به نام خدا

دانشکده علوم و فنون نوین

روش‌های یادگیری ماشین در پردازش زبان طبیعی

(۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۳-۱۴۰۲

تاریخ تحویل:
۱۴۰۲/۰۹/۱۳

تمرین شماره ۳

۱. (۱۰٪) **[پژوهش–مدل‌های زبانی بزرگ و تعبیه کلمات[۱]]** هدف از این بخش مطالعه تاثیر بافت جمله در بردار ویژگی کلمات و نحوه رفع چالش‌های چندمعنایی و هم‌نامی در مدل‌های زبانی بزرگ است. برای پردازش متن به کمک مدل‌های یادگیری لازم است که متن را به ساختاری قابل پردازش برای الگوریتم یادگیری تصویر کنیم. برای بدست آوردن مدل‌سازی مناسبی، لازم است تا به ویژگی‌هایی در سطوح مختلف زبان (ساختار نحوی جمله، ساختار معنایی کلمات و تفاوت معنا با توجه به زمینه و ...) توجه کرد. به تازگی با روش‌های تبدیل متن به بردارهای ویژگی آشنا شده‌اید. در NLP، دگرنماییِ واژه اصطلاحی است که برای بردن واژگان به فضایی برداری (نمایش آن‌ها با بردارهای عددی) به منظور تجزیه و تحلیل متن استفاده می‌شود، ساختار آن معمولاً به شکل یک بردار با مقادیر واقعی است که معنای هر کلمه را به گونه‌ای رمزگذاری می‌کند که کلماتی که در بردار به هم نزدیک‌تر هستند از نظر معنی نیز مشابه باشند. جاسازی کلمات را می‌توان با ترکیبی از عملیات مدل‌سازی زبان[۲] و تکنیک‌های یادگیری ویژگی[۳] به دست آورد که در آن نگاشتی از واژگان کلمات یا عبارات به بردارهای اعداد واقعی انجام می‌شود. این عمل از نظر مفهومی به معنای یک جاسازی ریاضی از فضایی با ابعاد بزرگ به فضای برداری پیوسته با ابعاد بسیار کمتر است. یکی از محدودیت‌های اصلی جاسازی کلمات (به‌طور کلی مدل‌های فضای برداری[۴]) این است که کلمات با معانی متعدد در یک نمایش واحد (یک بردار در فضای معنایی) ترکیب می‌شوند، به عبارت دیگر، چندمعنایی[۵] و هم‌نامی[۶] به درستی به کار گرفته نمی‌شود. با ظهور مدل‌های زبانی بزرگ، علیرغم وجود تعداد نمونه[۷] و اطاعات کافی در متن نتایج قابل توجه‌ای در تشخیص معنای درست کلمات از خود نشان داده‌اند.

مدل‌های زبانی بزرگ چگونه توانسته‌اند بدون داشتن دانشی (یا دانش کمی) از متفاوت بودن معنای کلمه مدنظر در جمله پاسخی حدودا درست و بعضا دقیق نسبت به معنای مدنظر کلمه در نظر بگیرند. می‌خواهیم بدانیم که LLMها چگونه متن را به برداری ویژگی‌ای تبدیل می‌کنند که رسیدن به چنین نتایج قابل توجه‌ای در تشخیص معنای کلمات (با فرض وجود یا عدم وجود دانش پیشین از معنای مدنظر کلمه) ممکن می‌شود! با بررسی مقالات موجود و انتخاب یک مقاله اخیر (از سال ۲۰۲۰ تا کنون) در مورد این موضوع، **دلیل انتخاب** مقاله مدنظر، خلاصه‌ای از **ایده اصلی**[۸]، **روش‌شناسی**[۹]، **انگیزه**[۱۰] و **نتایج** آن ارائه کنید و **پیشنهاد دهید** که چگونه مدل‌های یادگیری ماشینی سبک‌تری به کمک ایده بدست آمده می‌توان طراحی کرد که خروجی‌های مشابه با LLMها اما با تعداد بسیار کمتر پارامتر داشت؟

