توضیحات مهم:

- پروژه از طریق تماس اسکایپی تحویل گرفته می‌شود، زمانبندی مربوطه اعلام خواهد شد.
- تمامی مستندات خود شامل گزارش و کدهای خود را در یک فایل فشرده با فرمت zip ذخیره کرده و با عنوان studentID_Project.zip بارگذاری نمایید (به عنوان مثال 99131000_Project.zip).
- [redacted]
- استفاده از کتابخانه‌های رایج در یادگیری ماشین بلا مانع است.
- [redacted]
- ملاک اصلی انجام پروژه گزارش آن است و ارسال کد بدون گزارش فاقد ارزش است. یک فایل pdf تهیه کرده و برای هر سوال، ورودی، خروجی و توضیحات مربوطه را بصورت جامع گزارش کنید.
- تا حد ممکن سعی کنید اصول لازم برای گزارش مهندسی را رعایت نمایید. (به بهترین گزارش نمره‌ی تشویقی تعلق می‌گیرد.)
- [redacted]

بخش اول: دسته‌بندی تصاویر

مجموعه‌داده Fashion MNIST از ۶۰ هزار تصویر آموزش و ۱۰ هزار تصویر ارزیابی تشکیل شده‌است. هر تصویر در این مجموعه‌داده دارای ابعاد ۲۸×۲۸ است و متعلق به یکی از ۱۰ کلاس مختلف پوشاک می‌باشد. در تنسورفلو می‌توانید با دستور زیر مجموعه‌داده را فراخوانی و به دو دسته آموزش و آزمون تقسیم کنید.

```
fashion_mnist = tf.keras.datasets.fashion_mnist
(train_images, train_labels),(test_images, test_labels)=fashion_mnist.load_data()
```

همچنین عنوان کلاس‌های این مجموعه‌داده به شرح زیر است.

Label	Class
0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot

در این سوال با برخی از مسائل مربوط به دسته‌بندی تصاویر با شبکه‌های عصبی آشنا خواهید شد.

الف) از هر کلاس از دیتاست، به تصادف یک تصویر را انتخاب و در فایل گزارش نمایش دهید.

ب) یک شبکه عصبی تماما متصل[1] را با داده‌های آموزشی، آموزش دهید و دقت داده‌های ارزیابی را بر روی آن گزارش دهید.

ج) قسمت ب را به ازای تعداد لایه‌ها و تعداد نرون‌های مختلف در هر لایه تکرار و نتایج را گزارش و تفسیر کنید.(حداقل ۶ آزمایش)

د) نمودار تابع هزینه[2] و دقت در دو مجموعه آموزش و ارزیابی را برای بهترین شبکه قسمت ج ارائه و تفسیر کنید.

[1] fully connected
[2] loss function

ه) پس از آموزش کامل شبکه و ساخت مدل نهایی، تصاویر داده‌های ارزیابی را چهار پیکسل به سمت بالا شیفت دهید (شیفت حلقوی با دستور پیشنهادی numpy.roll). داده‌های ارزیابی جدید را مجددا بر روی شبکه آموزش دیده ارزیابی و نتیجه را تفسیر کنید.

و) تحقیق کنید که شبکه‌های کانولوشنی[3] چگونه از بروز مشکل پیش آمده جلوگیری می‌کند. (راهنمایی: translation invariant)

بخش دوم: حل یک مسأله واقعی

(Bureau of Meteorology)، یک سازمان اجرایی وابسته به دولت استرالیا است که اطلاعات مربوط به وضعیت آب‌وهوایی این کشور را ارائه می‌کند. یکی از مهم‌ترین وظایف این سازمان جمع‌آوری و ذخیره سوابق وضعیت هواشناسی در ایستگاه‌های مختلف است. در این سوال شما با کمک بخشی از این اطلاعات وضعیت بارندگی را پیش‌بینی خواهید کرد.

مجموعه‌داده این سوال شامل اطلاعات روزانه آب‌وهوایی در ایستگاه‌های مختلف کشور استرالیا بین سال‌های ۲۰۰۸ و ۲۰۱۷ می‌باشد. هر ردیف شامل چندین متغیر عددی و دسته‌ای (categorical) می‌باشد. همچنین مقدار برخی متغیرها در برخی روزها نامعلوم (null) می‌باشد. متغیر هدف در این مجموعه-داده RainTomorrow است که نشان می‌دهد آیا روز بعد باران خواهد بارید؟

این مجموعه‌ی داده را می‌توانید از طریق لینک زیر دانلود کنید.

https://drive.google.com/file/d/1uBXj5fGE1i3f7RqFhitoPPaaI2gjc967/view?usp=sharing

همچنین برای دانلود در پلتفرم کولب می‌توانید از دستور زیر استفاده کنید.

```
!gdown --id 1uBXj5fGE1i3f7RqFhitoPPaaI2gjc967
```

