هوش مصنوعی — تمرین کامپیوتری شماره‌ی ۳

مدرس: **دکتر فدایی**

طراح: محمد سعادتی، آرش محمدپور، مهسا همت‌پناه

مهلت تحویل: شنبه ۹ اردیبهشت ۱۴۰۲، ساعت ۲۳:۵۵

مقدمه

در این پروژه قصد داریم با استفاده از Naive Bayes Classifier به طبقه‌بندی تصاویر بپردازیم.

معرفی مجموعه داده

در این تمرین شما با مجموعه داده اعداد دست‌نویس فارسی کار خواهید کرد. مجموعه داده شامل تصاویری از ۱۰ رقم ۰ تا ۹ در زبان فارسی است. در مجموع ۱۰۲۳۵۲ تصویر در این مجموعه داده موجود است که ۶۰۰۰ تصویر از بین این تصاویر برای تمرین به شما داده شده است. اطلاعات بیشتر راجع به مجموعه داده و تعداد تصاویر هر رقم در مجموعه داده را می‌توانید اینجا بخوانید.

محتویات پوشه مجموعه داده به صورت زیر است.

```
dataset
├── data.pkl
└── label.pkl
```

مجموعه داده‌ها و برچسب‌های آن‌ها در قالب فایل‌های pkl در اختیار شما قرار داده شده‌اند تا حجم کمتری داشته باشند.

شما می‌توانید با استفاده از قطعه کد نمونه زیر فایل‌ها را بخوانید.

```
import pickle
pkl_file = open('file.pkl', 'rb')
data = pickle.load(pkl_file)
```

مجموعه داده به صورت لیستی از آرایه‌های پایتون و مجموعه برچسب‌ها به صورت لیستی از اعداد ۰ تا ۹ در اختیار شما قرار گرفته است.

همچنین مجموعه داده‌ها از ۶۰۰۰ تا عکس تشکیل شده است که باید به نسبت دلخواه و منطقی، از آن برای آموزش و تست مدل خود استفاده کنید.

فاز اول: بررسی و پیش‌پردازش داده

۱. مجموعه داده‌ها را با چه نسبتی به دو مجموعه train و test تقسیم کردید؟ دلیل عدد انتخابی خود برای نسبت تقسیم را توضیح دهید.

۲. یک تصویر در مجموعه داده train را به صورت رندوم بررسی کنید و نشان دهید (از کتابخانه matplotlib استفاده کنید)، مقادیر هر پیکسل در چه محدوده‌ای قرار می‌گیرد؟

۳. از آنجایی که اندازه تصاویر متفاوت است، اندازه تصاویر را به اندازه ۲۰ در ۲۰ تغییر دهید. علت یکسان‌سازی اندازه تصاویر را نیز ذکر کنید.

۴. در مجموعه داده train، از هر کلاس به دلخواه یک تصویر انتخاب کرده و نمایش دهید. برای هر تصویر، نوع آن را نیز به همراه تصویر نمایش دهید.

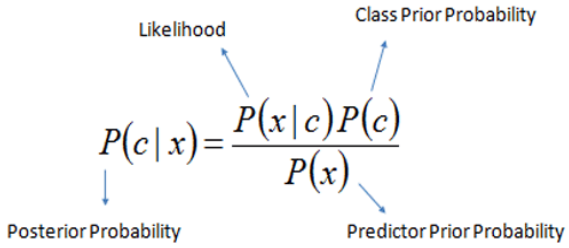
۵. تعداد تصاویر هر دسته را برای مجموعه داده train و test محاسبه کنید و برای آن‌ها نمودار میله‌ای رسم کنید.

۶. مقدار داده‌ها را به گونه scale کنید که قبل از دادن اطلاعات به Naive Bayes Classifier، مقدار هر pixel بین 0 تا 1 باشد. در صورت انجام ندادن این کار چه مشکلی ممکن است رخ دهد؟

فاز دوم: فرآیند مسئله

در این مسئله می‌خواهیم دو نوع الگوریتم Naive Bayes را برای دسته‌بندی تصاویر اعداد دست‌نویس پیاده سازی کنیم. این دو مدل الگوریتم‌های Gaussian Naive Bayes و Bernoulli Naive Bayes هستند.

