



به نام خدا

دانشکده مهندسی برق و کامپیوتر دانشگاه تهران

98-99 هوش مصنوعی، ترم پاییز

پروژه طبقه بند بیزی، مهلت ارسال: ۲۵ آبان

طراحان پروژه: پارسا قربانی، احسان آقازاده و کامران حسینی



## Saadi vs. Hafez the Shirazians

### مقدمه

در این پروژه شما باید با استفاده از naïve bayes، طبقه بندی<sup>1</sup> را طراحی کنید که بتواند با گرفتن یک بیت شعر، مشخص کند این شعر از سعدیست و یا حافظ.

### توضیح مسئله

فایل train\_test.csv که در اختیار شما قرار داده شده است، یک فایل ۵۰۰۰ نمونه است که هر نمونه دارای دو مقدار دارد، اولی یک بیت شعر و دومی شاعر این بیت است.

در این قسمت دو راه حل پیشنهادی برای مسئله بیان شده است. این مسئله راه حل های متفاوتی دارد و در صورتی که در کلاس درس راه حل متفاوتی آموخته اید، می توانید بر اساس آن عمل کنید.

در راه اول می توانید هر کلمه را به عنوان یک فیچر در نظر بگیرید که ممکن است در هر بیت وجود داشته باشد یا نداشته باشد. (طبعاً باید از کلماتی استفاده کنید که در دیتاست شعر وجود دارند). بنابراین باید دیتاستی که به شما ورودی داده شده است را بر این اساس به یک دیتافریم تبدیل کرده که مقدار هدف (target value) آن شاعر و ویژگی هایش (Feature) کلمات (یا پوزیشن آن ها) هستند. این دیتافریم احتمالاً<sup>۱</sup> باینری است.

در راه حل دیگر، انتخاب فیچر جایگاه به عنوان ویژگی اصلی است که مقدار آن می تواند کلمات مختلف باشد.

در گزارش خود ذکر کنید که در راه حل شما چه ویژگی هایی استفاده می شوند و طبق آن، ۴ مقدار مختلف قاعده بیزین (posterior, prior, likelihood, evidence) بیانگر چه مفهومی بوده و چگونه محاسبه می شوند.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Likelihood      Class Prior Probability

↓      ↓

Posterior Probability      Predictor Prior Probability

\*راهنمایی: نیازی به محاسبه عبارت evidence در مخرج کسر نیست و می توانید با مقایسه احتمال شاعر بودن سعدی و احتمال شاعر بودن حافظ، در مورد شاعر تصمیم گیری کنید.

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

<sup>1</sup> می توان به جای در نظر گرفتن وجود داشتن یا نداشتن یک کلمه در بیت، تعداد آن را محاسبه کرد. در این صورت ویژگی باینری نخواهد بود.

سپس با پیاده سازی کدی که عبارات بالا را محاسبه کند، مقادیر بیز posterior یعنی  $p(y|x)$  برای هر شاعر و فیچر محاسبه می شود. یعنی می دانید به ازای هر کلمه، احتمال حافظ بودن شاعر و سعدی بودن آن چه قدر است. نهایتاً مدل شما می تواند برای اشعار جدید این شناسایی را انجام دهد.

## معیار ارزیابی مدل

برای ارزیابی مدل باید ابتدا بخشی از داده ها، یعنی ۸۰ درصد آن ها را به عنوان ورودی به برنامه خود داده که بر اساس آن مقادیر لازم روابط بیزین محاسبه می شود. سپس شاعر ۲۰ درصد مابقی داده ها را بر اساس مدل به دست آمده، تشخیص دهید. نهایتاً با مقایسه کردن خروجی آن ۲۰ درصد با مقادیر واقعی، دقت مدل محاسبه می شود. برای تعریف دقت سه عدد recall, precision و accuracy را محاسبه کرده و در گزارش کار خود ذکر کنید.

$$Recall = \frac{Correct\ Detected\ Hafezes}{All\ Hafezes}$$

$$Precision = \frac{Correct\ Detected\ Hafezes}{Detected\ Hafezes\ (This\ also\ includes\ wrong\ detections)}$$

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

Correct Detected Hafezes: تعداد بیت هایی از حافظ که مدل شما درست تشخیص داده است.

Detected Hafezes: تعداد بیت هایی که مدل شما شاعر آنها را حافظ تشخیص داده است.

Correct Detected: تعداد بیت هایی که مدل شما شاعر آنها را به درستی مشخص کرده است.

Total: تعداد کل بیت های داده تست

- accuracy بالای ۷۶ درصد مطلوب است. (توجه کنید که مدل رندوم دقت حدود ۵۰ درصد دارد و ممکن است به طور اتفاقی ۶۰ درصد بشود. بنابراین اگر به دقت ۵۱ درصد رسیدید، ارزشی ندارد).
- مقدار precision و recall مطلوب حدود ۷۳ درصد است.

