دانشکده مهندسی کامپیوتر و فن‌آوری اطلاعات
دانشگاه صنعتی امیرکبیر

گزارش تمرین اول درس مدل‌های احتمالاتی گرافی

استاد درس:

دکتر نیک‌آبادی

نام دانشجو:

احمد اسدی

۹۴۱۳۱۰۹۱

فروردین ۱۳۹۵

فهرست مطالب

۱ پیش‌پردازش و استخراج ویژگی‌ها

در این بخش به بررسی عملیات پیش‌پردازش و استخراج ویژگی‌ها می‌پردازیم. تمامی مراحل مربوط به پیش‌پردازش اطلاعات و ویژگی‌های استخراج شده، شامل نحوه استخراج ویژگی‌ها، نحوه محاسبه توزیع‌های احتمالی شرطی[۱] و نحوه نگاشت داده‌ها به فضای برداری را در این قسمت بررسی خواهیم نمود.

۱.۱ پیش‌پردازش

قبل از بررسی عملیات مربوط به پیش‌پردازش داده‌ها، ابتدا باید ساختار داده‌ها را شناخت. در مجموعه‌داده مورد استفاده[۲] هر فایل خبری شامل دو بخش اصلی است:

سرآیند در این بخش، اطلاعات کلی شامل اطلاعاتی مانند شناسه فایل خبری، شناسه اخبار مرجع، رایانامه نویسنده خبر، نام نویسنده، موضوع خبر، نام سازمان مرتبط و مانند آن وجود دارد. از آنجا که بخش قابل توجهی از این اطلاعات مربوط به دسته‌خبری می‌باشند، نمی‌توان از آن‌ها در فرآیند دسته‌بندی استفاده نمود. با این حال، مواردی مانند نام نویسنده، رایانامه نویسنده، موضوع خبر، تعداد خطوط موجود در خبر و اطلاعاتی از این قبیل می‌توانند نماینده‌های خوبی به عنوان ویژگی باشند.

بدنه اصلی در این بخش، محتوا و متن خبر وجود دارد. این محتوا در تمام فایل‌ها با یک خط خالی از بخش سرآیند تفکیک شده است. این بخش، در تمام فرآیند دسته‌بندی مورد استفاده قرار می‌گیرد و همان‌طور که در ادامه توضیح داده خواهد شد، تمام ویژگی‌های مربوطه از این بخش استخراج می‌شوند.

با توجه به ساختار داده‌ها و از آنجا که دقت نهایی الگوریتم در این تمرین حائز اهمیت نمی‌باشد، از بخش سرآیند به طور کلی صرف نظر نموده و تمام عملیات‌های خود را روی بخش بدنه اصلی اخبار انجام می‌دهیم. در این بخش نیاز داریم تا کلمات معنادار اخبار را جدا نموده و از بین آن‌ها تعدادی را به عنوان ویژگی انتخاب کرده و مورد استفاده قرار دهیم. به همین منظور، عملیات پیش‌پردازش مناسب باید بتواند کلمات را از ارقام و علامت‌های نگارشی جدا نماید تا بتوان از بین آن‌ها کلمات کلیدی را با دقت بیشتری انتخاب نمود.

از این رو، در مرحله پیش‌پردازش، ابتدا کلمات هر فایل را با توجه به کاراکتر فاصله جدا می‌نماییم. در این پروژه، کلماتی را که شامل ارقام هستند به کلی حذف کرده و از آن‌ها استفاده نمی‌نماییم. جدول ۱ تعدادی از کلماتی را که در مجموعه‌داده وجود دارند و شامل اعداد هستند نمایش می‌دهد.

جدول ۱: نمونه‌هایی از کلماتی که در مجموعه‌داده وجود دارند و شامل ارقام هستند.

