دانشکده مهندسی سامانه‌های هوشمند

گروه علوم داده

یادگیری ماشین

تمرین سه

استاد درس

دکتر سامان هراتی‌زاده

زمان تحویل: [تاریخ تحویل]

[تاریخ در زمان تمرین]

یادگیری ماشین – تمرین «سه»

دکتر سامان هراتی‌زاده
دانشگاه تهران - دانشکده سامانه‌های هوشمند

دستیاران آموزشی
مجتبی شاعفی

نیم‌سال اول ۱۴۰۳-۱۴۰۴

ددلاین: [1403/9/10]

هدف تمرین:

در این تمرین ابتدا یک مجموعه داده شامل توییت‌های انگلیسی جهت آموزش طبقه‌بند بیز ساده پیش‌پردازش می‌شود. سپس طبقه‌بند بیز ساده پیاده‌سازی و آموزش داده می‌شود. در مرحله بعد با استفاده از از شبکه بیزی به پیش‌بینی ابتلا به سرطان بر اساس علائم و ویژگی‌ها افراد پرداخته می‌شود. در این تمرین دو روش بیز ساده و شبکه بیزی با یکدیگر مقایسه خواهند شد.

موارد مورد انتظار: پیش پردازش مجموعه‌های داده، پیاده سازی طبقه‌بند متن با استفاده از بیز ساده، پیاده سازی شبکه بیزی جهت پیش‌بینی سرطان، ارزیابی مدل‌ها و تحلیل خروجی‌ها (در این تمرین استفاده از توابع کتابخانه ای مجاز نیست)

مجموعه داده:

- مجموعه داده اول: این مجموعه شامل ۳۰۹۰ توییت در چهار کلاس شادی، غم، خشم و ترس است. تمرکز اصلی این توییت‌ها بر کووید-۱۹ است.
- مجموعه داده دوم: یک مجموعه داده ابتلا به سرطان برای افراد با ویژگی‌های مختلف

سوالات:

1. مجموعه داده توییت‌های پیرامون کووید-19 با چهار برچسب احساسی مختلف در اختیار شما قرار گرفته است :

توکن های شامل کلمات توقف و هشتگ و لینک و ... را حذف کرده و سپس نمودارهای تعداد تکرار 10 توکن با بیشترین و کمترین تکرار را رسم نمایید (می توانید از کتابخانه NLTK و Gensim استفاده نمایید).

یکی از روش های های بازنمایی متن به بردار استفاده از کیسه کلمات است در این روش تعداد تکرار هر توکن در یک توییت را در به عنوان یک ویژگی در نظر می‌گیریم و بردار ویژگی ها برای هر توئیت نشان دهنده وقوع/عدم وقوع هر یکی از کلمات (ویژگی ها) در آن توئیت است.

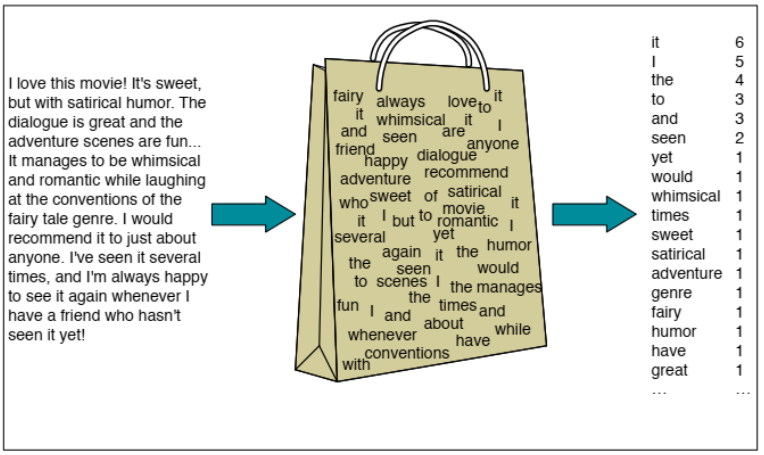
یادگیری ماشین – تمرین «سه»

دکتر سامان هراتی‌زاده
دانشگاه تهران - دانشکده سامانه‌های هوشمند

دستیاران آموزشی
مجتبی شاعفی

نیم‌سال اول ۱۴۰۳-۱۴۰۴

ددلاین: [1403/9/10]



(Fig.1) Daniel Jurafsky and James H. Martin. 2024. Speech and Language Processing.

