

تمرین سوم درس یادگیری ماشین

عنوان: کلاس‌بندی

استاد درس: دکتر محدث

موعد تحویل: ۱۰ روز پس از بارگذاری تمرین

مقدمه

در این تمرین به مباحث مربوط به روش‌های ماشین بردار پشتیبان[۱] (SVM) و بیز ساده[۲] که جزء روش‌های پایه کلاس‌بندی[۳] در یادگیری ماشین[۴] هستند، می‌پردازیم. در نظر داشته باشید می‌توانید این تمرین را به‌صورت گروهی انجام دهید. اگر تمرین را به‌صورت گروهی انجام می‌دهید، در صفحه اول فایل تحویلی اسم هم‌گروهی‌های خود را بنویسید. همچنین تنها پاسخ‌هایی بررسی خواهد شد که از طریق سامانه کورسز[۵] تحویل داده شوند، اگر پاسخ تمرین از هر طریق دیگری به‌غیر از سامانه کورسز ارسال شود، ترتیب اثر داده نخواهد شد و نمره تمرین برابر با صفر خواهد بود. علاوه بر این، عدم رعایت هر کدام از نکات مربوط به گزارش و پاسخ تمرین که در انتهای این فایل قرار داده‌شده‌است، باعث کسری نمره خواهد شد. منظور از گزارش کد در قسمت پیاده‌سازی این است که دلیل استفاده از دستوراتی که برای پیاده‌سازی تمرین استفاده کردید و نحوه عملکرد آن‌ها را توضیح دهید. در نظر داشته باشید که پیاده‌سازی‌هایی که بدون گزارش باشند، کمتر از ۵۰ درصد نمره را خواهند گرفت.

سؤالات تشریحی (۱۲۵)

مطالعه و جستجو (۴۰)

۱. (آ) ماتریس آشفتگی[۶] در یادگیری ماشین را توضیح دهید. (۵)

(ب) معیارهای دقت[۷]، صحت[۸]، پوشش[۹] و امتیاز f1[۱۰] را توضیح دهید. (۸)

(ج) برای معیارهای صحت، پوشش و امتیاز f1، میانگین‌گیری‌های micro و macro را توضیح دهید. تفاوت این دو در چیست؟ (۷)

۲. برخی اوقات در الگوریتم بیز ساده برای محاسبه احتمال به‌جای محاسبه مستقیم احتمال از لگاریتم احتمال استفاده می‌شود.

(آ) علت این مسئله چیست؟ (علت استفاده از لگاریتم را شرح دهید.) (۵)

(ب) در نهایت مسئله به‌چه‌صورت حل می‌شود؟ یعنی، احتمال پسین[۱۱] به‌چه‌صورت به‌دست می‌آید؟ (۵)

۳. یکی از مسائل کلاس‌بندی که در آن از روش بیز ساده استفاده می‌شود، کلاس‌بندی متن در پردازش متن طبیعی[۱۲] (NLP) است. در این مسئله هر کلمه یک ویژگی در نظر گرفته می‌شود. حال اگر کلماتی در مجموعه آزمایش[۱۳] وجود داشته باشند، در حالی‌که آن کلمات در مجموعه آموزش[۱۴] وجود نداشته باشند، در کلاس‌بندی متن NLP چه رویکردی برای این‌گونه کلمات در نظر گرفته می‌شود و به‌چه‌صورت مسئله را حل می‌کنند؟ (۱۰)

ماشین بردار پشتیبان (۴۰)

۱. در جدول ۱، داده‌های مسئله به‌همراه برچسب کلاس آن‌ها و جواب مسئله دوگان SVM داده شده است.