الف) متد info(). از کتابخانه pandas تعداد داده‌های non-null برای هر متغیر را نشان می‌دهد. استفاده از این متد گام ابتدایی مناسبی برای حل مسئله است. همانطور که می‌بینید در بعضی روزها مقدار متغیر هدف نامشخص می‌باشد و بنابراین استفاده از آن‌ها برای آموزش یا ارزیابی ممکن نیست. این روزها را از مجموعه‌داده حذف کنید. برای سایر مقادیری که مقدار آن‌ها null است، از روش‌های مناسب استفاده کنید.(راهنمایی)

[3] convolutional neural network

ب) برای قسمت‌های بعدی مجموعه‌داده را پس از شافل کردن، به سه مجموعه train، test و validation با نسبت های معقول تقسیم کنید.

ج) برای جلوگیری از بیش‌برازش مدل‌های یادگیری ماشین پیشنهاد می‌شود تعداد متغیرهای مسئله را کاهش دهید. یک راه مناسب، حذف یکی از جفت متغیرهایی است که correlation بالایی با متغیر دیگر دارد. آیا در این دیتاست جفت متغیرهای وجود دارد که دارای ضریب همبستگی بیش از ۰.۹۵ باشند؟ در صورت وجود، آیا وجود correlation با این مقدار بالا بین این جفت متغیرها توجیهی دارد؟ در صورت وجود، یکی از هر جفت را به دلخواه حذف کنید؟

د) همبستگی پایین با سایر متغیرها شرط کافی برای مناسب بودن یک متغیر نیست. به طور دلخواه با استفاده از یک روش انتخاب ویژگی، تعداد ویژگی‌های دیتاست را کاهش دهید.(انتخاب تعداد ویژگی به عهده خودتان است)

ه) یک مدل شبکه عصبی سه لایه با داده‌های کاهش یافته آموزش دهید و روی داده های تست ارزیابی کنید.

بخش سوم: آشنایی با تحلیل داده‌های متنی

مسئله تعریف شده در این پروژه در رابطه با تجزیه و تحلیل احساسات[4] 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.

مجموعه‌دادگانی که در اختیار شما قرار می‌گیرد، از۱۲ هزار متن تشکیل شده است که هرکدام دارای برچسب‌های +۱، +۲، ۰، -۱ و -۲ هستند. در این فاز شما باید از این مجموعه‌داده برای آموزش مدل‌هایتان استفاده کنید و مجاز به استفاده از هیچ مجموعه‌داده کمکی برای بهبود نتیجه‌هایتان نیستید. استفاده از هر مجموعه‌داده کمکی تخلف در نظر گرفته می‌شود. ارزیابی نهایی مدل‌های شما روی یک مجموعه‌داده مشابه از نظرات کاربران انجام

[4] Sentiment Analysis

می‌شود. این مجموعه‌داده در زمان انجام پروژه در اختیار شما قرار نـمی‌گیرد؛ بنابراین باید با استفاده از تکنیک‌های Validation و توجه به Overfit نـشدن مدل‌ها روی دادگان آموزش و حتی دادگان اعتبارسنجی، تلاش کنید مدل‌هایی با توان تعمیم پذیری بالا ارایه دهید.

این داده‌ها دارای دو ستون text و polarity هستند. ستون polarity دارای مقادیر 2 ,1 ,0 ,2- ,1- است که منفی یا مثبت بودن بار معنایی text را مشخص می‌کند. مجموعه‌ی داده را می‌توانید از طریق لینک زیر دانلود کنید.

https://docs.google.com/spreadsheets/d/1-9Lb3Q6x7BfYOfmcAmtDHLJ-nuswHxIQ/edit?usp=sharing&ouid=105359388001754195649&rtpof=true&sd=true

همچنین برای دانلود در پلتفرم کولب می‌توانید از دستور زیر استفاده کنید.

```
!gdown --id 1-9Lb3Q6x7BfYOfmcAmtDHLJ-nuswHxIQ
```

گام اول: تولید داده‌های مناسب و قابل پردازش

معمولا داده‌های متنی خام، بی نظمی و یا حالات خاصی را دارند که تحلیل آن‌ها را دشوار می‌کند. به همین علت برای استفاده از آن‌ها، ابتدا پیش پردازش‌هایی بر روی‌آن‌ها انجام می‌شود. همچنین برای استفاده از بیشتر مدل‌های یادگیری ماشین نیاز داریم که داده‌هایمان در قالب 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.

پیش پردازش 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 پیاده سازی این روش‌ها معمولا می‌توان از کتابخانه‌های آماده‌ای نظیر hazm در python کمک گرفت. روش‌های ابتدایی حذف اعداد یا تبدیلشان به حروف، حذف کاراکترهای اضافی و جداسازی کلمات از هم است. از پیش‌پردازش‌های سطح بالاتر می‌توان به حذف کلمات ایست[5] و انجام ریشه‌یابی[6] و بن‌واژه‌سازی[7] روی کلمات بدست آمده اشاره کرد.

