتریس درهم‌ریختگی[۷] و دقت سیستم توسعه‌داده شده در مجموعه داده آزمون را گزارش کنید.

نکات مهم:

I.نمره سوالات **پیاده سازی و تحلیلی با کد** به صورت زیر در نظر گرفته میشود:

۱.کدها ٪۴۰

۲.گزارش ٪۳۰

۳.تحویل حضوری ٪۳۰

II.نمره سوالات **تحلیلی بدون کد** به صورت زیر در نظر گرفته میشود:

۱.گزارش ٪۶۰

۲.تحویل حضوری ٪۳۰

III.نمرات کدها، گزارش و تحویل حضوری منوط به ارسال به موقع **کدها و گزارش** است

[۶] Stop words
[۷] Confusion matrix