به نام خالق یکتا

تمرین سوم مبانی یادگیری ماشین

پاییز 1402

بخش تشریحی

سوال اول)

جدول زیر شامل اطلاعات یک مجموعه داده مربوط به تشخیص گربه و سگ است. همانطور که می‌بینید سه ویژگی اصلی برای این مجموعه داده در نظر گرفتیم. این سه ویژگی اصلی قد، وزن و طول بدن حیوان می‌باشد. با توجه به این مجموعه داده و اطلاعات داده شده از شما می‌خواهیم به پرسش‌های زیر پاسخ دهید. (مقدار k را برای بخش اول برابر 5 در نظر بگیرید.)

الف) تشخیص دهید که با استفاده از الگوریتم K nearest neighbors نمونه زیر سگ یا گربه است؟

کلاس	قد	وزن	طول بدن حیوان
؟؟؟؟؟	15	5	25

ب) مقدار k را کم و زیاد کنید و تاثیر ان را در پاسخ مسئله توضیح دهید.

کلاس	قد	وزن	طول بدن حیوان
گربه	10	2	20
گربه	15	3	30
سگ	15	5	60
گربه	21	12	55
سگ	30	24	54
سگ	50	25	60
گربه	12	2.75	23
سگ	17	5.75	33
گربه	16	4	34
گربه	10.5	13	35

سوال دوم)

فرض کنید در حال تلاش برای یادگیری درخت تصمیم هستیم. داده‌های ورودی ما شامل N نمونه است که هر کدام دارای k ویژگی (N>>k) هستند. ما عمق یک درخت را حداکثر تعداد گره‌های بین ریشه و هر یک از گره‌های برگ (شامل برگ، نه ریشه) تعریف می‌کنیم.

الف) اگر همه صفات باینری باشند، حداکثر تعداد گره‌های برگ (تصمیم) که می‌توانیم در درخت تصمیم برای این داده داشته باشیم چقدر است؟ حداکثر عمق ممکن درخت تصمیم برای این داده چقدر است؟

ب) اگر همه صفات پیوسته باشند، حداکثر تعداد گره‌های برگ که می‌توانیم در درخت تصمیم برای این داده‌ها داشته باشیم چقدر است؟ حداکثر عمق ممکن برای درخت تصمیم در این کار چقدر است؟

ج) فرض کنید یک مجموعه اعتبارسنجی به صورت زیر داریم. خطای مجموعه آموزشی و خطای مجموعه اعتبارسنجی درخت چه خواهد بود؟ پاسخ‌های خود را به عنوان تعداد نمونه‌هایی که به اشتباه طبقه بندی می‌شوند، بیان کنید .

سوال سوم)

مجموعه داده زیر برای یادگیری درخت تصمیم برای پیش بینی خوراکی بودن یا نبودن قارچ بر اساس شکل، رنگ و بو استفاده می‌شود.

Shape	Color	Odor	Edible
C	B	2	No
D	B	2	No
C	W	2	Yes
C	B	1	YES
D	B	1	YES
D	W	1	YES
D	W	2	YES
C	B	2	YES
D	B	2	NO
D	G	2	NO
C	U	2	NO
C	B	3	NO
C	W	3	NO
D	W	3	NO

الف) انتروپی را H(Edible| Order =1 or Odor=3) به دست اورید.

ب) الگوریتم حریصانه بالا به پایین (ID3) کدام ویژگی را برای استفاده از ریشه درخت (بدون هرس) انتخاب می‌کند؟

ج) درخت تصمیم کاملی را که برای این داده‌ها آموخته می‌شود، رسم کنید (بدون هرس).

سوال چهارم)

شبکه‌ی عصبی را در نظر بگیرید که در آن هر نورون تابع فعال‌سازی خطی دارد، یعنی خروجی هر نورون $\frac{1}{n}\sum_{i=1}^{n} W_i\, x_i$ g(x)=c+b است، که در آن b و c دو عدد حقیقی ثابت و n تعداد لینک‌های ورودی به آن نورون است.

الف) فرض کنید یک نورون منفرد با تابع فعالسازی خطی g() مانند بالا و ورودی $x_0,..,x_n$ و وزن‌های $W_0,...,W_n$ دارید. اگر خروجی واقعی یک عدد اسکالر y باشد، تابع خطای مربع را برای این ورودی بنویسید، سپس قانون به روزرسانی وزن نورون را بر اساس نزول گرادیان روی این تابع خطا بنویسید.