برای حل این مسئله به صورت کلی از naive bayes استفاده می‌کنیم که مفهوم پشت آن با توجه به مفاهیم احتمالی زیر قابل بحث است.



$$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

۷. در گزارش کار خود، توضیح دهید که هر کدام از (evidence, likelihood, prior, posterior) بیانگر چه مفهومی در این مسئله هستند و چگونه محاسبه می‌شوند.

دقت کنید که نیازی نیست عبارت Evidence در مخرج کسر به صورت مستقیم محاسبه شود.

Gaussian Naive Bayes

در این الگوریتم فرض می‌شود مقادیر pixel های تصویر هر عدد دارای یک توزیع گوسی یا نرمال هستند. بدین ترتیب اگر c کلاس داشته باشیم می‌توانیم برای هر کلاس میانگین و واریانس را محاسبه کرده و پارامترهای توزیع نرمال را برای آن‌ها برآورد کنیم. اگر μ_c را میانگین و σ_c^2 را واریانس کلاس c ام در نظر بگیریم، برای محاسبه مقدار likelihood خواهیم داشت:

$$P(X|Y=c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{\frac{-(x-\mu_c)^2}{2\sigma_c^2}}$$

Bernoulli Naive Bayes

در این الگوریتم فرض می‌شود که مقادیر پیکسل هر رقم binary است (سیاه یا سفید). ابتدا باید مقادیر پیکسل عکس‌ها را با استفاده از یک threshold به binary تبدیل کنید و سپس با استفاده از مجموعه داده train، احتمال سیاه یا سفید بودن هر پیکسل را برای هر کلاس تخمین بزنید. اگر x_i را مقدار پیکسل i ام و $P(x_i|c)$ را احتمال مشاهده پیکسل i ام با مقدار 1 (سیاه) در نظر بگیریم، برای محاسبه مقدار likelihood خواهیم داشت:

$$P(X|Y=c) = \prod_{i=1}^{n} P(x_i|c)^{x_i}(1-P(x_i|c))^{(1-x_i)}$$

Additive Smoothing

در Naive Bayes مشکلی که ممکن است در بدست آوردن دسته‌ها به آن برخورد کنید این است که با ویژگی‌هایی (مقدار یک پیکسل در یک تصویر) روبرو شوید که در داده train برای دسته‌ی خاصی وجود نداشته است.

برای مثال ممکن است در کلاس ۸ مقدار پیکسل ۱۸۰ ام در داده های train ،صفر بوده باشد، در صورتی که در داده test تصاویری برای کلاس ۸ موجود باشند که مقدار پیکسل ۱۸۰ ام در آن ها مقدار یک را داشته باشد. بنابراین در این حالت، احتمال صفر بدست می آید.

همچنین در حالت خاص Gaussian Naive Bayes در برخی موارد، مقدار واریانس ویژگی برای یک کلاس خاص می‌تواند صفر باشد و در این حالت، استفاده از واریانس در محاسبه احتمال می‌تواند منجر به تقسیم بر صفر شود. برای رفع این مشکل یک مقدار ثابت کوچک به عنوان Smoothing به واریانس برای هر ویژگی و هر کلاس اضافه می‌شود.

۸. در گزارش خود با در نظر داشتن Naive Bayes توضیح دهید چرا این اتفاق رخ می‌دهد.

۹. درباره روش Smoothing تحقیق کنید و با پیاده‌سازی آن در پروژه و در هر دو نوع الگوریتم Naive Bayes، این مشکل را برطرف کنید.

در گزارش خود این روش را توضیح دهید و بگویید دقیقا چطور به حل این مشکل کمک می‌کند. (در بخش ارزیابی، تفاوتی که استفاده از این روش بر دقت می‌گذارد را باید گزارش کنید.)

فاز سوم: ارزیابی

برای ارزیابی هر دو مدل خود باید از 4 معیار زیر استفاده کنید.

$$Accuracy = \frac{Correct\ Detected}{Total}$$

$$Precision = \frac{Correct\ Detected\ Class}{All\ Detected\ Class\ (Including\ Wrong\ Ones)}$$

$$Recall = \frac{Correct\ Detected\ Class}{Total\ Class}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Correct Detected Class: تعداد عکس‌هایی که به درستی در دسته‌بندی مورد نظر تشخیص داده شده‌اند.