۲. (۱۰٪) **[نظری–درخت تصمیم و مقادیر گم شده]** هدف از این بخش ساخت درخت تصمیم به کمک معیارهای انشعاب مختلف و مواجه با چالش مقادیر گم شده در ساخت درخت است. در جدول زیر ۲۰ داده آموزشی به همراه خروجی مورد نظر هر یک از آن‌ها داده شده است. هر یک از سطرهای این جدول مشخص کننده یکی از نمونه‌های آموزشی با ۳ مشخصه «ویژگی ۱»، «ویژگی ۲» و «ویژگی ۳» به همراه خروجی «برچسب» است. هر یک از سه مشخصه می‌توانند یکی از مقادیر اسمی ۰، ۱ یا ۲ را اختیار کنند. خروجی نیز با توجه به مقدار مشخصه‌ها می‌تواند یکی از مقادیر اسمی ۰، ۱ یا ۲ را داشته باشد (مقادیر مفقود شده با • مشخص شده‌اند).

[1] Word Embedding
[2] Language Model
[3] Feature Learning or Representation Learning
[4] Vector Space Model
[5] Polysemy
[6] Homonymy
[7] Zero-Shot, One-Shot, Few-Shot
[8] Objective
[9] Methodology
[10] Motivation

h.veisi@ut.ac.ir

به نام خدا

روش‌های یادگیری ماشین در پردازش زبان طبیعی

(۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۳-۱۴۰۲

تاریخ تحویل:
۱۴۰۲/۰۹/۱۳

تمرین شماره ۳

جدول ۱: دادگان ارائه شده برای ساخت درخت تصمیم

ویژگی ۱	ویژگی ۲	ویژگی ۳	برچسب
2	2	1	2
2	0	1	0
0	0	2	0
1	2	0	2
0	2	•	0
1	2	1	1
1	1	1	0
1	0	1	1
0	1	1	0
0	•	1	0
0	0	0	0
2	2	2	2
1	2	1	1
2	2	2	0
0	1	2	2
2	1	2	2
1	1	1	2
2	1	2	0
1	2	1	1
1	0	0	1
1	2	2	0

فرض کنید قصد داریم یک درخت تصمیم برای این داده‌ها تولید نماییم. درخت تصمیم مورد نظر را بدست آورده و ترسیم نمایید. مراحل بدست آوردن درخت به همراه محاسبات مرتبط با آن را به طور دقیق تشریح نمایید. معیار انتخاب مشخصه را هر یک از ۳ حالت زیر در نظر بگیرید و برای هر حالت درخت را به طور جداگانه ترسیم نمایید.

۱. Information Gain
۲. Gini Index
۳. Gain Ratio

جدول2: دادگان آزمون

ویژگی ۱	ویژگی ۲	ویژگی ۳	برچسب
0	2	•	2
1	0	2	2
0	•	•	0
•	•	•	1
2	1	1	0
2	2	0	2

با توجه به دادگان آزمون داده شده، پیش‌بینی‌های هر مدل را برای دادگان آزمون بدست آورید، سپس ماتریس درهم‌ریختگی و معیارهای ارزیابی (دقت، صحت و فراخوانی) را به طور دستی محاسبه کنید.

۳. (۱۰٪) **[نظری–ماشین بردار تصمیم]** هدف از این تمرین بررسی دسته‌بند حداکثر حاشیه[1] است (**راهنمایی**: برای رسم نیز می‌توانید از زبان‌های برنامه‌نویسی کمک بگیرید.).

۱. فرض کنیم که ۷ نمونه دو بعدی به همراه برچسب به ما داده شده است.

[1] Maximal Margin Classifier

h.veisi@ut.ac.ir

به نام خدا

دانشکده علوم و فنون نوین

روش‌های یادگیری ماشین در پردازش زبان طبیعی
(۸۳۰۴۳۶۸)
نیم‌سال اول ۱۴۰۳-۱۴۰۲

تاریخ تحویل:
۱۴۰۲/۰۹/۱۳

تمرین شماره ۳

جدول 3: مشاهدات

مشاهده	X_1	X_2	برچسب
۱	3	4	قرمز
۲	2	2	قرمز
۳	4	4	قرمز
۴	1	4	قرمز
۵	2	1	آبی
۶	4	3	آبی
۷	4	1	آبی

این نقاط را با توجه به برچسب‌هایشان در یک فضای دو بعدی رسم کنید.