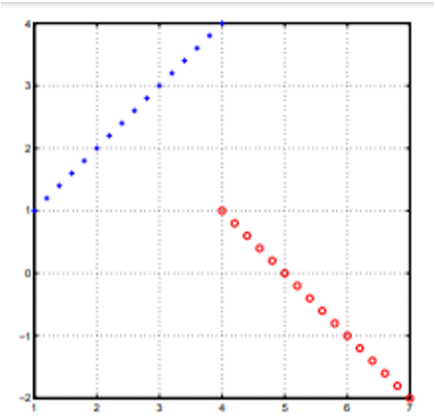
ب) اکنون شبکه‌ای از نورون‌های خطی با یک لایه پنهان از m واحد، n واحد ورودی و یک واحد خروجی را در نظر بگیرید. برای یک مجموعه معین از وزن‌ها $w_{k,j}$ در لایه پنهان ورودی و W_j در لایه خروجی پنهان، معادله واحد خروجی را به عنوان تابعی از $w_{k,j}$، W_j و ورودی x یادداشت کنید. نشان دهید که یک شبکه خطی تک لایه بدون واحدهای پنهان وجود دارد که همان تابع را محاسبه می‌کند.

سوال پنجم)

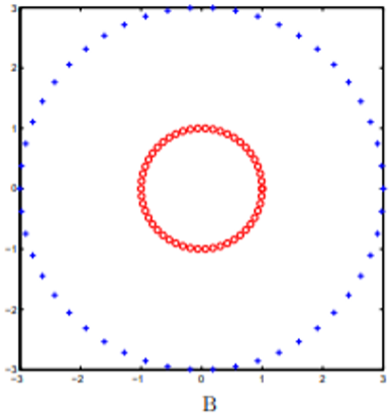
مجموعه داده toydata1 را در شکل 1 و toydata2 را در شکل 2 در نظر بگیرید.

در هر یک از این مجموعه داده‌ها دو کلاس '+' و 'o' وجود دارد .هر کلاس تعداد نمونه‌های یکسانی دارد. هر نقطه داده دارای دو ویژگی با مقادیر حقیقی است، یعنی مختصات X و Y.

برای هر یک از این مجموعه داده‌ها، مرز تصمیم‌گیری را که طبقه‌بندی‌کننده Gaussian Naive Bayes یاد می‌گیرد، ترسیم کنید. (به یاد داشته باشید که همه کلاس‌ها دارای تعداد یکسانی هستند و بنابراین ما نباید نگران Prior باشیم.)



شکل ۱ - مجموعه داده toydata1



شکل ۲ - مجموعه داده toydata2

بخش پیاده سازی

سوال اول)

1. مقدمه

در این سوال، هدف استفاده از درخت تصمیم برای طبقه بندی گل ها در دیتا ست آیریس و مقایسه عملکرد آن با استفاده از روش KNN از تمرین قبل است. می‌توانید از پیاده سازی های آماده درخت تصمیم برای این سوال استفاده کنید.

3. ارزیابی و آپلود

معیار محاسبه accuracy را به همراه confusion matrix گزارش کنید.

در انتها نتایج مدل خود را در فایلی به فرمت csv ذخیره کنید. این فایل باید به تعداد سطرهای داده ی تست پیشبینی داشته باشد که در یک ستون با نام Species ذخیره شده است. فایل نتایج را به علاوه ی کد در پوشه ی مربوط به سوال در ریپازیتوری گیت هاب قرار دهید.

سوال دوم)

1. مقدمه

کلاسیفایر بیزی: این نوع از الگوریتم‌های یادگیری ماشین بر اساس قاعده بیز کار می‌کنند. این قاعده رابطه‌ای ریاضی است که احتمال وقوع یک رویداد را با توجه به اطلاعات پیشین محاسبه می‌کند. شما در این تمرین می بایست با استفاده از sklearn، یک مدل multiclass generative bayes classifier را پیاده سازی کنید.

2. معرفی مجموعه داده

این دیتاست شامل اطلاعات مربوط به محصولات موجود در یک پلتفرم فروش آنلاین است. ستون‌ها در این دیتاست عبارتند از:

- Product ID: شناسه منحصر به فرد هر محصول.
- Product Title: نام یا عنوان محصول، که می‌تواند شامل اطلاعاتی در مورد مدل، ویژگی‌ها، رنگ، و غیره باشد.
- Merchant ID: شناسه فروشنده یا تامین‌کننده محصول.
- Cluster ID: شناسه خوشه‌بندی شده برای محصولات، که احتمالا برای گروه‌بندی محصولات مشابه با یکدیگر استفاده می‌شود.
- Cluster Label: برچسب خوشه، که نشان‌دهنده دسته‌بندی یا طبقه‌بندی خاصی از محصولات است.
- Category ID: شناسه دسته‌بندی محصول.
- Category Label: برچسب دسته‌بندی محصول، که دسته بندی کلی محصول را مشخص می‌کند، مانند " Mobile Phones".