کلمه	شماره فایل	دسته‌خبری	مجموعه‌داده
ISBN0-910309-26-4	۴۹۹۶۰	alt.atheism	20NewsGroups
D-3000Hannover	۴۹۹۶۰	alt.atheism	20NewsGroups
416-629-7000/629-7044	۳۸۴۸۹	comp.graphics	mini_newsgroups
S1/S2	۱۵۴۲۳	sci.crypt	mini_newsgroups

در مرحله دوم پیش‌پردازش، تمام علائم نگارشی از محتوا حذف می‌شوند. علاوه بر این، ممکن است کلماتی در متن خبر وجود داشته باشند که شامل علائمی غیر از حروف انگلیسی باشند. در این موارد نیز تمام علامت‌های به غیر از حروف الفبا حذف شده و کلمه اصلی حفظ می‌شود. جدول ۲ نمایش‌دهنده نمونه‌هایی از کلامتی از این دست در کنار قالب پردازش‌شده آن‌ها می‌باشد.

مرحله دیگری که در بخش پیش‌پردازش انجام می‌شود، حذف کلمات توقف[۳] است. در این مورد،از آنجا که این کلمات در تمام دسته‌های خبری به طور یکنواخت تکرار می‌شوند و کمکی به دسته‌بندی صحیح نمی‌کنند،برای کاهش حجم محاسبات،آن‌ها را در این مرحله حذف می‌کنیم. در صورتی‌که این کلمات

[۱] Conditional Probability Distribution (CPD)
[۲] 20 NewsGroups
[۳] Stop Words

جدول ۲: نمونه‌هایی از پیش‌پردازش کلماتی که شامل علامت‌های نگارشی هستند.

کلمه اصلی	نتیجه پیش‌پردازش
joke>	joke
><disclaimer:	disclaimer

در این مرحله حذف نشوند،تاثیری در عملکرد بخش استخراج ویژگی نخواهند داشت؛ زیرا در مرحله استخراج ویژگی‌ها، کلماتی را که به طور یکنواخت در دسته‌های خبری تکرار شده‌اند به عنوان ویژگی،انتخاب نمی‌کنیم. این مرحله فقط برای کاهش حجم پردازش‌ها می‌باشد.

۲.۱ استخراج ویژگی‌ها

پس از مرحله پیش‌پردازش، لیستی از کلمات موجود در مجموعه‌داده بدست می‌آوریم. به منظور دست‌یابی به ویژگی‌های مناسب، یک شاخص[۱] از تعداد حضور هر ویژگی در هر دسته‌خبری تولید می‌نماییم. این شاخص در قالب یک ماتریس بیان می‌شود که ستون‌های آن، کلمات بدست‌آمده در مرحله قبل و سطرهای آن، دسته‌های خبری موجود را نمایندگی می‌کنند. در هر سلول از این داده‌ساختار، تعداد حضور کلمه مربوطه بین تمام داده‌های دسته خبری متناظر آن قرار می‌گیرد.

شاخص تولید شده علاوه بر این‌که برای انتخاب ویژگی استفاده می‌شود، خود به نوعی نمایش‌دهنده توزیع احتمال توام هر ویژگی و دسته‌های خبری می‌باشد. رابطه ۱ فرم کلی شاخص را نمایش می‌دهد. در این ماتریس k تعداد دسته‌های خبری موجود و n تعداد کل کلمات می‌باشد. W_{ij} نمایش‌دهنده تعداد تکرار کلمه j در میان اخبار موجود در دسته‌خبری i است.

$$W = \begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ W_{k1} & W_{k2} & \cdots & W_{kn} \end{bmatrix} \quad (1)$$

با محاسبه ماتریس شاخص به شکل روبرو، در هر سطر از ماتریس می‌توان کلماتی را که در یک دسته‌خبری بیشتر از بقیه کلمات تکرار شده‌اند را بدست آورد. از طرفی در هر ستون این ماتریس، می‌توان وضعیت تکرار کلمه مشخصی را در دسته‌های خبری مختلف مشاهده نمود. اگر کلمه‌ای در یکی از دسته‌های خبری تعداد تکرار به مراتب بیشتری نسبت به کلمه‌های دیگر داشته باشد، آن کلمه را به عنوان یکی از ویژگی‌ها انتخاب می‌کنیم. با این توضیح، کافیست ماتریس شاخص را به صورت ستونی پیمایش نماییم و اگر در یک ستون از ماتریس، جهش قابل مشاهده‌ای در مقدار یک سلول وجود داشت، کلمه متناظر آن ستون را به عنوان یکی از ویژگی‌ها برمی‌گزینیم.

