بنام آنکه عزت ازآن اوست

دانشگاه صنعتي اميركبير

(پلی تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

درس یادگیری ماشین

تمرین مبحث مدل های بیز ساده

استاد محترم

جناب آقای دکتر شیری قیداری

نمرهء این تمرین از 110% محاسبه میگردد به این معنا که با ارائهء هر گونه ایدهء جدید و یا کاراضافه، شما میتوانید نمره ای مازاد بر نمرهء اصلی را به خود اختصاص دهید.

همچنین دراین تمرین شما میتوانید از کمک افراد مجرب، اینترنت و یا هر منبع مناسب دیگری بهرمند شوید، با این شرط که منابع دقیقا ذکر شده و این استفاده بصورت غیر مستقیم باشد. در این راستا هر گونه کپی برداری یا واگذاری تمرین به اشخاص دیگر تقلب محسوب شده و پیامد آن نه تنها نمرهء تمرین ذیل، بلکه نمرهء سایر تمارین و نمرهء پایانی شما را نیز تحت تاثیر قرار میدهد.

فایلی که برای حل تمرین ارائه میدهید باید شامل گزارش شخصی شما با فرمت PDF بعلاوهء کدهای مربوطه باشد. این فایل را با نام خودتان، شمارهء دانشجویی، شمارهء تمرین (به عنوان مثال Sheykhi.9326169.HW5) به صورت فایل زیپ شده، در زمان مقرر ارسال نمایید. هر روز دیرکرد شما در ارسال این فایل، موجب از دست دادن 10% از نمرهء اصلی خواهد شد.

موفق باشید.

1. مفهوم مدل های مولد (generative) و مدل های تمایزی (discriminative) را شرح داده و این دو مدل را در پنج مورد مقایسه کنید.

2. با در نظر گرفتن دومدل مولد و تمایزی، صحت و عدم صحت موارد زیر را شرح دهید.

الف. مدل دسته بندی کننده logistic regression یک مدل تمایزی است.

ب. در صورت داشتن large dataset بکارگیری مدل‌های تمایزی ارجح بر مدل‌های مولد است.

ج. در مورد missing data مدل مولد رفتار بهتری از خود نشان میدهد.

د. در مدل مولد وابستگی بین پارامترها تاثیر کمتری دارد.

ه. میزان بایاس در مدل‌های تمایزی بیشتر است.

و. ریسک overfitting در مدل‌های تمایزی بیشتر است.

3. در صورتی که x نشان‌دهندهء مقدار پارامترها باشد و y کلاس دسته را نشان دهد، با داشتن مدل دسته بندی کننده باینری که ۷ ویژگی دارد، مدل تمایزی و مدل مولد متناظر را رسم کرده و تعداد پارامترها را در هر یک محاسبه کنید.

4. هر یک از مفاهیم knowledge ، objective function ، hypothesis space discovery ، likelihood ، latent variable وhidden layer را توضیح دهید.

5. شرح دهید که در یک مدل بیز ساده، متغیر پنهان به چه معناست و چه کاربردی دارد؟

6. هدف از آموزش در مدل بیز ساده چیست؟

7. نقطه ضعف اصلی مدل هایی که بر اساس likelihood استنتاج میکنند چیست؟

پیاده سازی

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

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

الف) ابتدا از مجموعه داده معرفی شده، ویژگی‌های مورد نظر خود را استخراج کنید. چگونگی انتخاب ویژگی و 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 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؟

کدهای پیاده سازی خود و همچنین جعبه ابزار استفاده شده را همراه با فایل گزارش در قالب فایل PDF در یک فایل فشرده ارسال نمایید.