## سوال های اضافی

۱. چنانچه برای ارزیابی خروجی یک مدل ماشین لرنینگ فقط به مقدار Precision توجه شود چه مشکلی پیش می آید؟ برای مثال یک مدل ماشین لرنینگ معرفی کنید که Precision بالایی دارد ولی خوب کار نمی کند.
۲. چرا مقدار Accuracy به تنهایی برای تشخیص خوبی مدل کافی نیست؟ برای مثال یک دیتاست معرفی کنید که یک مدل ضعیف ماشین لرنینگ روی آن به Accuracy بالایی برسد بدان آن که عملاً تشخیصی انجام دهد.

ارزیابی، نهایی، مدل

برای ارزیابی مدل، در مرحله قبل روی بخشی از داده ها عملیات یادگیری را انجام دادید و با بخش دیگر تست کردید. حالا فایل `evaluate.csv` به شما داده شده است که شامل دو ستون است یکی `id` و دیگری یک بیت شعر پس از نهایی شدن مدل و یادگیری از داده `train` مدل خود را با داده های فایل `evaluate.csv` اجرا کرده و شاعر هایی که مدل شما به هر کدام از بیت ها تخصیص میدهد را به فرمت یک فایل `csv` با نام `output.csv` با دو ستون یکی `id` که همان `id` مربوط به بیت در فایل `evaluate.csv` است و دیگری نام شاعری که مدل شما خروجی داده است را همراه با گزارش کار و کد های خود آپلود کنید. دقت این خروجی در نمره پروژه تاثیر مستقیم دارد. شما دقت خود را روی این فایل نمی دانید، ولی اگر از قسمت قبل به دقت خوبی رسیده باشید، طبیعتاً نباید مشکل ایجاد کند.

تحلیل عمیق تر (اختیاری: این قسمت (بین دو خط آبی) حذف شده و نمره ای ندارد).

توجه کنید که هر مدل ماشین لرنینگ یک برداشت (interpretation) دارد که بیان می کند این موجود هوش مصنوعی چگونه تصمیم گیری می کند. در این مسیله عامل هوشمند با نگاه به کلمات حدس می زند که شاعر این شعر کیست. حالا شما باید دقیق تر مغز این موضوع را بررسی کنید. ابتدا یک Tag Cloud آماده کنید که نشان می دهد کلماتی که بیشترین تاثیر را در انتخاب شاعر به عنوان حافظ دارند، کدام ها هستند. مثل شکل مقابل، در شکل شما کلماتی که باعث شاعر شدن حافظ می شوند، بزرگ تر نمایش داده می شوند.

\* کتابخانه های آماده ای برای پایتون این کار را انجام می دهند.

سپس دو کلمه ای که بسیار تاثیر گذار هستند را انتخاب کنید و برای هر کدام یک نمودار میله ای بکشید که در صورت رخداد آن کلمه احتمال شاعر بودن حافظ و احتمال شاعر بودن سعدی چه قدر است. (واضح است جمع این دو احتمال باید یک باشد).

لا يلاس

ممکن است کلماتی وجود داشته باشند که فقط در اشعار یک شاعر (و حتی فقط یک بار) به کار رفته باشند. برای مثال کلمه گیاه فقط در یک شعر حافظ به کار رفته باشد اما در شعر سعدی نباشد. انسان تشخیص می دهد آمدن گیاه به معنای شاعر بودن حافظ نیست. در گزارش کار بیان کنید که چرا وجود این کلمه در یک بیت باعث می شود که مدل نایو بیز آن را با احتمال قطعی به حافظ نسبت دهد. (حتی اگر عوامل بسیاری نشان دهنده متعلق بودن شعر به سعدی باشند و نبود این کلمه در اشعار قبلی حافظ صرفا اتفاقی و به خاطر نادر بودن خود کلمه باشد). سپس با جست و جو یا ایده راه حلی برای آن ارایه کنید. این راه حل را ساده سازی کرده و تفاوتی را در دقت ها گزارش کنید.

## گزارش کار (جمع بندی)

در گزارش خود چارچوب کلی الگوریتم پیاده سازی شده را شرح دهید. به طور مجزا جواب سوالات را با هدر مناسب مشخص کنید. مطمئن شوید مواردی که در متن بیان شده را ثبت کرده اید. این موارد شامل دقت ها، تحلیل ها، و تاثیرات قسمت آخر هستند.

## آپلود

گزارش کار، تمام کدها که شامل کدهای پیاده سازی نایوبیز و تحلیل ها و نمودار ها است به همراه فایل `evaluate.csv` را حتما آپلود کنید. در غیر این صورت به بخش های ناقص نمره ای تعلق نمی گیرد.

## نکات

- کل پروژه های درس به زبان پایتون باید انجام شود.
- تاخیر به ازای روز اول و دوم هر کدام ۱۰ درصد و روز سوم به بعد هر روز ۱۵ درصد خواهد بود. برای مثال سه روز تاخیر ۳۵ درصد از نمره دریافتی شما را کم میکند.
- برای ما مهم است که حاصل کار خودتان را به ما تحویل دهید. در صورت تقلب برای بار اول به هر دو طرف نمره -۱۰۰ تعلق میگیرد و بار دوم معرفی به دانشگاه و ثبت نمره ۰/۲۵ به عنوان تقلب انجام می شود.

هر گونه ابهام و سوالی را در فروم و یا با ارسال ایمیل به نشانی های موجود در اطلاعیه تمرین با ما در میان بگذارید.

موفق باشید