علاوه بر این، اگر مقدار موجود در هر خانه ماتریس را به مجموع مقادیر هم ستون آن خانه تقسیم نماییم، احتمال رخداد توام کلمه مربوطه را در دسته‌خبری متناظر بدست می‌آوریم. رابطه ۲ این عملیات را نمایش می‌دهد. از آنجا که در این رابطه، تعداد تکرار کلمه i در دسته خبری j به مجموع تعداد تکرار این کلمه در تمام دسته‌های خبری تقسیم شده است، این رابطه احتمال وجود کلمه i را در دسته‌خبری j محاسبه می‌نماید. با توجه به رابطه ۳ مجموع احتمال حضور کلمه i در تمام دسته‌های خبری برابر یک بوده و بنابر این، رابطه احتمال ۲ یک توزیع احتمال معتبر را نمایش می‌دهد.

$$P(X_i, C_j) = \frac{W_{ij}}{\Sigma_{h=1}^{k} W_{ih}} \quad (2)$$

$$\Sigma_{l=1}^{k} P(X_i, C_l) = \Sigma_{l=1}^{k} \frac{W_{il}}{\Sigma_{h=1}^{k} W_{ih}} = \frac{\Sigma_{l=1}^{k} W_{il}}{\Sigma_{h=1}^{k} W_{ih}} = 1 \quad (3)$$

[۱] Index

در حالت برداری، می‌توان این عملیات را به شکل زیر انجام داد. در رابطه ۴، $sum(W)$ برداری است که هر سلول آن مجموع مقادیر ستون متناظر در ماتریس W را نمایش می‌دهد.

$$W = \frac{W}{sum(W)} \tag{۴}$$

همان‌طور که گفتیم، کلماتی که بیشتر در یکی از دسته‌های خبری تکرار شوند و در دسته‌های دیگر کمتر دیده شوند، انتخاب‌های مناسبی به عنوان ویژگی هستند. به همین دلیل پس از محاسبه احتمال حضور کلمات در دسته‌های خبری مختلف، مطابق با رابطه ۵ کلماتی که احتمال حضور آن‌ها بیشتر از α درصد از احتمال توزیع یکنواخت بین دسته‌های خبری باشد به عنوان ویژگی انتخاب می‌شود.

$$\Omega = \{X_i | (۱ + \alpha) \cdot min(W_{۱:n,۱:k}) < P(X_i, C_j)\} \tag{۵}$$

استفاده از رابطه ۵، به ما این امکان را می‌دهد که با تغییر دادن مقدار α، تعداد ویژگی‌های انتخاب شده را تغییر دهیم. اگر $\alpha = ۰$ انتخاب شود، تمام کلمات متن، به عنوان ویژگی انتخاب می‌شوند که بیشترین تعداد ممکن است. همین‌طور اگر $\alpha = \frac{max(W_{۱:n,۱:k})}{min(W_{۱:n,۱:k})} - ۱$ انتخاب شود، فقط کلماتی که دارای بیشترین انحراف معیار هستند به عنوان ویژگی انتخاب می‌شوند و در صورتی که α بزرگتر از این مقدار تعیین شود،هیچ کلمه‌ای به عنوان ویژگی انتخاب نخواهد شد. بنابراین، مقدار این پارامتر در آزمایش‌های مختلف تغییر می‌کند تا به بهترین مقدار ممکن آن دست پیدا کنیم.

۲ ماتریس درهم‌ریختگی و تاثیر تعداد کلمات کلیدی منتخب از دسته‌های خبری مختلف

جدول ۳ ماتریس درهم‌ریختگی را برای حالت دو کلاسه نمایش می‌دهد. در این آزمایش، از مجموعه داده دوم [۱] استفاده شده و تمام عملیات فقط به ازای داده‌های این دو کلاس انجام شده است. از هر کلاس، تعداد ۱۰۰ داده در کل مورد استفاده قرار گرفته است که در مجموع ۱۴۰ داده برای آموزش و ۶۰ داده برای آزمایش به‌کار گرفته شده است. لازم به ذکر است در این آزمایش، مقدار $\alpha = ۰/۳$ انتخاب شده و تعداد ویژگی‌های بدست آمده برابر با ۳۶۷۴ ویژگی بوده است.

