دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین بهار ۱۴۰۰

پروژه عملی - تحلیل احساسات (فاز اول)

مدرس: دکتر محمّدحسین رهبان زمان تحویل: ۲۸ خرداد

هدف از این پروژه به کارگیری مطالب آموخته شده در طول ترم، روی یک مسئله واقعی و بررسی روش‌های مختلف در یادگیری ماشین روی مجموعه دادگانی خاص است. در طراحی این پروژه تلاش کرده‌ایم تا مسیری که ممکن است در حل مسائل واقعی یادگیری ماشین طی کنید را در بستری کنترل شده به شما نشان دهیم.

زبان برنامه‌نویسی قابل استفاده در این پروژه python است و استفاده از زبان‌های برنامه‌نویسی دیگر مجاز نیست. تمامی کدهای خود را باید در فضای Jupyter Notebook بنویسید. با توجه به اینکه در بخش‌هایی از پروژه ممکن است حجم محاسباتی زیادی احتیاج داشته باشید، پیشنهاد می‌شود از سرویس Google Colaboratory استفاده کنید. برای الگوریتم‌های یادگیری ماشین فقط مجاز به استفاده از پکیج scikit-learn هستید. با اینکه محدودیتی در استفاده از پکیج‌های پردازش داده‌های متنی ندارید ولی با توجه به مطالبی که در طول ترم آموخته‌اید و حجم و گستردگی منابع موجود، پیشنهاد می‌شود برای پردازش متن‌ها از پکیج NLTK استفاده کنید. استفاده از پکیج‌های یادگیری ژرف مانند TensorFlow و PyTorch در این فاز از پروژه مجاز نیست.

۱ معرفی مسئله و مجموعه دادگان

مسئله تعریف شده در این پروژه در رابطه با تجزیه و تحلیل احساسات[1] در متن‌های واقعی است. هدف این مسئله دسته‌بندی، که جزو مسائل حوزه پردازش زبان طبیعی قرار می‌گیرد، استخراج مثبت یا منفی بودن بار معنایی و احساسی یک متن می‌باشد. هدف این پروژه، ایجاد مدل‌هایی است که یک متن را به عنوان ورودی گرفته و مثبت یا منفی بودن بار احساسی آن را تشخیص می‌دهند.

مجموعه دادگانی که در اختیار شما قرار می‌گیرد، متشکل از ۴۵ هزارتا از نظرات کاربران یک شبکه اجتماعی است که دارای برچسب‌های مثبت و منفی هستند. در فاز اول پروژه شما باید از این مجموعه داده برای آموزش مدل‌هایتان استفاده کنید و مجاز به استفاده از هیچ مجموعه داده کمکی برای بهبود نتیجه‌هایتان نیستید. استفاده از هر مجموعه داده کمکی تخلف در نظر گرفته می‌شود. ارزیابی نهایی مدل‌های شما روی یک مجموعه داده مشابه از نظرات کاربران انجام می‌شود. این مجموعه داده در زمان انجام پروژه در اختیار شما قرار نمی‌گیرد؛ بنابراین باید با استفاده از تکنیک‌های Validation و توجه به Overfit نشدن مدل‌ها روی دادگان آموزش و حتی دادگان اعتبارسنجی، تلاش کنید

[1] Sentiment Analysis

مدل‌هایی با توان تعمیم‌پذیری[1] بالا ارایه دهید.

مجموعه دادگان فاز اول را می‌توانید از اینجا دریافت کنید. این داده‌ها دارای دو ستون comment و sentiment هستند. ستون sentiment دارای مقادیر negative و positive است که منفی یا مثبت بودن بار معنایی comment را مشخص می‌کند. در تمامی مدل‌سازی‌ها، کلاس منفی را کلاس ۰ و کلاس مثبت را کلاس ۱ در نظر بگیرید.

۲ تولید داده مناسب و قابل پردازش

معمولا داده‌های متنی به صورت خام، بی‌نظمی و یا حالات خاص مختلفی دارند که تحلیل آن‌ها را دشوار می‌کند. به همین علت برای استفاده از آن‌ها، ابتدا پیش‌پردازش‌هایی رویشان انجام می‌شود. همچنین برای استفاده از بیشتر مدل‌های یادگیری ماشین نیاز داریم که داده‌هایمان در قالب[2] عدد بیان شوند. به همین علت در چنین مسائلی که داده‌ی متنی داریم، ابتدا باید آن‌ها را در قالب عددی بیان کنیم. برای این تبدیل، روش‌های زیادی وجود دارد که در ادامه با برخی از آن‌ها بیشتر آشنا می‌شویم.