۲. معادله ابرصفحه[1] بهینه را بدست آورید و رسم کنید.

۳. یک قانون دسته‌بندی برای این دسته‌بند حداکثر حاشیه بدست آورید.

۴. در ابرصفحه‌ای که رسم کرده‌اید، حاشیه و بردارهای پشتیبان را مشخص کنید.

۵. بحث کنید که اگر نمونه ۷ کمی جابه‌جا شود، تاثیری در ابرصفحه حداکثر حاشیه نمی‌گذارد.

۶. به مجموعه مشاهدات، یک نمونه دیگر اضافه کنید به طوری که توسط هیچ ابرصفحه‌ای نتوان این دو برچسب را از یکدیگر تفکیک کرد.

---

۴. (۳۰٪) **[پیاده‌سازی – تخمین‌گرها[2]، دسته‌بندی و ایموجی‌ها[3]]** هدف از این تمرین استفاده از مدل‌های یادگیری ماشین از جمله SVM، Decision Tree و Random Forest برای دسته‌بندی ایموجی‌ها است. مجموعه دادگان داده شده شامل ایموجی‌های ۴ شرکت اپل، فیس‌بوک، گوگل و توییتر است. به ترتیب مراحل زیر را انجام دهید. خروجی قابل ارائه تکمیل نوت‌بوک آماده شده و پاسخ به تمامی سوالات خواسته شده در صورت سوال و دیگر سوال‌های نوت‌بوک و ارائه تحلیل‌ها و نتایج خواسته شده در گزارش است.

آماده‌سازی دادگان: در این قسمت به دریافت و آماده‌سازی دادگان می‌پردازیم، موارد لازم در نوت‌بوک Emoji.ipynb در اختیار شما قرار گرفته است.

۱. **ساخت مجموعه دادگان ایموجی:** در این بخش می‌خواهیم با استفاده از لیست emoji-meta که از بخش قبل برای عکس‌های ایموجی چهار شرکت بدست آوردیم، یک مجموعه داده به صورت (X,y) برای آموزش، اعتبارسنجی و آزمون مدل یادگیری ماشین خود بسازیم (N باید ۷۵۰۰ باشد).

۱. یک آرایه نامپای با نوع داده float32 با سایز N در D که N تعداد کل ایموجی‌ها (N برای ۴ شرکت اپل، فیس‌بوک، گوگل و توییتر در نظر گرفته شود) و D تعداد کل پیکسل‌ها برای هر ایموجی است. هر عکس ایموجی باید از RGBA (۴ کاناله) به RGB (۳ کاناله) توسط تابع rgba_to_rgb که در اختیار شما قرار گرفته است، تبدیل شود. برای اطمینان از درست بودن ساخت X، عکس X[0] را نمایش دهید. (توجه: ممکن است نیاز به reshape کردن آن داشته باشید!)

۲. در این قسمت یک بردار هدف y برای برچسب‌های ایموجی‌ها می‌سازیم. ابتدا به کمک فیلد category، تمام برچسب‌های

[1] Hyperplane
[2] Estimators
[3] Emojies

h.veisi@ut.ac.ir

به نام خدا

روش‌های یادگیری ماشین در پردازش زبان طبیعی
(۸۳۰۴۳۶۸)
نیم‌سال اول ۱۴۰۳-۱۴۰۲

دانشکده علوم و فنون نوین

تاریخ تحویل:
۱۴۰۲/۰۹/۱۳

تمرین شماره ۳

موجود را از لیست emoji-meta بدست آورید. سپس یک آرایه نامپای y با نوع داده int32 که مقادیر ۰ تا M-1 را اختیار می‌کند، بسازید (M تعداد کل برچسب‌های بدست آمده). در ادامه برای مجموعه‌دادگان X، آرایه y را برای تمام دادگان مقداردهی کنید. در پایان سایز y برابر با (N,) است.

