بنام خدا

الگوشناسی آماری (CE-725)

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف

پروژه پایانی درس – بهار ۱۳۹۱

به نکات زیر توجه فرمائید:

- این پروژه شامل سه فاز است که زمان اجرای فاز اول از امروز شروع می‌شود.
- خروجی‌های هر فاز را قبل از مهلت تعیین شده با عنوان SPR-Prj-Phx-stdnum (مثلا SPR-Prj-Ph1-90200000) و در یک فایل فشرده به همین نام به آدرس Muhammadi@dml.ir ارسال نمایید.
- برای پیاده سازی‌های خود تنها از Matlab استفاده نمایید. همچنین توجه فرمایید که کدهای Matlab بدون توضیحات (comment) پذیرفته نخواهند شد.
- گزارش شما باید کامل و جامع بوده و تمامی فعالیت‌های انجام شده توسط شما را پوشش دهد. نتایج، مقایسات و تحلیل‌ها نقش مهمی در ارزیابی فعالیت شما دارند.

اهداف پروژه

موضوع این پروژه **پالایش اخبار**[1] و مجموعه داده استفاده شده در آن، بخشی از مجموعه داده پر کاربرد 20 Newsgroups [1] است. داده‌های مورد نظر را می‌توانید از سایت درس دانلود کنید. هدف از این پروژه بررسی یک مساله کاربردی و مقایسه انواع روش‌های نه چندان پیچیده برای حل آن است.

در انتهای این پروژه انتظار می‌رود که دانشجو با استخراج مشخصه و استفاده از انواع کلاسه‌بندهای مناسب در حل مسائل و چگونگی به‌روز رسانی مدل‌های یاد گرفته شده و استفاده مناسب از آن‌ها در الگوشناسی آشنا شده و بتواند تحلیلی از تأثیر عوامل مختلف در شناسایی موفق الگوها داشته باشد.

توضیح پروژه و پایگاه داده

داده‌هایی که در اختیار شما قرار خواهند گرفت، داده‌های متنی برگرفته شده از بیست گروه خبری و شامل شش موضوع خواهد بود. شما بایستی با توجه به داده‌های آموزشی که علایق یک کاربر مجازی را به موضوع‌های مختلف نشان می‌دهند، اخبار را پالایش کنید تا کاربر تنها اخبار مورد علاقه خود را دریافت نماید. در واقع بایستی یک مساله کلاسه‌بندی دو کلاسه را انجام دهید. در این پروژه در ابتدا نیاز خواهید داشت از داده‌های متنی دریافتی، مشخصه‌های مناسب را استخراج کرده و کاهش بعد مناسب را روی آن‌ها انجام دهید. سپس از کلاسه‌بندهای مناسب استفاده نموده و عملیات پالایش اخبار را انجام دهید و نتایج روش‌های مختلف را با یکدیگر مقایسه کنید. در ابتدا فرض می‌کنید داده‌های دریافتی نشان‌دهنده علایق فرد در زمانی مشخص است و عملیات کلاسه‌بندی را به صورت کلاسیک انجام خواهد داد. در نهایت مساله را با در نظر گرفتن ترتیب زمانی برای اخبار و تغییر در علایق کاربر بررسی خواهید نمود (بخش اختیاری پروژه).

علاوه بر کیفیت کد، مستندات، نتایج گزارش شده و نحوه گزارش، نتیجه آزمونی که توسط کدهای ارسالی شما بدست می‌آید و نیز زمان اجرای برنامه شما، در نمره‌دهی نهایی تأثیر خواهند داشت. آزمون بر روی مجموعه‌هایی تصادفی از آموزش و آزمون، انتخاب شده از مجموعه‌داده 20 Newsgroup انجام

[1] News Filtering

خواهد شد.

داده‌هایی که در اختیار شما قرار می‌گیرند، شامل ۶ پوشه (forsale، motorcycles، hockey، crypt، electronics و space) که معرف گروه‌های خبری هستند، خواهد بود. در فاز دوم از پروژه، فرض کنید علایق کاربر شامل گروه‌های خبری electronics، crypt و hockey است. در فاز سوم، فرض کنید ترتیب ورود داده‌ها به این صورت است: داده‌های اول ۶ گروه، داده‌های دوم ۶ گروه و ... . به علاوه ۶ بازه زمانی مساوی با برچسب ۰ تا ۵ برای ورود داده‌ها در نظر بگیرید. اگر برچسب‌های ۰ تا ۵ را برای گروه‌های خبری به ترتیب حضور در پوشه در نظر بگیریم، گروه‌های خبری مورد علاقه کاربر در بازه زمانی i ام، گروه‌های خبری (i, i+1, i+2) mod 6 خواهند بود. دسته‌بندی اخبار در فاز دوم و سوم، بایستی با توجه به برچسب‌های ذکر شده در این بخش باشد.

کلاسه‌بند پایه

در این پروژه شما می‌بایست روش پیشنهادی خود در فازهای دو و سه را با یک کلاسه‌بند پایه پیاده شده روی یک مشخه پایه مقایسه کنید. کلاسه‌بند پایه مورد استفاده، کلاسه‌بند بند بیز ساده[۲] است. مشخصه پایه مورد استفاده نیز به این صورت است که هر خبر را به صورت یک بردار دودویی نشان‌دهنده وقوع یا عدم وقوع کلمات در آن خبر مدل می‌کنید. بدیهی است بخشی از انجام پروژه، پیاده‌سازی کلاسه‌بند پایه بر روی مشخصه پایه خواهد بود.