۱.۲ پیش‌پردازش داده‌ها

پیش‌پردازش داده‌های متنی وابسته به منبعی که داده از آن بدست آمده‌اند و میزان نویزی که در داده‌های خام وجود دارد، می‌تواند دارای مراحل مختلفی باشد. برای پیاده‌سازی این روش‌ها معمولا می‌توان از کتابخانه‌های آماده‌ای نظیر NLTK در python کمک گرفت. روش‌های ابتدایی پیش‌پردازش شامل تبدیل حروف به حالت lower-case و حذف اعداد یا تبدیلشان به حروف، حذف کاراکترهای اضافی و جداسازی کلمات از هم[3] است. از پیش‌پردازش‌های سطح بالاتر می‌توان به حذف Stop-Wordها و انجام Stemming و Lemmatization روی کلمات بدست آمده اشاره کرد.

چنین پردازش‌هایی به این خاطر انجام می‌گیرند که خیلی از کلمات و اجزای متن لزوما ویژگی‌های خوبی نیستند و اطلاعات اضافی را به همراه ندارند. به عنوان مثال حروف بزرگ درون کلمات اطلاعات خاصی را به همراه ندارند و مشابه حالتی هستند که همه حروف کوچکند؛ به همین خاطر تمام حروف را lower-case می‌کنیم. یا مثلا کلمات پرتکراری همانند the، am، is و ... در بیشتر جملات وجود دارند و در نتیجه ویژگی خوبی محسوب نمی‌شوند.

در بخش سوم صورت پروژه که مدل‌سازی‌های مورد نظر نوشته شده است، باید از این روش‌های پیش‌پردازش برای بهبود عملکرد مدل‌ها استفاده کنید. برای مشاهده تاثیر عملکرد این روش‌ها باید مدل‌سازی‌های خواسته شده را برای سه حالت بدون پیش‌پردازش داده‌ها، پیش‌پردازش ابتدایی و پیش‌پردازش سطح بالا انجام داده و نتایج را مقایسه نمایید.

[1] Generalization
[2] Format
[3] Tokenization

۲.۲ تبدیل به بردار

پس از مرحله پیش‌پردازش، نوبت به اختصاص برداری عددی به هر نمونه متنی می‌رسد. برای این منظور از روش‌هایی مانند Word2Vec و Bag of Words استفاده می‌شود. این روش‌ها هر کلمه را به یک بردار عددی با طول ثابت تبدیل می‌کنند. سپس می‌توان با روش‌های مختلفی بردار مربوط به کل متن را از روی بردار کلماتش بدست آورد. برای مثال یکی از این روش‌ها، میانگین گرفتن از بردار کلمات است.

۳ مسئله دسته‌بندی نظرات کاربران (یادگیری با نظارت)

۱.۳ بررسی عملکرد روش‌های پیش‌پردازش

در این بخش ابتدا می‌خواهیم تاثیر پیش‌پردازش متن‌ها را روی عملکرد مدل‌ها بسنجیم. به این منظور سه مدل مبتنی بر روش‌های Logistic Regression، k-NN[۱] و SVM[۲] را در سه حالت مختلف آموزش داده و نتایج را مقایسه کنید. همچنین برای بردارسازی، روش Bag of Words را بکار ببرید.

حالت اول بدون پیش‌پردازش متن

حالت دوم پیش‌پردازش ابتدایی شامل تبدیل حروف به حالت lower-case، حذف اعداد، حذف کاراکترهای اضافی و جداسازی کلمات

حالت سوم پیش‌پردازش سطح بالا شامل تمامی مراحل **حالت دوم** به اضافه حذف Stop-Wordها، Stemming و Lemmatization.

علاوه بر این روش‌ها، برای این بخش به مراجع موجود در اینترنت برای پیش‌پردازش داده‌های متنی مراجعه کنید و اگر روش‌های دیگری نیز یافتید که به نتیجه مدل‌سازی کمک کند، آن‌ها را در این بخش پیاده‌سازی کنید. همچنین دلایل استفاده از این روش‌ها را در گزارش خود شرح دهید (مراجع استفاده شده در این بخش را در گزارش خود بیاورید).