3. استفاده از دیتاست در کلاسیفایر بیز مولد چندکلاسه

در این تمرین، با استفاده از این داده‌ها میبایست با استفاده از عنوان محصول و اطلاعات دسته‌بندی، یک مدل بسازید که قادر به پیش‌بینی دسته‌بندی (مانند "Mobile Phones") برای محصولات جدید باشد.

4. ارزیابی و آپلود

معیار محاسبه accuracy را به همراه confusion matrix

در انتها نتایج مدل خود را در فایلی به فرمت. csv ذخیره کنید. این فایل باید به تعداد سطرهای داده ی تست پیشبینی داشته باشد که در یک ستون با نام Category Label ذخیره شده است. فایل نتایج را به علاوه ی کد در پوشه ی مربوط به سوال در ریپازیتوری گیت هاب قرار دهید.

سوال سوم)

1. مقدمه

در این پروژه، هدف ما توسعه یک سیستم تشخیص اسپم است که قادر به تمیز دادن ایمیل‌های اسپم از ایمیل‌های عادی (غیر اسپم) با استفاده از روش کلاسیفایر بیز ساده است. این روش بر پایه اصول احتمالاتی بیز و فرض استقلال ویژگی‌ها (کلمات در متن) عمل می‌کند.کلاسیفایر بیز ساده به دلیل سادگی و کارایی بالا در تشخیص دسته‌بندی متن‌ها، به ویژه در تشخیص اسپم، بسیار محبوب است. این روش با فرض اینکه ویژگی‌های مختلف (در این مورد، کلمات) به طور مستقل از هم بر نتیجه تأثیر می‌گذارند، عمل می‌کند و از این رو برای مجموعه داده‌های بزرگ و پیچیده مناسب است.سازی کنید. شما در این تمرین می بایست با استفاده از sklearn، یک مدل Naïve Bayes را پیاده سازی کنید.

2. معرفی مجموعه داده

این دیتاست دیتاست spam_ham_dataset.csv که برای تشخیص اسپم استفاده می‌شود، شامل چند ستون اصلی است:

- label: این ستون برچسب‌های دستی را نشان می‌دهد که نشان می‌دهد آیا یک ایمیل اسپم است یا خیر ('ham' برای ایمیل‌های عادی و 'spam' برای ایمیل‌های اسپم).
- text: این ستون متن ایمیل را شامل می‌شود. این داده‌های متنی اصلی هستند که برای تجزیه و تحلیل و آموزش مدل کلاسیفایر بیز ساده استفاده می‌شوند.
- Label_num: این ستون برچسب‌های عددی را نشان می‌دهد، که معمولاً 0 برای 'ham' و 1 برای 'spam' است. این ستون برای آموزش مدل‌های یادگیری ماشین که نیاز به برچسب‌های عددی دارند، مفید است.

3. استفاده از دیتاست در کلاسیفایر بیز ساده

برای استفاده از این دیتاست در یک پروژه کلاسیفایر بیز ساده، مراحل زیر دنبال می‌شوند:

- پیش‌پردازش متن: متن ایمیل‌ها باید پیش‌پردازش شود که شامل تبدیل به حروف کوچک، حذف نشانه‌های نگارشی و حذف کلمات بی‌معنی (stop words) است.
- استخراج ویژگی‌ها: متن پیش‌پردازش شده به ویژگی‌های عددی تبدیل می‌شود. این معمولاً با استفاده از مدل "کیسه کلمات" (Bag of Words) یا TF-IDF انجام می‌شود.
- تقسیم داده‌ها: دیتاست به دو بخش آموزش و تست تقسیم می‌شود.
- آموزش مدل کلاسیفایر بیز ساده: با استفاده از داده‌های آموزش، مدل کلاسیفایر بیز ساده آموزش داده می‌شود.
- ارزیابی مدل: با استفاده از داده‌های تست، عملکرد مدل ارزیابی می‌شود.

4. ارزیابی و آپلود

معیار محاسبه accuracy را به همراه confusion matrix

در انتها نتایج مدل خود را در فایلی به فرمت csv ذخیره کنید. این فایل باید به تعداد سطرهای داده ی تست پیشبینی داشته باشد که در یک ستون با نام Spam ذخیره شده است. فایل نتایج را به علاوه ی کد در پوشه ی مربوط به سوال در ریپازیتوری گیت هاب قرار دهید.

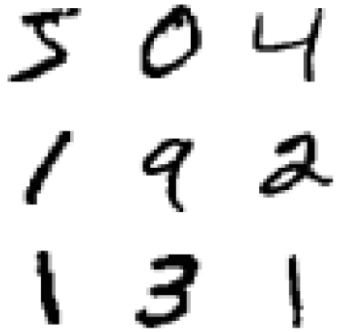
سوال چهارم)

1. مقدمه

در این سوال، هدف پیاده‌سازی شبکه عصبی پیشخور برای پیش بینی برچسب متناظر با تصاویر دیتاست MNIST می باشد. در این تمرین شما می بایست با استفاده از یکی از دو فریم ورک Pytorch یا Tensorflow یک شبکه عصبی متناسب با دیتاست MNIST طراحی کنید.