جدول ۳: ماتریس درهم‌ریختگی حالت دو کلاسه

	کلاس اول داده آزمایشی	کلاس دوم داده آزمایشی	کلاس اول داده آموزشی	کلاس دوم داده آموزشی
تعداد واقعی موجود	۲۷	۳۳	۷۳	۶۷
درست مثبت [۲]	۲۶	۱۹	۷۳	۴۶
غلط مثبت [۳]	۱	۱۴	۲۱	۰
نرخ درست مثبت [۴]	۰.۹۶۳۰	۰.۵۷۵۸	۱	۰.۶۸۶۶
نرخ غلط مثبت [۵]	۰.۵۱۸۵	۰.۰۳۰۳	۰.۲۸۷۶	۰

جدول ۴ ماتریس درهم‌ریختگی را برای حالت بیست کلاسه و با داده‌های آموزشی نمایش می‌دهد. در این آزمایش مقدار $\alpha = ۰/۳$ انتخاب شده و تعداد ویژگی‌های بدست آمده برابر با ۱۷۱ ویژگی می‌باشد. همان‌طور که انتظار می‌رفت، از آنجا که تعداد کلاس‌ها در این حالت ۱۰ برابر حالت قبل است، با در نظر گرفتن $\alpha = ۰/۳$ فقط ویژگی‌هایی انتخاب می‌شوند که تفاوت بسیار چشم‌گیری با سایر ویژگی‌های کاندید دارند.

[۱] mini_newsgroups

جدول ۴: ماتریس درهم‌ریختگی حالت بیست کلاسه با داده‌های آموزشی و $\alpha = ۰/۳$

شماره کلاس	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶	۱۷	۱۸	۱۹	۲۰
تعداد واقعی موجود	۶۳	۶۹	۷۰	۷۲	۶۸	۷۰	۷۴	۷۷	۶۹	۶۸	۶۸	۷۴	۶۵	۶۷	۷۲	۷۰	۷۵	۶۸	۷۲	۶۹
درست مثبت	۲۰	۴۳	۲۵	۲۹	۱۴	۱۱	۵	۶۸	۱۹	۱	۴	۲۸	۵	۳۳	۰	۳۸	۳۳	۲۷	۳۵	۲۳
غلط مثبت	۸	۹	۹	۲۵	۹	۴	۷	۷۴۴	۷	۰	۱	۲۰	۱	۲۸	۰	۲۰	۱۱	۱	۲۸	۷

با توجه به نتایج این جدول می‌توان دریافت، تعداد کلمات کلیدی کلاس ۸ که به عنوان ویژگی انتخاب شده‌اند، تفاوت چشم‌گیری نسبت به بقیه ویژگی‌ها داشته‌اند. به همین دلیل، الگوریتم بیشتر پیش‌بینی‌هایش را برابر با کلاس ۸ انجام داده است. علاوه بر این، تعداد کلمات کلیدی از دسته‌های خبری ۱۰، ۱۵، ۱۱، ۷ و ۱۳ که به عنوان ویژگی انتخاب شده‌اند برعکس دسته‌خبری ۸، کم است و الگوریتم به سختی قادر به تشخیص آن‌ها می‌باشد.

جدول ۵ ماتریس درهم‌ریختگی را به ازای داده‌های آزمایشی نشان می‌دهد. تاثیر تعداد کلمات کلیدی انتخاب شده از هر دسته‌خبری در این ماتریس به خوبی مشاهده می‌شود. همان‌طور که قبلا ذکر شد، الگوریتم به سختی قادر به تشخیص داده‌های دسته‌خبری ۱۰ و ۱۵ می‌باشد. در این ماتریس، الگوریتم قادر به تشخیص هیچ‌کدام از داده‌های این دسته نشده است. علاوه بر این، نرخ درست مثبت الگوریتم در دسته‌خبری ۸ بیشترین مقدار خود را دارد و به ۰.۸۷ می‌رسد.