در این بخش از پروژه زمان زیادی برای تنظیم دقیق ابرپارامتر[۳]های مسئله نگذارید. برای تعداد مناسبی حالت، این مدل‌سازی‌ها را انجام داده و بهترین نتیجه هر بخش را گزارش کنید. برای آشنایی بیشتر با شیوه پیش‌پردازش داده‌های متنی به منابع زیر مراجعه کنید:

- NLP Text Preprocessing: A Practical Guide and Template
- Text Preprocessing for NLP, Beginners to Master

[1] k Nearest Neighbors
[2] Support Vector Machine
[3] Hyperparameter

۲.۳ بررسی عملکرد روش‌های بردارسازی و تنظیم دقیق ابرپارامترها

در این بخش تاثیر هر یک از روش‌های بردارسازی داده‌های متنی را بر کیفیت مدل‌سازی بررسی می‌کنیم. الگوریتم‌های SVM، k-NN و Logistic Regression را با دو روش بردارسازی Bag of Words و Word2Vec بکار گرفته و نتایج مدل‌سازی را در گزارش خود بیان کنید. برای پیش‌پردازش این بخش از پیش‌پردازش **حالت سوم** بخش قبل استفاده کنید. همچنین با تنظیم دقیق ابرپارامترهای مدل‌ها، برای هر یک از سه الگوریتم یادگیری، بهترین مدل بدست آمده را گزارش کنید. در گزارش خود مقدار این پارامترها را به همراه روش بردارسازی‌ای که به بهترین نتیجه می‌رسد، ارایه دهید.

مدل‌های پکیج scikit-learn توسط پیمانه Pickle [۱] در python، قابل ذخیره‌سازی هستند. برای هر یک از سه الگوریتم بیان شده در بالا، مدل با بیشترین دقت دسته‌بندی را با نام‌های LR.pkl، kNN.pkl و SVM.pkl ذخیره کنید.

۳.۳ یافتن بهترین مدل (امتیازی)

در این بخش از الگوریتم MLP [۲] برای مدل‌سازی استفاده کنید. پارامترهایی که برای شبکه در نظر می‌گیرید (تعداد لایه و نورون‌های هر لایه، تابع‌های فعال‌سازی و ...) را به طور کامل در گزارش خود بنویسید. در این بخش، محدودیتی در پیش‌پردازش داده‌ها ندارید و برای بردارسازی نیز می‌توانید روش‌هایی غیر از Bag of Words و Word2Vec را بکار ببرید. شرح کوتاهی از شیوه کارکرد روش‌هایی که در هر پیمانه استفاده می‌کنید را با ذکر مرجع گزارش دهید. در پایان بهترین مدلی که یافته‌اید را با نام best.pkl ذخیره کنید.

۴ ثبت خروجی مدل‌ها

برای تمام مدل‌هایی که در این فاز از پروژه آموزش می‌دهید، خروجی تابع analysis(. , .) که تعریف آن در ادامه می‌آید را برای داده‌های آموزش و آزمون در کد و گزارش خود ثبت کنید. این تابع، تحلیلی از عملکرد دسته‌بندی که آموزش داده‌اید را نشان می‌دهد و معیار خوبی برای بررسی عملکرد مدل‌هاست.

```
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

def analysis(labels, predictions):
    print( Report: "Classification\n", classification_report(labels,
            predictions, target_names=["positive", "negative"]))
    print( Matrix: "Confusion\n", confusion_matrix(labels, predictions))
    print( "Accuracy:\n", accuracy_score(labels, predictions))
```

[1] Module

[2] MultiLayer Perceptron

۵ ارزیابی فاز اول

کد هر بخش از پروژه را در Cell متفاوتی از فضای Jupyter Notebook قرار دهید تا ارزیابی آن‌ها دقیق‌تر انجام شود. به منظور صرفه‌جویی در زمان خود، افزایش خوانایی و کمک به ارزیابی تیم دستیاران آموزشی، کدهای خود را به صورت پیمانه‌ای[1] بنویسید. گزارش خود را در قالب یک فایل pdf به همراه کدها و مدل‌های نهایی‌ای که ذخیره کرده‌اید، ارسال کنید. در نگارش گزارش، به موارد زیر توجه کنید:

- برای **حالت سوم** بخش ۱.۳، شرح کوتاهی از روش‌هایی که استفاده کرده‌اید را بیان کنید.
- در بخش ۲.۳، شرح کوتاهی از شیوه عملکرد روش‌های مختلف بردارسازی را بیان کرده و بررسی کنید که کدام روش با کدام الگوریتم به نتیجه بهتر رسیده است. همچنین نتایج بدست آمده از الگوریتم‌های مختلف را با هم مقایسه کنید.
- در بخش ۳.۳، اگر روشی غیر از روش‌های بکار رفته در بخش‌های ۱.۳ و ۲.۳ را استفاده کرده‌اید، آن را شرح دهید.
- در پایان به این نکته توجه کنید که نگارش دقیق و کامل گزارش، تاثیر مثبتی در نمره شما از این فاز پروژه دارد.

پاینده باشید

[1] Modular