(آ) با توجه به داده‌های جدول، ابرصفحه[۱۵] مدل SVM را به‌دست‌آورید. دقت کنید که منظور از ابرصفحه مدل SVM رابطه روبه‌رو است: $f(\boldsymbol{z}) = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{z} + b$. (راهنمایی: برای به‌دست‌آوردن مقدار b به‌ازای تمام نقاطی که در آن مقدار α_i برابر با صفر است، معادله متناظر را بنویسید و سپس از آن‌ها میانگین بگیرید.) (۲۰)

(ب) با استفاده از معادله‌ای که برای ابرصفحه SVM به‌دست‌آورده‌اید، کلاس دو نقطه $\boldsymbol{x} = (۳, ۳)^{\mathsf{T}}$ و $\boldsymbol{x} = (۳, ۲)^{\mathsf{T}}$ را مشخص کنید. (با دلیل) (۱۰)

(ج) نقاط جدول ۱ به‌همراه ناحیه حاشیه[۱۶] و ابرصفحه SVM را در یک صفحه رسم کنید. (۱۰)

---

[۱] Support Vector Machine
[۲] Naïve Bayes
[۳] Classification
[۴] Machine Learning
[۵] https://courses.aut.ac.ir/
[۶] Confusion Matrix
[۷] Accuracy
[۸] Precision
[۹] Recall
[۱۰] f1-score
[۱۱] Posterior
[۱۲] Natural Language Processing
[۱۳] Test Set
[۱۴] Train Set
[۱۵] Hyperplane
[۱۶] Margin

جدول ۱: داده‌ها و جواب مسئله دوگان SVM

ردیف	x_{i1}	x_{i2}	y_i	α_i
۱	۴	۲/۹	۱	۰/۴۱۴
۲	۴	۴	۱	۰
۳	۱	۲/۵	-۱	۰
۴	۲/۵	۱	-۱	۰/۰۱۸
۵	۴/۹	۴/۵	۱	۰
۶	۱/۹	۱/۹	-۱	۰
۷	۳/۵	۴	۱	۰/۰۱۸
۸	۰/۵	۱/۵	-۱	۰
۹	۲	۲/۱	-۱	۰/۴۱۴
۱۰	۴/۵	۲/۵	۱	۰

بیز ساده (۴۵)

۱. فرض کنید می‌خواهیم از کلاس‌بند بیز ساده برای دسته‌بندی افراد به دو دسته بیمار و سالم استفاده کنیم. مجموعه داده موجود در جدول ۲ اطلاعات برچسب‌دار در مورد فشارخون و دو ویژگی دیگر که مقادیر True و False می‌گیرند را برای ۸ نمونه نشان می‌دهد. فرض کنید که فشار خون یک متغیر پیوسته است که دارای توزیع گاوسی در هر کلاس است. با استفاده از کلاس‌بند بیز ساده، برچسب کلاس نمونه نهم را مشخص کنید (تمام مراحل را با جزئیات بیان کنید). (یادآوری: تابع توزیع گاوسی به‌صورت روبه‌رو می‌باشد: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-m)^2}{\sigma^2})$) (۲۰)

جدول ۲: مجموعه داده افراد بیمار و سالم

ردیف	فشار خون	ویژگی اول (F_1)	ویژگی دوم (F_2)	کلاس
۱	۱۱	F	T	سالم
۲	۱۲	T	F	بیمار
۳	۱۴	F	T	بیمار
۴	۱۳	F	F	سالم
۵	۱۰/۵	T	T	سالم
۶	۱۱/۵	T	F	بیمار
۷	۱۴/۵	F	F	بیمار
۸	۱۳/۵	T	T	سالم
۹	۳/۷۵	T	F	؟

۲. در جدول ۳ زیرمجموعه‌ای از نظرات کاربران در رابطه با یک محصول را مشاهده می‌کنید و برچسب هر کلاس نیز مشخص شده‌است.