All Detected Class: تعداد تمام عکس‌هایی که در دسته‌بندی مورد نظر تشخیص داده شده‌اند. (حتی اگر به اشتباه)

Total Class: تعداد تمام عکس‌هایی که در مجموعه داده تست در آن دسته‌بندی خاص بودند.

به جای Class می‌توانید هرکدام از دسته‌بندی‌های موجود را بگذارید.

۱۰. در گزارش کار خود توضیح دهید که چرا مقدار Precision و Recall هر کدام به تنهایی برای ارزیابی مدل کافی نیست؟ برای هر کدام مدلی را مثال بزنید که در آن، این معیار مقدار بالایی دارد ولی مدل خوب کار نمی‌کند.

۱۱. در گزارش کار خود توضیح دهید معیار F1 از چه نوع میانگین‌گیری بین Precision و Recall استفاده می‌کند؟ تفاوت آن نسبت به میانگین‌گیری عادی چیست و در اینجا چرا اهمیت دارد؟

۱۲. با توجه به اینکه مسئله ما بیشتر از ۲ کلاس دارد در مورد multi-class metrics تحقیق کنید. در گزارش کار خود، سه حالت میانگین‌گیری macro و micro و weighted را شرح دهید. برای تحقیق می‌توانید از این سایت[1] استفاده کنید.

[1] https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1

مدل های خود را که با استفاده از naive bayes و براساس داده‌ی train ساخته‌اید، روی داده‌ی test اجرا کنید و برای هر کدام از سطرهای آن، تشخیص مدل‌هایتان را بدست آورید. سپس براساس آن، معیارهای بالا را برای هر کلاس به صورت تنها و سپس با استفاده از سه نوع میانگین‌گیری گفته شده برای تمام کلاس‌ها محاسبه کنید. (برای محاسبه معیارها **نباید** از کتابخانه‌ها استفاده شود اما برای مطمئن شدن از محاسباتتان می‌توانید از توابعی مثل classification_report[2] استفاده کنید.)

مقدار accuracy در حالت 'ب' باید بیشتر از ۸۵ درصد باشد.

۱۳. در گزارش خود، معیارها را به ازای هر دو حالت زیر به دست آورید (نمونه‌ای از معیارهایی که باید گزارش کنید در ادامه آمده است. توجه کنید که این فقط یک مثال از نحوه ارائه نتایج است.)

الف. نتایج بدون استفاده از Additive Smoothing

ب. نتایج با استفاده از Additive Smoothing

	0	1	2	3	4	5	6	7	8	9	All Classes
Precision											-
Recall											-
F1-score											-
Accuracy	-	-	-	-	-	-	-	-	-	-	
Macro Avg	-	-	-	-	-	-	-	-	-	-	
Micro Avg	-	-	-	-	-	-	-	-	-	-	
Weighted Avg	-	-	-	-	-	-	-	-	-	-	

۱۴. در گزارش خود، مقادیر بدست آمده در بخش قبل را تحلیل کنید.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

۱۵. در گزارش خود ۵ مورد از عکس هایی که در داده‌ی تست هستند و مدل شما دسته اشتباهی برای آن‌ها تشخیص داده است بیاورید.

نکات پایانی

- دقت کنید که هدف پروژه تحلیل نتایج است؛ بنابراین از ابزارهای تحلیل داده مانند نمودارها استفاده کنید و توضیحات مربوط به هر بخش از پروژه را به طور خلاصه و در عین حال مفید، در گزارش خود ذکر کنید.
- نتایج و گزارش خود را در یک فایل فشرده با عنوان AI_CA3_<#SID>.zip تحویل دهید. محتویات پوشه باید شامل فایل jupyter-notebook، خروجی html و فایل‌های مورد نیاز برای اجرای آن باشد. توضیح و نمایش خروجی‌های خواسته شده بخشی از نمره این تمرین را تشکیل می‌دهد. از نمایش درست خروجی‌های مورد نیاز در فایل html مطمئن شوید.
- در صورتی که سوالی در مورد پروژه داشتید بهتر است در فروم درس یا در گروه تلگرام مطرح کنید تا بقیه از آن استفاده کنند؛ در غیر این صورت از طریق ایمیل با طراحان در ارتباط باشید.
- هدف از تمرین، یادگیری شماست. لطفا تمرین را خودتان انجام دهید.

موفق باشید!