جدول ۵: ماتریس درهم‌ریختگی حالت بیست کلاسه با داده‌های آزمایشی و $\alpha = ۰/۳$

شماره کلاس	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶	۱۷	۱۸	۱۹	۲۰
تعداد واقعی موجود	۳۷	۳۱	۳۰	۲۸	۳۲	۳۰	۲۶	۲۳	۳۱	۳۲	۳۲	۲۶	۳۵	۳۳	۲۸	۳۰	۲۵	۳۲	۲۸	۳۱
درست مثبت	۱۱	۱۲	۴	۱۴	۳	۰	۱	۲۰	۸	۰	۱	۱۲	۳	۱۲	۰	۱۵	۱۰	۱۲	۱۳	۷
غلط مثبت	۲	۳	۵	۸	۴	۱۰	۵	۳۴۷	۶	۰	۱	۸	۰	۱۲	۰	۷	۱۲	۰	۸	۴

۳ تاثیر تعداد ویژگی‌ها

در این بخش، برای بررسی تاثیر تعداد ویژگی‌های انتخابی بر عملکرد الگوریتم، نتیجه دسته‌بندی را با نتایج ارائه شده در بخش قبلی و با مقادیر مختلف برای α مقایسه خواهیم نمود. همان‌طور که قبلا ذکر شد، هرچه مقدار α بزرگتر باشد، تعداد ویژگی‌های انتخاب شده کاهش می‌یابند. در جدول ۳ بخش قبل، ماتریس درهم‌ریختگی برای حالت دو کلاسه نمایش داده شده است. در این جدول، مقدار $\alpha = ۰/۳$ انتخاب شده است. برای حالت دو کلاسه، این مقدار مناسب است. در جدول ۶ نتیجه همان آزمایش، با مقدار $\alpha = ۰/۸$ که منجر به انتخاب ..... ویژگی می‌شود، گزارش شده است.

جدول ۶: ماتریس درهم‌ریختگی برای حالت دو کلاسه با $\alpha = ۰/۸$

	کلاس اول	کلاس دوم
تعداد واقعی		
درست مثبت		
غلط مثبت		
نرخ درست مثبت		
نرخ غلط مثبت		

همان‌طور که مشاهده می‌شود با کاهش تعداد ویژگی‌ها، .....

علاوه بر این، در بخش قبل در جدول ۴ و جدول ۵ به ترتیب، ماتریس‌های درهم‌ریختگی مرتبط با حالت بیست‌کلاسه و $\alpha = ۰/۳$ گزارش شده‌اند. مقدار $\alpha = ۰/۳$ برای حالت بیست‌کلاسه مقداری بزرگ است؛ چون در این حالت، مقادیر احتمال‌های حضور کلمات در دسته‌های خبری کوچم و به هم نزدیک

است. به همین دلیل، بیشتر احتمال‌ها در یک بازه نزدیک به مقدار ۰.۰۵ قرار دارند. این مشکل باعث شده فقط ۱۷۱ کلمه در محدوده قابل قبول قرار بگیرند و به عنوان ویژگی انتخاب شوند.

در این بخش با تعیین مقدار ..... $= \alpha$ تعداد ویژگی‌های انتخاب شده را افزایش می‌دهیم. در این حالت چون تعداد ویژگی‌های انتخاب شده به ..... رسیده است، تعداد کلمات کلیدی انتخاب شده از هر دسته‌خبری در ویژگی‌ها تقریبا با هم برابر می‌شود و انتظار داریم نتایجی بهتر از نتایج ذکر شده در جداول ۴ و ۵ مشاهده نماییم. جدول ۷ ماتریس درهم‌ریختگی را به برای داده‌های آموزشی و جدول ۸ این ماتریس را برای داده‌های آزمایشی، نمایش می‌دهند.