(آ) با استفاده از بیز ساده و بدون استفاده از mـ تخمین احتمال[۱] برچسب نمونه ششم جدول ۳ را به‌دست‌آورید. (تمام مراحل را با جزئیات بیان کنید) (۱۵)

(ب) قسمت قبل را با استفاده از mـ تخمین احتمال دوباره تکرار کنید. (راهنمایی: در NLP مقدار m را برابر با تعداد کلمات منحصربه‌فرد[۲] یا همان تعداد واژگان[۳] (V) در مجموعه آموزش در نظر می‌گیرند، یعنی $m = |V|$. همچنین برای p معکوس اندازه واژگان را در نظر می‌گیرند.) (۱۰)

[۱] m-estimate of probability [۲] Unique [۳] Vocabulary

جدول ۳: نظرات کاربران

ردیف	نظر	برچسب
۱	بسیار ضعیف	–
۲	کاملاً قابل پیش‌بینی و فاقد انرژی است	–
۳	بسیار جذاب و دارای خنده بسیار است	+
۴	هیچ غافلگیری جالبی ندارد و خنده خیلی کمی دارد	–
۵	جالب‌ترین فیلم تابستان	+
۶	قابل پیش‌بینی و بدون خنده	؟

پیاده‌سازی (۸۰ + ۳۶)

در این قسمت قرار است با استفاده از کلاس‌بند بیز ساده، یک کلاس‌بندی چندکلاسه روی متن انجام گیرد. در ادامه به توضیحات مجموعه‌داده و جزئیات مختلف هر قسمت پیاده‌سازی می‌پردازیم. الگوریتم بیز ساده با استفاده از mـ تخمین احتمال در اسلاید ۴۴ تدریس‌شده درس وجود دارد.

نکته ۱: در این تمرین می‌توانید از ساختار کلاس و تابع در پایتون برای راحت‌تر پیاده‌سازی کردن هر قسمت و کاراتر بودن برنامه استفاده کنید.

نکته ۲: در هر بخش کتابخانه‌هایی که مجاز به استفاده از آن هستید، در ابتدای بخش بیان می‌شود. تنها می‌توانید از کتابخانه‌های گفته‌شده استفاده کنید یا کلا از کتابخانه‌ای استفاده نکنید و خودتان پیاده‌سازی را انجام دهید که توصیه نمی‌شود.

نکته ۳: فایل تحویلی باید در محیط نوت‌بوک Jupyter یا Google Colab نوشته شود. در واقع فایل تحویلی شما باید پسوند .ipynb داشته باشد.

مجموعه‌داده

کتابخانه مجاز: برای فراخوانی مجموعه‌داده از کتابخانه pandas استفاده کنید.

در فایل مربوط به مجموعه داده چهار فایل در اختیار شما قرار گرفته‌است: train.csv، test.csv، train_clean.csv و test_clean.csv. دو فایلی که با پسوند clean مشخص شده‌اند، فایل‌هایی هستند که پیش‌پردازش‌های لازم متن بر روی آن‌ها انجام شده‌است و می‌توانید برای ادامه کار از آن‌ها استفاده کنید. قسمت پیش‌پرداز متن و تمیز کردن فایل‌های خام آموزش و آزمایش اختیاری بوده و نمره امتیازی خواهد داشت. مجموعه‌داده اصلی دارای سه ستون شامل Id، Comment و Topic است. ستون Comment داده متنی و ستون Topic نشان‌دهنده برچسب کلاس داده است. در مجموعه‌داده تمیزشده، ستون Id حذف شده‌است و ستون دیگری که متن‌های تمیزشده در آن قرار دارد، اضافه شده‌است. دقت داشته باشید اگر خودتان مراحل پیش‌پردازش را انجام می‌دهید، حتما باید به فایل تمیزشده‌ای که به شما تحویل داده‌شده‌است برسید.

پیش‌پردازش متن (امتیازی) (۳۶)

کتابخانه‌های مجاز: ۱. string، ۲. re و ۳. unidecode

مراحل پیش‌پردازش

۱. مقابله کردن با داده‌های گم‌شده[1]: حذف سطرهایی که متن آن‌ها موجود نمی‌باشد.

