آمار و احتمال مهندسی

پروژه درس آمار و احتمال مهندسی

دکتر مرادیان

نیمسال دوم ۱۴۰۲-۱۴۰۳

یادگیری با استفاده از الگوریتم Naive Bayes

آمار و احتمال مهندسی

دکتر مرادیان

توضیح پروژه

در این پروژه هدف پیاده‌سازی یک فرآیند یادگیری ساده با استفاده از الگوریتم Naïve Bayes است. همانطور که در ادامه خواهیم دید این الگوریتم از اصول احتمالات شرطی و قائده بیز برای دسته‌بندی داده‌ها استفاده می‌کند.

فرآیند یادگیری ذکر شده در دو مرحله پیاده‌سازی می‌شود. در مرحله اول شما یک مجموعه داده (Dataset) به نام X دریافت می‌کنید که مجموعه‌ای از نمونه‌ها است. هر نمونه توسط یک بردار n-بعدی $x = (x_1, \dots, x_n) \in X$ نشان داده می‌شود و هر مؤلفه بردار یک ویژگی نامیده می‌شود. بنابراین هر بردار شامل n ویژگی است. همچنین هر یک از xها به یکی از m دسته $c_1, c_2, \dots, c_m$ تعلق دارند. برای مثال هر عضو از مجموعه داده dogs.csv در این پروژه نشان دهنده ویژگی های یک سگ (مثلا ارتفاع و عرض) است و تعداد دسته‌بندی‌ها سگ‌ها 3 تا است. با توجه به اطلاعات موجود در این مرحله یعنی مجموعه نمونه‌ها و کلاس هر نمونه، شما می‌توانید توزیع ویژگی‌های هر دسته را به دست آورید. سپس در مرحله بعد با استفاده از این توزیع‌ها و همچنین الگوریتم Naïve Bayes می‌توانید داده‌های جدید را که قبلا در مجموعه داده شما حضور نداشتند را دسته‌بندی کنید. در ادامه ابتدا بخش یادگیری و سپس الگوریتم Naïve Bayes با جزئیات توضیح داده می‌شود.

مرحله اول: بخش یادگیری (Training)

در این مرحله شما مجموعه داده X را دریافت می‌کنید که شامل نمونه‌های x و دسته بندی آن‌ها است. سپس توزیع ویژگی‌ها را در هر دسته به دست می‌آورید. به بیان دقیق‌تر هدف، به دست آوردن توزیع احتمال زیر برای هر ویژگی $1 \leq i \leq n$ و هر دسته‌بندی $1 \leq k \leq m$ است:

$$P(x_i|c_k)$$

آمار و احتمال مهندسی

دکتر مرادیان

در صورتی که x_i یک متغیر تصادفی گسسته باشد، می‌توانید تعداد نمونه‌هایی در دسته c_k را که دارای ویژگی $x_i = \alpha$ هستند را به تعداد کل نمونه‌ها در این دسته تقسیم کنید و مقدار $P(x_i = \alpha|c_k)$ را به دست آورید. در این پروژه نوع توزیع به شما داده می‌شوند و شما باید صرفا پارامترهای توزیع را به دست آورید. توزیع‌های زیر در این پروژه به کار می‌آیند:

الف) تابع uniform PDF با پارامترهای a و b :

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a,b]. \\ 0, & \text{otherwise.} \end{cases}$$

ب) تابع Gaussian PDF با پارامترهای μ و σ :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

ج) تابع PMF را برای متغییر تصادفی دو جمله ای با پارامتر n و p :

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 1, 2, \ldots, n$$

مرحله دوم: بخش دسته‌بندی و الگوریتم Naive Bayes

در این بخش الگوریتم Naïve Bayes به منظور دسته‌بندی داده‌های جدید توضیح داده می‌شود. فرض کنید x برداری باشد به صورت $x = (x_1, x_2, \ldots, x_n)$ که در آن n تعداد ویژگی های هر نمونه باشد.

همچنین همانطور که پیش از این ذکر شد فرض کنید هر داده x عضو یکی از m دسته بندی $c_1, c_2, \ldots, c_m$ است. میخواهیم مشخص کنیم اگر x خاصی به ما داده شود که در مجموعه داده ما نیست متعلق به کدام دسته است. برای اینکار دسته ای را انتخاب می‌کنیم که احتمال $P(c_k \mid x)$ را بیشینه کند. در واقع دسته‌ای را انتخاب می‌کنیم که احتمال متعلق بودن به آن دسته بیشینه شود. بنابراین بهینه سازی زیر را حل می‌کنیم:

$$\hat{k} = argmax_k \, P(c_\mathrm{k} \mid x)$$

که در رابطه بالا $\hat{k}$ دسته انتخابی است. برای حل بهینه‌سازی بالا احتمال $P(c_\mathrm{k} \mid x)$ را طبق قانون بیز به صورت زیر محاسبه می‌کنیم:

$$P(c_\mathrm{k} \mid x) = \frac{P(x \mid c_\mathrm{k})P(c_\mathrm{k})}{P(x)}$$

میدانیم مخرج کسر برای همه دسته بندی ها یکسان می باشد. پس برای محاسبه دسته مطلوب کافی است صورت کسر را بیشینه کنیم. رایج است تا فرض کنیم هر کدام از کلاس ها احتمال یکسانی دارند $P(c_1) = P(c_2) = \ldots = P(c_\mathrm{m})$. پس نیازمند این هستیم که تا مقدار $P(x \mid c_\mathrm{k})$ را بیشینه کنیم. همچنین فرض میشود به شرط دانستن دسته، ویژگی ها از هم مستقل هستند (به همین دلیل به این الگوریتم Naive یا ساده گفته می‌شود):

$$P(x \mid c_\mathrm{k}) = P(x_1 \mid c_\mathrm{k})P(x_2 \mid c_\mathrm{k}) \ldots P(x_n \mid c_\mathrm{k})$$

اگر x_i گسسته باشد برای به دست آوردن $P(x_i = \nu \mid c_\mathrm{k})$ کافی است تعداد نمونه هایی در c_k را که ویژگی x_i آن‌ها برابر با ν است $(n_{i,k})$ را به تعداد کل نمونه‌ها در دسته c_k یعنی n_k تقسیم کنیم:

$$P(x_i = \nu | c_k) = \frac{n_{i,k}}{n_k}$$

اگر x_i پیوسته باشد فرض میکنیم که مقدار $P(x_i \mid c_\mathrm{k})$ از توزیع مربوط به آن ویژگی به دست می‌آید. برای مثال اگر ویژگی x_i در دسته c_k از توزیع گوسی با پارامترهای $\mu_{i,k}$ و $\sigma_{i,k}^2$ پیروی کند مقدار $P(x_i \mid c_\mathrm{k})$ به صورت زیر تعریف می‌شود:

آمار و احتمال مهندسی

دکتر مرادیان

$$P(x_i|c_k) = \frac{1}{\sqrt{2\pi\sigma_{i,k}^2}} e^{-\frac{(x_i-\mu_{i,k})^2}{2\sigma_{i,k}^2}}$$

در این پروژه توزیع ویژگی‌ها به شما داده می‌شود و شما باید تنها پارامترهای توزیع را به دست آورید. در ادامه شما الگوریتم یادگیری معرفی شده را برای دو حالت ویژگی‌های پیوسته و گسسته پیاده‌سازی خواهید کرد.

بخش پیوسته

برای این بخش از فایل dogs.csv استفاده کنید. این فایل در واقع مجموعه داده X است که شامل بردارهای نمونه x و کلاس‌های آن‌ها می‌باشد. بخشی از ویژگی ها پیوسته و یکی گسسته است. به طور خاص، height و weight از توزیع گاوسی پیروی میکنند. bark_days از توزیع دو جمله ای و ear_head_ratio از توزیع یکنواخت پیروی میکند.

بخش اول: یادگیری) در این بخش با توجه به توزیع هر ویژگی که داده شده است، پارامترهای هر توزیع را به دست آورید. برای این منظور با توجه به دسته بندی هر ویژگی پارامتر مخصوص آن را بدست می آوریم. برای سادگی کار میتوان از کتابخانه numpy در python کمک گرفت.

بخش دوم: دسته‌بندی) اکنون تابعی بنویسید که ویژگی های مدنظر را بگیرد (میتواند به صورت آرایه یا لیست باشد) و با توجه به توابع توزیعی که در بخش اول به دست آورده‌ایم و همچنین به کارگیری الگوریتم Naïve Bayes دسته مربوط به نمونه دریافتی را به عنوان خروجی ارائه دهد.

آمار و احتمال مهندسی

دکتر مرادیان

بخش گسسته

برای این بخش از فایل emails.csv به عنوان مجموعه داده استفاده کنید. در اینجا هر متن ایمیل یک نمونه است که در آن هر کلمه یک ویژگی است. پس بدین منظور تعداد ویژگی های هر نمونه میتواند یکسان نباشد. هر نمونه میتواند اسپم باشد یا نباشد.

بخش اول: یادگیری) در این بخش تابعی بنویسید که لیست همه کلمات را پیدا کند و سپس احتمال آن را بیابید که هر کلمه خاص در ایمیل‌های spam و غیر spam به کار رفته باشد. در واقع شما دو دسته‌بندی $c_0 = ham$ و $c_1 = spam$ در این مسئله دارید و قرار است احتمالات $P(x_i|c_0)$ و $P(x_i|c_1)$ را به دست آورید که x_i در آن یک کلمه است. برای اینکار تعداد دفعاتی را که کلمه x_i در ایمیل های spam و غیر spam به کار رفته است را بشمارید و به تعداد کل کلمات در هر گروه تقسیم کنید.

بخش دوم: دسته‌بندی) در این بخش یک رشته تحت عنوان یک جمله به شما داده می‌شود و با توجه به کلمات استفاده شده در این جمله مشخص کنید اسپم است یا خیر. (برای اینکار باید احتمال اسپم بودن جمله (یا ایمیل) را به دست آورید.)