جدول ۷: ماتریس درهم‌ریختگی حالت بیست کلاسه با داده‌های آموزشی و .... $= \alpha$

شماره کلاس	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶	۱۷	۱۸	۱۹	۲۰
تعداد واقعی موجود	۳۷	۳۱	۳۰	۲۸	۳۲	۳۰	۲۶	۲۳	۳۱	۳۲	۳۲	۲۶	۳۵	۳۳	۲۸	۳۰	۲۵	۳۲	۲۸	۳۱
درست مثبت	۱۱	۱۲	۴	۱۴	۳	۰	۱	۲۰	۸	۰	۱	۱۲	۳	۱۲	۰	۱۵	۱۰	۱۲	۱۳	۷
غلط مثبت	۲	۳	۵	۸	۴	۱۰	۵	۳۴۷	۶	۰	۱	۸	۰	۱۲	۰	۷	۱۲	۰	۸	۴

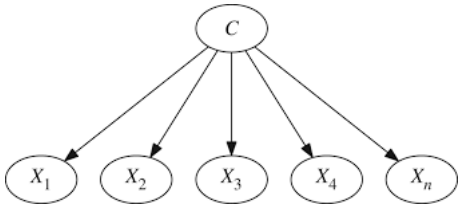
همان‌طور که مشاهده می‌شود.....

جدول ۸: ماتریس درهم‌ریختگی حالت بیست کلاسه با داده‌های آزمایشی و ..... $= \alpha$

شماره کلاس	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶	۱۷	۱۸	۱۹	۲۰
تعداد واقعی موجود	۳۷	۳۱	۳۰	۲۸	۳۲	۳۰	۲۶	۲۳	۳۱	۳۲	۳۲	۲۶	۳۵	۳۳	۲۸	۳۰	۲۵	۳۲	۲۸	۳۱
درست مثبت	۱۱	۱۲	۴	۱۴	۳	۰	۱	۲۰	۸	۰	۱	۱۲	۳	۱۲	۰	۱۵	۱۰	۱۲	۱۳	۷
غلط مثبت	۲	۳	۵	۸	۴	۱۰	۵	۳۴۷	۶	۰	۱	۸	۰	۱۲	۰	۷	۱۲	۰	۸	۴

۴ بررسی استقلال شرطی ویژگی‌ها به شرط دانستن کلاس

در مدل بیز ساده، فرض بر این است که تمامی ویژگی‌های استفاده شده، به شرط دانستن کلاس، از یک‌دیگر مستقل هستند و تنها استقلال شرطی موجود همین است. گراف معادل برای توزیع احتمالی با این ویژگی معادل شکل ۱ می‌باشد.



شکل ۱: گراف معادل با توزیع احتمالی با شرط مفروض در مدل بیز ساده.

فرض مذکور در این مدل را می‌توان با رابطه ۶ نمایش داد. برای اثبات درستی یا نادرستی فرض، به ازای تمام زوج‌ویژگی‌های موجود، باید دو طرف این رابطه را جداگانه محاسبه و سپس نتایج را باهم مقایسه کرد.

$$\forall i \neq j; (X_i \perp X_j|C) \longleftrightarrow P(X_i, X_j|C) = P(X_i|C) \cdot P(X_j|C) \quad (۶)$$

برای اثبات یا رد فرض، مقادیر $P(X_i|C)$ را به ازای تمام ویژگی‌ها باید محاسبه کرد که تمام توزیع‌های احتمالی شرطی موجود بین یک ویژگی و یک دسته‌خبری قبلا محاسبه شده‌اند. برای محاسبه $P(X_i, X_j|C)$ نیز از ماتریس شاخص که در بخش پیش‌پردازش توضیح داده شد، استفاده می‌کنیم.

۵ تاثیر تعداد داده‌های آموزشی

۶ توضیحات

* سورس کد مربوط به پروژه در ضمیمه این گزارش ارسال شده است. همین‌طور این کد از این لینک ، قابل دریافت می‌باشد.

* آدرس لینک برای دریافت کد:

https://github.com/ahmad-asadi/PGM/tree/master/BayesianNetwork