دانشکده مهندسی کامپیوتر و فن‌آوری اطلاعات

دانشگاه صنعتی امیرکبیر

درس

مدل‌های احتمالاتی گرافی

پروژه اول

دسته بندی متون با استفاده از مدل ساده بیز

اسفندماه ۹۴

شرح پروژه :

مجموعه داده 20 NewsGroups یکی از مشهورترین مجموعه‌ها در حوزه کاربردهای متنی در یادگیری ماشین مثل دسته‌بندی و خوشه‌بندی متن است که شامل مجموعه ای از ۲۰۰۰۰ پیام بر گرفته شده از ۲۰ گروه خبری است که به ازای هر گروه خبری ۱۰۰۰ پیام وجود دارد. برخی از این گروه‌ها شباهت بیشتری نسبت به یکدیگر دارند و برخی از آن‌ها به هیچ عنوان به هم شبیه نیستند. لیست گروه‌های مختلف این مجموعه داده را در زیر مشاهده می‌کنید:

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

این مجموعه داده را می‌توانید از لینک زیر دریافت کنید:

https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups

هدف این پروژه دسته بندی متن به کمک مدل ساده بیز (Naïve Bayes) می‌باشد.

الف) ابتدا از مجموعه داده معرفی شده، ویژگی‌های مورد نظر خود را استخراج کنید. چگونگی انتخاب ویژگی مورد نظر را شرح داده و تمامی عملیات پیش پردازش بر روی داده اولیه را توضیح دهید.

ب) سپس داده‌ها را به دو مجموعه داده آموزشی و آزمایشی (از ۷۰ درصد داده برای آموزش و از ۳۰ درصد آن-ها برای تست استفاده کنید) تقسیم کرده و مدل ساده بیز (Naïve Bayes) را با استفاده از داده آموزشی، آموزش دهید (کلیه بخش های الگوریتم یادگیری مدل پیاده سازی شود). ماتریس درهم ریختگی (confusion matrix) را برای داده های آزمایشی محاسبه کرده و ضمن ارائه آن به تحلیل نتایج بپردازید.

ج) بررسی کنید که تعداد ویژگی‌های انتخاب شده چه تاثیری بر دقت دسته‌بندی می گذارد؟

د) با فرض مشخص بودن کلاس، برای زوج‌های مختلف از ویژگی‌های انتخابی در قسمت الف، شرط وابستگی یا استقلال ویژگی‌ها را مورد بررسی قرار دهید و تحلیل خود را در این زمینه بیان کنید.

ه) انتخاب چندین ویژگی (مثلا کلمه کلیدی) از یک دسته خاص از اسناد، بر روی نتیجه نهایی دسته‌بندی چه تاثیری خواهد گذاشت؟ تحلیل خود را در مورد آن بیان کنید.

و) تعداد داده‌های آموزشی چه تاثیری بر دقت دسته‌بندی روی داده‌های آزمایشی خواهد داشت؟

ز) یک جعبه ابزار (toolbox) آماده برای پیاده سازی مدل های احتمالاتی گرافی انتخاب کرده و از آن برای ایجاد یک مدل ساده بیز برای دسته بندی اسناد مورد نظر استفاده کنید. نتایج حاصل را با نتایج به دست آمده در پیاده سازی خودتان مقایسه کنید.

فرمت گزارش :

گزارش بایستی به زبان فارسی و در قالب فایل PDF باشد. گزارش حداکثر در ۸ صفحه ارائه شود. فایل گزارش خود را به شکل "PGMS16_P1-Report_Stdnumber" نام گذاری نمایید.

(مثال PGMS16_P1-Report_93131020)

کدهای پیاده سازی خود و همچنین فایل جعبه ابزار استفاده شده را همراه با فایل گزارش که طبق فرمت فوق تهیه شده، در قالب یک فایل فشرده در سایت درس بارگذاری نمایید. فایل فشرده را به شکل "PGMS16_P1_Stdnumber" نام گذاری نمایید.

سؤالات خود در مورد این پروژه را می‌توانید از طریق ایمیل به آدرس Sadegh.Etemad@gmail.com ارسال نمایید. لطفاً در عنوان ایمیل کلمه PGMS16_Project_1 را قید فرمایید.