به نام خدا

بازیابی پیشرفته اطلاعات
نیم‌سال اول ۹۷-۹۸
مدرس: دکتر حمید بیگی

پروژه دوم **تاریخ تحویل: ۹۷/۰۹/۰۲**

در این پروژه شما باید الگوریتم های k-NN و Naive bayes را برای دسته بندی داده ها از پایه پیاده سازی کنید. همچنین باید روش SVM و به صورت امتیازی Random Forest را برای دسته بندی داده ها پیاده سازی کنید. (استفاده از کتابخانه های آماده تنها برای دو روش آخر مجاز است.) در آخر این روش ها با معیار های خواسته شده مقایسه و نتایج را در فایل pdf جداگانه گزارش کنید.

مجموعه داده ها

داده هایی که در اختیارتان قرار گرفته است، مجموعه ای بیش از یک میلیون مقاله خبری است که توسط یک موتور جستجوی آکادمیک به نام ComeToMyHead از بیش از ۲۰۰۰ منبع خبری جمع آوری شده است. داده ها در دو مجموعه ی آموزش و تست در اختیار شما قرار گرفته است که هر فایل از سه ستون تشکیل شده است که به ترتیب شماره دسته، عنوان خبر و متن خبر می باشد. این داده ها در چهار دسته ی 1. word و 2.sport و 3.business و 4.science/technology است.

پیاده سازی (۴۰ + ۲۰ امتیازی)

در این بخش ابتدا باید اسناد را که در واقع متن خبر ها می باشد، به فضای برداری tf-idf و در حالت ntn ببرید و سپس موارد زیر را پیاده سازی کنید.

Naive bayes

الگوریتم Naive bayes را از پایه پیاده سازی و روی داده آموزش اجرا کنید. (۱۰ نمره)

K-NN

الگوریتم K-NN را از پایه پیاده سازی کنید و روی داده های آموزش اجرا کنید. (K به عنوان ورودی گرفته شود.) (۱۰ نمره)
این الگوریتم را با k های ۱ و ۵ و ۱۰ پیاده سازی و نتیجه را گزارش دهید. (۵ نمره)

SVM

الگوریتم SVM را با استفاده از کتابخانه های موجود برای حالت soft margin پیاده سازی کنید. (پارامتر C به عنوان ورودی گرفته شود.) (۱۰ نمره)
این الگوریتم را با C های 2 ، $\frac{3}{2}$ ،1، $\frac{1}{2}$ پیاده سازی و نتیجه را گزارش دهید. (۵ نمره)

Random Forest

الگوریتم Random Forest را به طور مختصر توضیح دهید و آن را برای داده های آموزش به کمک کتابخانه های موجود پیاده سازی نمایید. (۲۰ نمره امتیازی)

نکته : در مراحلی که مقایسه پارامتر ها خواسته شده است از ده درصد اول داده های آموزش به عنوان داده ی validation استفاده نمایید. سپس الگوریتم را برای هر پارامتر روی بقیه ی داده آموزش اجرا و با مقایسه روی داده ی validation بهترین پارامتر را گزارش کنید.

ارزیابی (۲۰ نمره)

در این بخش باید برای تمام الگوریتم های پیاده سازی شده (با بهترین پارامتر به دست آمده) معیار های زیر را روی داده های آموزش و تست گزارش کنید.

معیار ها : accuracy و F1 با $\alpha = 0.5$ و $\beta = 1$ و برای هر کلاس precision و recall

نکات

1. امکان تغییر بارم‌بندی وجود دارد.
2. برای ارسال پروژه، کد و فایل گزارش خود را به صورت zip شده در سایت کوئرا بارگذاری نمایید.
3. می‌توانید سوالات و اشکالات خود را در سایت پیاتزا زیر پست مربوط به این تمرین بپرسید.
4. تمام قسمت‌های پروژه را خودتان به تنهایی پیاده‌سازی کنید. در صورت مشاهده تقلب، طبق قوانین دانشکده با شما برخورد خواهد شد.

(: موفق باشید