2. معرفی مجموعه داده

دیتاست MNIST مجموعه‌ای از تصاویر دست‌نویس اعداد است که به طور گسترده‌ای برای آموزش و آزمایش در زمینه یادگیری ماشین و پردازش تصویر استفاده می‌شود. این دیتاست شامل ۷۰,۰۰۰ تصویر است که ۶۰,۰۰۰ تصویر برای آموزش و ۱۰,۰۰۰ تصویر برای آزمایش در نظر گرفته شده است. هر تصویر در این دیتاست نمایانگر یک عدد دست‌نویس از ۰ تا ۹ است و اندازه هر تصویر ۲۸ در ۲۸ پیکسل است.



تصاویر در MNIST به صورت سیاه و سفید (خاکستری) هستند و هر پیکسل در این تصاویر مقداری بین ۰ تا ۲۵۵ دارد که نشان‌دهنده شدت رنگ است. این دیتاست به دلیل سادگی و اندازه مناسب برای تحقیقات و توسعه مدل‌های یادگیری عمیق و هوش مصنوعی محبوب است. MNIST به عنوان یک "هلو ورلد" در زمینه یادگیری ماشین شناخته شده و برای کسانی که تازه وارد این حوزه می‌شوند، معرفی خوبی است.

برای کار با این دیتاست شما می توانید به شکل دستی از سایت کگل یا گیت هاب از طریق لینک های زیر دانلود کرده یا به شکل مستقیم با استفاده از فریم ورک های Tensorflow یا Pytorch از آن استفاده کنید.

https://drive.google.com/file/d/11ZiNnV3YtpZ7d9afHZg0rtDRrmhha-1E/view

https://www.kaggle.com/datasets/hojjatk/mnist-dataset

در صورت دانلود مستقیم این دیتاست لازم به ذکر است که فورمت دیتاست را از idx3-ubyte با استفاده از کد زیر تغییر دهید.

```python
import numpy as np
import struct

def read_idx(filename):
    with open(filename, 'rb') as f:
        zero, data_type, dims = struct.unpack('>HBB', f.read(4))
        shape = tuple(struct.unpack('>I', f.read(4))[0] for d in range(dims))
        return np.frombuffer(f.read(), dtype=np.uint8).reshape(shape)

# Load the data
train_images = read_idx('train-images.idx3-ubyte')
train_labels = read_idx('train-labels.idx1-ubyte')
test_images = read_idx('t10k-images.idx3-ubyte')
test_labels = read_idx('t10k-labels.idx1-ubyte')
```

برخی از نکات این پیاده‌سازی به شرح زیر است:

- استفاده از مدل‌های از قبل آموزش داده شده و روش‌های مبتنی بریادگیری انتقالی مجاز نمی باشد.
- شما قادر به استفاده از کانوولوشن در معماری خود نیستید.
- استفاده از تابع فعال سازی softmax الزامی است.
- برای Loss از Crossentropy و برای optimizer از adam استفاده کنید.
- برای این تمرین ماتریس درهم ریختگی (Confusion Matrix) مربوطه را رسم کنید و در سند نهایی قرار دهید.

3. آپلود

در انتها نتایج مدل خود را در فایلی به فرمت. csv ذخیره کنید. این فایل باید به تعداد سطرهای داده ی تست پیشبینی داشته باشد که در یک ستون با نام Class ذخیره شده است. فایل نتایج را به علاوه ی کد در پوشه ی مربوط به سوال در ریپازیتوری گیت هاب قرار دهید.

آنچه باید تحویل دهید:

لازم است که موارد زیر را به عنوان موارد مورد تحویل هر سوال در پوشه ی مربوط به آن سوال داخل ریپازیتوری گیت های قرار دهید:

- فایل جواب سوالات تشریحی را نیز در پوشه با شماره سوال در گیت هاب آپلود کنید.
- فایل ژوپیتر نوت بوک به فرمت. ipynb که شامل کد پاسخ و توضیحات مربوط به آن است.
- فایل پیش بینی های خروجی مدل که به فرمت Q#.csv خواهد بود و # شماره ی سوال می باشد.
- تحویل همه موارد فقط و فقط از گیت هاب صورت میگیرد. لینک تمرین:

https://classroom.github.com/a/vv-tTzXs

- مهلت تمرین تا جمعه 1 دی، ساعت 3 بامداد می باشد.

موفق باشید