فازهای انجام پروژه

الف) فاز یک، آشنایی با مساله شامل:
(۲۵٪ نمره، مهلت تحویل: ۱۳۹۰/۱۲/۲۱)

- جستجو و مشخص کردن اولیه مشخصه‌های مورد استفاده
- جستجو و مشخص کردن اولیه کلاسه‌بندهای مورد استفاده
- روش‌های انتخاب داده‌های آموزش و آزمایش
- مشخص کردن روش‌های صحت‌سنجی (confusion matrix و ...) و انتخاب پارامترها (cross validation، marginal likelihood maximization و ...)

خروجی: یک فایل pdf در قالب مناسب.

مهم‌ترین بخش این فاز، جستجو برای یافتن ویژگی‌های مناسب داده‌هاست. شما می‌توانید از روش‌های معروف موجود در زمینه کلاسه‌بندی متون استفاده کنید یا با جستجوی مقالات موجود در این زمینه مشخصه‌های مناسب را پیدا کنید.

ب) فاز دو، پالایش اخبار با توجه به علایق فرد در یک زمان خاص:
(۷۵٪ نمره، مهلت تحویل: ۱۳۹۱/۲/۳۱)

- کلاسه‌بندی اخبار با استفاده از کلاسه‌بند پایه و مشخصه پایه
- کلاسه‌بندی اخبار با استفاده از کلاسه‌بندها و مشخصه‌های انتخاب شده در فاز قبل (استفاده از یک کلاسه‌بند پایه و یک مشخصه پایه اجباری است).
- مقایسه چگونگی کارکرد روش‌های مختلف استخراج مشخصه و کلاسه‌بندهای مختلف بر اساس چارچوب ارائه شده در فاز اول.
- ارائه گزارش کامل بر اساس چهارچوب ارائه شده در بخش قبل
  - مقایسه کامل نتایج روش پایه با روش پیشنهادی و ارائه جداول و نمودارهای مناسب
  - ارائه دلایل بهبود عملکرد روش پیشنهادی و بحث در نتایج بدست آمده
  - تحلیل کامل پارامترها و میزان حساسیت الگوریتم به آنها با نمودارها یا جدول‌های مناسب

[۲] Naïve Bayes

- ذکر نقات قوت و ضعف روش پیشنهادی و جاهایی که این روش بد کار می‌کند
- ارائه پیشنهاداتی برای بهبود روش پیشنهادی

خروجی:

- یک فایل pdf به فرمت مناسب شامل گزارش پروژه و نتایج حاصله از انجام پروژه و تحلیل نتایج
- کدهای پیاده‌سازی شده
  - یک فایل main.m بدون هیچ‌گونه پارامتر ورودی. با اجرای این فایل، بایستی پارامترهای ورودی در صورت نیاز از کاربر گرفته شوند.
  - کد شما حین اجرا بایستی آدرس مجموعه داده مورد استفاده (شامل کل داده‌های آموزش و آزمایش) را که یک پوشه خواهد بود دریافت کند، روش پیشنهادی را اجرا نماید و نتایج را در قالب مناسب نمایش دهد.
- سایر کدهای مورد نیاز
  - جعبه‌ابزارهای استفاده شده که به صورت پیش‌فرض در Matlab وجود ندارند.
- توجه کنید که هیچ فایل داده‌ای میانی از شما پذیرفته نخواهد شد. کدها بایستی با گرفتن مجموعه داده ورودی، خودشان مشخصه‌ها و سایر موارد مورد نیاز را استخراج نمایند.

در صورت تغییر هر یک از موارد تعیین شده در فاز اول، می‌بایست در گزارشی جداگانه تغییرات انجام شده را به همراه دلیل ذکر نمایید.

پ) فاز سه، پالایش اخبار دارای ترتیب زمانی و به همراه تغییر در علایق کاربر: (۲۰ درصد نمره اضافه، مهلت تحویل: ۱۳۹۱/۳/۳۱)

- در این فاز فرض خواهید کرد، اخبار دارای ترتیب زمانی هستند. شیوه آزمایش این فاز به این صورت خواهد بود که هر خبر پس از ورود بایستی توسط برنامه کلاسه‌بندی شود. سپس می‌توانید از برچسب واقعی خبر استفاده کرده و کلاسه‌بند خود را به روز نمایید. بنابراین در این فاز داده‌های آموزش و آزمایش جدا از هم نخواهند بود. علاوه بر آن نمی‌توانید فرض کنید همه داده‌ها در ابتدا در اختیار شما هستند و لذا یادگیری به صورت افزایشی[3] انجام خواهد شد. ترتیب ورود اخبار، ترتیب نمونه‌ها در داده‌های آموزشی خواهند بود.
- بررسی یک روش مناسب جهت حفظ کارایی در صورت تغییر در علایق کاربر: در این قسمت نیازمند به یک ساز و کار فراموشی جهت تطابق با داده‌های جدید خواهید بود. توجه کنید که مهمترین قسمت این فاز، تعیین ساز و کار فراموشی مناسب خواهد بود.
- مقایسه روش خود با نتایج کلاسه‌بند پایه: در این قسمت در ابتدا داده‌ها را به دو قسمت آزمایش و آزمون تقسیم کنید. سپس کلاسه‌بند پایه را روی آن اجرا نمایید.
- در قسمت مراجع، چند منبع [2,3] در باره این فاز ارائه شده است که می‌توانید به آنها مراجعه کنید.

مراجع

[1] Asuncion, A., Newman, D.: UCI machine learning repository (2007).

[2] Widmer, G. and M. Kubat, *Learning in the presence of concept drift and hidden contexts*. Machine Learning, 1996. 23(1): pp. 69-101.

[3] Klinkenberg, R., *Learning drifting concepts: Example selection vs. example weighting*. Intelligent Data Analysis, 2004.**8** (3): pp. 281-300.

[3] incremental