۳. به کمک تابع train_test_split از کتابخانه Sk-Learn دادگان را به سه بخش، آموزش، اعتبارسنجی و آزمون تقسیم کنید. سپس به کمک روش‌های تغییر مقیاس ویژگی[1] (Standard Scaler) داده‌های هر بخش را نرمالایز کنید. حال برای مشاهده تغییرات در عکس پس از نرمال‌سازی، یک عکس به دلخواه از مجموعه‌دادگان آموزش انتخاب و آن را پیش و پس از نرمال‌سازی رسم کنید.

۲. **آموزش دسته‌بندها و شناسایی بهترین پارامترها:** در این بخش عملکرد دسته‌بندهای گفته شده را بر روی مجموعه دادگان بدست آمده ارزیابی می‌کنیم.

۱. ابتدا به تکمیل تابع train_estimators بپردازید. این تابع آرگومان base-estimator که یک دسته‌بند (مانند DecisionTreeClassifier) از کتابخانه Sk-Learn را دریافت می‌کند، سپس برای پارامتر param_name و مقادیر مختلف آن param_vals، مدل را بر روی دادگان (X,y) آموزش می‌دهد. خروجی این تابع باید یک لیست از مدل‌های آموزش داده شده برای تمام مقادیر مختلف آن پارامتر باشد. برای اطمینان از کد خود، این تابع برای DecisionTreeClassifier با پارامتر max-depth و مقادیر مختلف [1,5,10] فراخوانی کنید (بر روی دادگان آموزش).

خروجی شما باید به شکل زیر باشد:

```
[DecisionTreeClassifier(max_depth=1, random_state=0, splitter='random'),
DecisionTreeClassifier(max_depth=5, random_state=0, splitter='random'),
DecisionTreeClassifier(max_depth=10, random_state=0, splitter='random')]
```

۲. تابع score_estimators را تکمیل کنید. این تابع یک لیستی از تخمین‌زن‌ها را به عنوان ورودی دریافت و دقت هر کدام را بر روی دیتای (X,y) بدست می‌آورد. خروجی تابع یک لیست از این دقت‌ها است. برای درخت‌های تصمیم آموزش داده‌شده در قسمت قبل، دقت آن‌ها را برای سه مجموعه داده آموزش، اعتبارسنجی و آزمون بدست آوردید. درخت تصمیم با کدام مقدار max_depth بهترین عملکرد را داشته است؟ دلیل خود را توضیح دهید.

۳. تابع plot_estimator_scores را تکمیل کنید. این تابع باید دقت تخمین‌زن‌ها را بر اساس مقادیر مختلف پارامتر مورد نظر در سه مجموعه داده آموزش، اعتبارسنجی و آزمون نمایش دهد. به کمک این تابع دقت درخت تصمیم‌های بدست آمده در بخش‌های قبل را رسم کنید.

۴. مراحل گفته شده در قسمت‌های قبل را برای مدل SVM با مقدار پارامترهای (C = [0.001,0.01,0.1,1.0] و kernel = [rbf, linear]) و مدل Random Forest را با مقدار پارامترهای (max-depth =[1,5,10,50,100] و criterion = [gini, entropy]) انجام دهید. همچنین مدل DecisionTree را با مقدار پارامترهای (max-depth = [1,5,10,50,100] و criterion =[gini, entropy]) بررسی کنید.

۵. طبق رسم‌های انجام شده، آیا می‌توان گفت که درکل عملکرد مدل در اعتبارسنجی تخمینی از عملکرد مدل در بخش آزمون است؟ توضیح دهید.

۶. کدام یک از دسته‌بندها بهترین عملکرد در آزمون را نسبت به بهترین تنظیمات خود در اعتبارسنجی دارد؟ نام دسته‌بند و مقدار پارامترهای آن را ذکر کنید.