۲. حذف متن‌هایی که تعداد کلماتشان کمتر یا مساوی پنج است.

۳. حذف نشانگر خط جدید (\n) در همه نمونه‌ها

۴. حذف لینک‌های موجود در متن

۵. تبدیل عبارت â€™ به علامت '

[1] Missing Value

۶. تبدیل عبارت â€¦ به علامت !

۷. تبدیل نمادهای غیرانگلیسی به حروف انگلیسی

۸. تبدیل فرم‌های اختصار[1] به فرم‌های رسمی، به‌طور مثال تبدیل I'm به I am

۹. حذف علائم نگارشی (هر چیزی به‌غیر از حروف الفبا و اعداد)

۱۰. حذف چندین فاصله و فاصله‌های ابتدا و انتهای هر متن

۱۱. تبدیل حروف بزرگ به کوچک

۱۲. اطمینان یافتن از خالی نشدن یک متن!

آماده‌سازی متن برای ورودی الگوریتم (۲۳)

کتابخانه مجاز: در این قسمت هیچ کتابخانه‌ای مجاز نیست.

از آن‌جایی‌که در روش بیز ساده به احتمال و در نتیجه تعداد رخداد هر ویژگی (در اینجا کلمه) نیاز داریم، در ابتدا باید تعداد کلمات را به‌دست‌آورید. کلمات منحصربه‌فرد مجموعه‌های آموزش و آزمایش و تعداد تکرار هر کدام از کلمات منحصربه‌فرد را در مجموعه‌های آموزش و آزمایش به‌دست‌آورید. (برای این‌کار می‌توانید از ساختار دیکشنری در پایتون استفاده کنید.) همچنین تعداد کل کلمات موجود (تمام کلمات نه تنها کلمات منحصربه‌فرد) در مجموعه‌های آموزش و آزمایش را به‌دست‌آوردید. در فایل کد برای هر کدام از مجموعه‌های آموزش و آزمایش تعداد کلمات منحصربه‌فرد و تعداد کل کلمات را چاپ کنید و آن‌را در گزارش خود نیز بیاورید. علاوه بر آن ۱۰۰ کلمه منحصربه‌فرد پرتکرار مجموعه آموزش را در یک فایل ذخیره کنید و به‌همراه فایل کد و گزارش ارسال کنید. دقت داشته باشید نام قایل را Frequency_tokens.txt قرار دهید و در هر خط کلمه و با یک فاصله تعداد تکرار آن قرار گرفته باشد. روند پیاده‌سازی کردن این قسمت را در گزارش خود بیاورید. (۸)

در ادامه باید هر کلمه منحصربه‌فرد مجموعه آموزش را به یک عدد نگاشت کنیم. برای این‌کار نیز می‌توانید از ساختار دیکشنری استفاده کنید. برای این‌کار ابتدا دو دیکشنری Token2Index و Index2Token بسازید. اولین دیکشنری هر کلمه را به یک عدد نظیر می‌کند و دومین دیکشنری برعکس این‌کار را انجام می‌دهد. دقت داشته باشید که اعداد متناظر با کلمات منحصربه‌فرد نیز باید منحصربه‌فرد باشد. در نوتبوک تحویلی و گزارش خود ۵ نمونه از دیکشنری‌های ساخته‌شده خود را چاپ کنید. دقت داشته باشید که در ادامه و همچنین هنگام آموزش مدل به دیکشنری Token2Index مجموعه آموزش نیاز دارید، پس حتما آن‌را ذخیره کنید. (۸)

در نهایت برای هر متن موجود در مجموعه‌داده آموزش و آزمایش که یک رشته است را به لیستی از کلمات تبدیل کرده و در نهایت برای هر متن، لیستی از اعداد نظیر شده به‌هر کلمه به‌دست‌آمده از مرحله قبل را ذخیره کنید. نمونه‌ای از این‌کار را برای مجموعه آموزش و نمونه‌ای را برای مجموعه آزمایش در نوت‌بوک تحویلی چاپ کرده و در فایل گزارش خود نیز بیاورید. (۷)

آموزش مدل بیز ساده برای متن (۴۵)

کتابخانه مجاز: تنها کتابخانه مجاز در این بخش، کتابخانه math است.