چنین پردازش‌هایی به این منظور انجام می‌گیرند که خیلی از کلمات و اجزای متن لزوما ویژگی‌های مناسبی نـــیستند و اطلاعات اضافی را به همراه ندارند. به عنوان مثال "ها" در کلمه "کتاب‌ها" اطلاعات خاصی را به همراه

[5] Stop words
[6] Stemming
[7] Lemmatization

ندارد، به همین خاطر تمام کلمات را ریشه‌یابی می‌کنیم، یا مثلا کلمات پرتکراری همانند "است"، "از"، "در"، "به" و ... در بیشتر جملات وجود دارند و در نتیجه ویژگی خوبی محسوب نمی‌شوند.

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

گام دوم: تبدیل به بردار

پس از مرحله پیش پردازش، نوبت به اختصاص برداری عددی به هر نمونه متنی می‌رسد. برای این منظور از روش هایی 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.

گام سوم: مسئله‌ی دسته‌بندی

در این بخش ابتدا می‌خواهیم تاثیر پیش‌پردازش متن‌ها را روی عملکرد یک مدل شبکه عصبی تماما متصل بسنجیم. به این منظور این مدل شبکه‌ی عصبی را با لایه‌ها و نورون‌های مختلف آموزش داده و نتایج را مقایسه کنید. برای بردارسازی از روش Bag of words استفاده کنید. برای این کار می‌توانید از کد زیر که مربوط به کتابخانه sklearn است استفاده کنید:

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer()
train_vectors = vectorizer.fit_transform(train_data_samples)
test_vectors = vectorizer.transform(test_data_samples)
```

لینک زیر را برای استفاده بهتر از TfidfVectorizer مطالعه کنید. از طریق تنظیم پارامترهای مناسب، می‌توانید برخی از کارهای پیش‌پردازش را به سادگی انجام دهید.

https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

حالت اول: بدون پیش پردازش متن

حالت دوم: پیش پردازش متن شامل روش‌های مختلفی که در گام اول به آن‌ها اشاره شد (همچنین روش‌های بیشتری را با جست‌وجو در منابع موجود می‌توان یافت، که می‌توانید برای بهتر کردن دقت، از آن‌ها نیز استفاده کنید.).

بخش چهارم: تعریف یک مسئله‌ی مرتبط با رشته و گرایش تحصیلی

یک مسئله‌ی یادگیری ماشین(دسته بندی یا رگرسیون) که مرتبط با رشته یا گرایش تحصیلی و یا به طور خاص پروژه‌ی پایان‌نامه‌ی شما باشد، تعریف کنید. بعد از تعریف کامل مسئله و چالش‌های آن، راهکار(ها)یی برای حل این مسئله ارائه داده و آن(ها) را پیاده‌سازی کنید. در راستای حل مسئله، یک مجموعه‌داده که شامل تعداد قابل توجهی نمونه برای آموزش مدل باشد، انتخاب کنید. این مجموعه داده باید همراه با یکسری چالش‌ها باشد که با پیش‌پردازش بتوان این چالش‌ها را مدیریت کرد و یا از میان برداشت، علت و نحوه‌ی انجام هر پیش‌پردازشی که در این راستا انجام می‌دهید را توضیح دهید و میزان اثربخشی پیش پردازش‌ها را در نتیجه‌ی نهایی بررسی کنید.

- توجه داشته باشید که تعداد راهکارهایی که برای حل مسئله ارائه می‌دهید و میزان ابتکاری که در این راهکارها وجود دارد می‌تواند نمره‌ی امتیازی برای شما به همراه داشته باشد.
- دقت کنید که در فایل گزارش، تنها بهترین راهکار را ذکر نکنید، بلکه راهکارهای مختلف آزمایش‌شده را که در نهایت از راه‌حل‌های دیگر ضعیف‌تر عمل کردند، همراه با نتیجه آزمایش‌ها بیاورید.
- مجموعه‌ی داده‌ها و نتایج آزمایشات را برای درک بهتر تصویرسازی کنید.
- می‌توانید از لینک زیر برای تعریف بهتر مسئله و انجام گام به گام راه حل کمک بگیرید.

https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

بخش امتیازی

در راستای مسئله‌ای که تعریف کرده اید، مقاله‌ای را جستجو کرده و کار خود را با آن مقاله مقایسه کنید. با ذکر شباهت‌ها و تفاوت‌های دو روش، مزایا و معایب هر روش را توضیح دهید. اگر مجموعه‌ی داده توسط شما یا آزمایشگاهی که در آن عضو هستید، به تازگی جمع‌آوری شده و مقاله‌ای روی آن چاپ نشده است، مقاله‌ای را انتخاب کنید که روی مجموعه‌داده‌ای مشابه با کار شما انجام شده باشد.