[1] Feature Scaling

h.veisi@ut.ac.ir

به نام خدا

دانشکده علوم و فنون نوین

روش‌های یادگیری ماشین در پردازش زبان طبیعی
(۸۳۰۴۳۶۸)
نیم‌سال اول ۱۴۰۳-۱۴۰۲

تاریخ تحویل:
۱۴۰۲/۰۹/۱۳

تمرین شماره ۳

۷. کدام یک از دسته‌بندها کمترین بیش‌برازش را دارد؟ با فرض اینکه بیش‌برازش را تفاوت مطلق بین دقت آموزش و آزمون در نظر بگیریم، نام دسته‌بند و مقدار پارامترهای آن را ذکر کنید.

۳. ارزیابی مدل و تشخیص خطاها: پس از شناسایی بهترین دسته‌بندها، به بررسی عملکرد آن‌ها به طور دقیق‌تر می‌پردازیم.

۱. بهترین دسته‌بندها از هر کدام از مدل‌های SVM، Random Forest و DecisionTree انتخاب کنید و ماتریس درهم‌ریختگی[1] و معیارهای ارزیابی (دقت، صحت، فراخوانی) را برای مجموعه دادگان آزمون رسم کنید.

۲. برای هر دسته‌بند، تمامی نمونه‌هایی را که به اشتباه دسته‌بندی شده‌اند را شناسایی و آن‌ها را رسم کنید. پس از مشاهدات این نمونه‌ها، دلیل بیاورید که چرا این دسته‌بند این نمونه را به اشتباه دسته‌بندی کرده است.

---

۵. (۴۰٪) **[پیاده‌سازی–ماشین بردار پشتیبان و ابعاد با اهمیت]** هدف از این بخش بررسی تاثیر روش‌های مختلف استخراج ویژگی از متن در عملکرد مدل SVM برای وظیفه دسته‌بندی متن است. مجموعه دادگان نظرات Snapp Food در اختیار شما قرار گرفته است. قصد داریم تا با آموزش مدل خود بر نظرات، نظرات راضی و ناراضی را تشخیص دهیم. به کمک کتابخانه Sk-learn پیاده‌سازی‌های خواسته‌شده را انجام داده و تحلیل‌های لازم را گزارش کنید.

۱. **پیش‌پردازش**: به کمک روش‌های استخراج ویژگی Unigram، Bigram، TF و TF-IDF متن را به بردار ویژگی تبدیل کنید. به این منظور لازم است برای هر روش تابعی در کدهای تحویلی پیاده‌سازی کرده و روند تبدیل را در گزارش توضیح دهید.

۲. **اجرای الگوریتم**: به کمک کرنل‌های خطی[2]، چندجمله‌ای[3]، پایه شعاعی گوسی[4] و سیگموید[5] مدل‌های خود برای هر روش استخراج ویژگی آموزش داده و پارامترها و نتایج هر مدل را در گزارش خود ذکر کنید. همچنین مدلی که بهترین دقت بر مجموعه‌دادگان آزمون دارد را انتخاب کرده و پارامترهای آن (کرنل و دیگر پارامترهای مدل) را ذکر کنید، علاوه بر آن، ماتریس درهم‌ریختگی و معیارهای ارزیابی (دقت، صحت و فراخوانی) را برای آن مدل گزارش کرده، در نهایت مرز تصمیم این مدل را رسم کنید.

۳. **آزمایش بیشتر**: قصد داریم با نمونه‌گیری تصادفی ۱۰ درصد از مجموعه دادگان آموزش، «مجموعه آموزش انتخاب شده جدیدی» ایجاد می‌کنیم، که به آن «نمونه‌های پرس و جو» می‌گوییم. نمونه‌های انتخاب شده را از دادگان آموزش اولیه حذف کنید (قرار است که از باقی دادگان برای پیدا کردن دور ترین نمونه‌ها به نمونه‌های پرس و جو استفاده کنیم). برای تبدیل متن به بردار ویژگی از TF-IDF استفاده کنید. پس از آن، برای هر کلاس (در اینجا مساله دسته‌بندی دو کلاسه است) مجموعه‌ای[6] ایجاد کنیم که نمونه‌های حاصل این عملیات را به آن اضافه می‌کنیم. سپس تجزیه مقدارهای منفرد[7] را روی دو ماتریس انجام می‌دهیم: یکی از نمونه پرس و جو و دیگری از مجموعه آموزشی جدید (دادگان آموزش بدون نمونه‌های پرس و جو). به دنبال این، شباهت کسینوسی بین این ماتریس‌ها را محاسبه می‌کنیم و جفت‌هایی (شباهت بین هر پرس و جو و دادگان آموزش) با شباهت بیش از ۵۵ درصد را حذف می‌کنیم. سپس جفت‌های باقی‌مانده و پرس و جوهای اولیه را به عنوان بخشی از مجموعه