توصیه می‌شود این بخش را به‌صورت تابع برای هر مرحله پیاده‌سازی کنید.

ابتدا تعداد تکرار هر کلمه منحصربه‌فرد را در هر کلاس به‌صورت جداگانه به‌دست‌آورید. (۹)

سپس لگاریتم ۲ احتمال رخداد هر کلمه در هر کلاس را محاسبه کنید. برای جلوگیری از صفر شدن احتمالات از روش mـ تخمین احتمال استفاده کنید. (۱۲)

پس از آن لگاریتم ۲ احتمال رخداد هر کلاس را بر اساس مجموعه‌داده محاسبه کنید. (۱۲)

در نهایت، برای هر متن، پیش‌بینی تعلق به کلاس را انجام دهید. برای این‌کار لگاریتم ۲ احتمال پسین را محاسبه کنید و براساس لگاریتم ۲ احتمال به‌دست‌آمده برای هر متن به‌ازای هر کلاس، کلاس تخصیص یافته به متن را مشخص کنید. در نوت‌بوک تحویلی ۳ نمونه از این مرحله را به‌صورت زیر چاپ کنید و در گزارش خود نیز بیان کنید: (۱۲)

سطر اول: متن

سطر دوم: لیستی از مقادیر لگاریتم احتمال در هر کلاس

سطر سوم: کلاس اصلی متن و کلاس پیش‌بینی‌شده توسط مدل

---

[1] Contraction

ارزیابی مدل (۱۲)

کتابخانه مجاز: تنها در این قسمت می‌توانید از کتابخانه sklearn و مترهای موجود در این کتابخانه استفاده کنید.
معیارهای دقت، صحت، پوشش و امتیاز f1 را برای هر دو مجموعه آموزش و آزمایش به‌دست‌آورید و مقادیر به‌دست‌آمده را تحلیل کنید. یعنی تحلیل خود را از ارزیابی مدل بنویسید.

نکات تحویل تمرین

۱. گزارش کد و جواب‌های سؤالات تشریحی حتماً باید به‌صورت **تایپ** شده باشند. تنها اگر پاسخی از سؤالات تشریحی دارای فرمول‌های زیاد ریاضی بود که شاید تایپ آن‌ها برای شما سخت باشد، تنها پاسخ آن سؤال می‌تواند به‌صورت دستی باشد. پاسخ‌هایی که به‌صورت دستی هستند نیز باید به‌صورت مرتب، خوانا و بدون خط خوردگی باشد.

۲. مشاهدات و استنباط‌های خود را به‌صورت دقیق بیان کنید.

۳. در گزارش‌نویس کد و سؤالات تشریحی، پاسخ باید واضح و بدون هیچ‌گونه ابهامی باشد. نکته مهم، پاسخ واضح است و نه حجم زیاد پاسخ! اگر فرضی برای حل مسئله استفاده می‌کنید، حتما آن‌را ذکر کنید.

۴. هر گونه شباهت در پاسخ تشریحی و گزارش و کد، به منزله تقلب می‌باشد و نمره کل تمرین برابر صفر خواهد شد.

۵. اگر برای حل مسائل خود، از منبعی استفاده می‌کنید، حتما به آن ارجاع دهید.

۶. گزارش کد و پاسخ تشریحی در کنار هم در قالب یک فایل pdf باشد.

۷. فایل pdf، فایل نوت‌بوک و فایل Frequency_tokens.txt را به‌صورت یک‌جا در قالب یک فایل zip در سامانه آپلود کنید.