موارد خواسته شده را جهت بازنمایی کلمات و پیاده سازی طبقه‌بند احساس با استفاده از بیز ساده انجام دهید.

a. با استفاده از روش کیسه کلمات برای توئیت ها بردارهای ویژگی با طول های 100 تا 3000 (با گام های 100 تایی) بسازید. اولویت انتخاب ویژگی‌ها، تکرار توکن در کل مجموعه داده پیش‌پردازش شده می‌باشد. به ازای هر طول بردار و به کمک هموارسازی لاپلاس و الگوریتم بیز ساده یک طبقه بند احساس آموزش دهید و آن را ارزیابی کنید. برای این کار از از اعتبارسنجی متقابل 5 لایه و معیارهای صحت و میانگین امتیاز F1 استفاده کنید.
b. نمودار مقدار شاخص صحت و میانگین امتیاز F1 را روی داده آموزش و ارزیابی بر حسب طول بردار ویژگی ترسیم کنید.

2. در مجموعه داده (ASIA) ویژگی افراد مورد مطالعه و ابتلای آنان به سرطان آورده شده است.
a. از 70 درصد داده به عنوان داده آموزش و از 30 درصد آن به عنوان داده ارزیابی استفاده کنید.
b. الگوریتم K2 را به ازای مقادیر 1، 2، 3، 4 و 5 برای حداکثر تعداد والدهای هر راس روی داده آموزش اجرا کنید تا 5 شبکه بیزی را بسازید. با استفاده از هر یک از شبکه های بیزی حاصل و شبکه مربوط به بیز ساده مدل طبقه بند را آموزش دهید و با معیارهای صحت، و امتیاز F1 روی هر دو داده آموزش و تست، ارزیابی نمایید. این کار را پنج بار تکرار کنید (پنج بار با حداکثر 1 والد، پنج بار با حداکثر دو والد، ...) و میانگین نتایج ارزیابی (روی داده آموزش و تست) را به شکلی گزارش کنید که اثر حداکثر تعداد والد در شبکه (1، 2، 3، 4 و 5) بر کارایی مدل (روی داده تست و آموزش) قابل مشاهده باشد

یادگیری ماشین – تمرین «سه»

دکتر سامان هراتی‌زاده

دانشگاه تهران - دانشکده سامانه‌های هوشمند

دستیاران آموزشی

مجتبی شاعفی

نیم‌سال اول ۱۴۰۳-۱۴۰۴

ددلاین: [1403/9/10]

c. **اطلاعات اضافه:**

- بخش عمده ارزیابی شما به درک شما از الگوریتم ها و بخش مختلف کدتان مانند محاسبه احتمال در الگوریتم های بیزی مرتبط است.
- در پیاده سازی تمرین‌ها در بخش پیاده سازی الگوریتم اصلی مانند کیسه کلمات، طبقه بند بیز ساده و شبکه بیزی و K2 استفاده از کتابخانه آماده آنها مجاز نیست و تنها جهت مقایسه می توانید از آنها کمک بگیرید. جهت پیش پردازش داده می توانید از کتابخانه های مختلف استفاده نمایید.
- در حل این تمرین، شما مجاز به استفاده از مدل‌های زبانی بزرگ (LLM) برای کمک به نوشتن کد یا حل مسائل هستید. با این حال، شما باید به‌طور کامل به تمامی کدی که تحویل می‌دهید تسلط داشته باشید و قادر به توضیح عملکرد آن باشید. استفاده از ابزارها و مدل‌های کمکی به این معنا نیست که بتوانید بدون درک کافی از کد آن را ارائه دهید؛ هدف این است که دانش و درک عمیقی از مفاهیم و راه‌حل‌ها داشته باشید و صرفا پاسخ نهایی کافی نیست.
-
- لطفا گزارش، کدها و سایر ضمایم را در یک پوشه با نام زیر قرار داده و آن را فشرده سازید و سپس در سامانه elearn بارگذاری نمایید:
  - HW[Number]_[Lastname]_[StudentNumber].zip
- در صورت وجود مشکل و یا ابهامی در مورد سوالات می‌توانید از طریق ایمیل (shaefi@ut.ac.ir) و یا گروه درسی سوال خود را مطرح کنید.