[1] Confusion Matrix
[2] Linear
[3] Polynomial
[4] Radial Basis Function (RBF)
[5] Sigmoid
[6] Set
[7] Singular Value Decomposition (SVD)

h.veisi@ut.ac.ir

آموزشی جدید به مجموعه ما اضافه می‌شوند. این فرایند را برای همه نمونه‌های پرس و جو اعمال کنید، مجموعه‌دادگان آموزش نهایی یک مجموعه بالقوه کوچکتر می‌شود. حال با استفاده از TF-IDF متن را به بردارهای ویژگی تبدیل می‌کنیم و با استفاده از کرنل‌های خطی، چندجمله‌ای، پایه شعاعی گوسی و سیگموید مدل SVM خود را بر دادگان بدست آمده، آموزش دهید. مدلی که بهترین دقت بر مجموعه‌دادگان آزمون دارد را انتخاب کرده و پارامترهای آن (کرنل و دیگر پارامترهای مدل) را ذکر کنید، علاوه بر آن، ماتریس درهم‌ریختگی و معیارهای ارزیابی (دقت، صحت و فراخوانی) را برای آن مدل گزارش کرده، در نهایت مرز تصمیم این مدل را رسم کنید.

۴. **همبستگی شباهت**: قصد داریم تا از صحت روند انتخاب نمونه‌های متفاوت مطمئن شویم. به این منظور، همانند بخش «آزمایش بیشتر» با این تفاوت که یک نمونه از دادگان آموزش که با پرس و جو مدنظر بیش از ۹۶ درصد شباهت دارد را یک جفت اختیار کرده و میزان شباهت (شباهت کسینوسی) پرس و جو مدنظر و آن نمونه در دادگان آموزش را در ماتریس کاهش ابعاد یافته به عنوان شباهت معیار[1] در نظر می‌گیریم. حال شباهت (معیار انتخاب شباهت را شباهت کسینوسی و جاکارد در نظر بگیرید) جفت‌های بدست آمده را به کمک روش‌های استخراج ویژگی TF و TF-IDF محاسبه کرده و میزان ضریب همبستگی[2] بین نتایج شباهت معیار و نتایج شباهت به کمک روش‌های استخراج ویژگی گفته شده را محاسبه کنید (برای روش TF-IDF هر جمله را معادل یک سند در نظر بگیرید.).

راهنمایی: برای محاسبه میزان ارتباط نتایج حاصل و امتیازات جفت جملات انتخاب شده از ضریب همبستگی استفاده می‌شود که نحوه محاسبه آن از طریق رابطه زیر امکان‌پذیر است.

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

در این رابطه x_i مقدار تشابه محاسبه شده توسط شما برای جفت جمله ilم و y_i مقدار تشابه واقعی بدست آورده شده از طریق شباهت جفت‌ها در ماتریس کاهش ابعاد یافته برای این جفت جمله است. مقدار $\bar{x}$ برابر میانگین مقادیر تشابه محاسبه شده توسط شما برای همه جفت جملات است و $\bar{y}$ میانگین مقادیر عددی تشابه واقعی نمونه‌ها است. در نهایت با ارائه جدولی مشابه جدول زیر نتایج را تحلیل کنید که کدام معیار و کدام روش نمایش نتایج بهتری ارائه می‌دهد.

جدول ۴: نتایج حاصل از تحلیل شباهت جفت جمله‌ها

استخراج ویژگی / شباهت	فراوانی کلمه	فراوانی کلمه–معکوس فراوانی سند
کسینوسی		
جاکارد		

[1] Gold Standard
[2] Correlation Coefficienct

h.veisi@ut.ac.ir