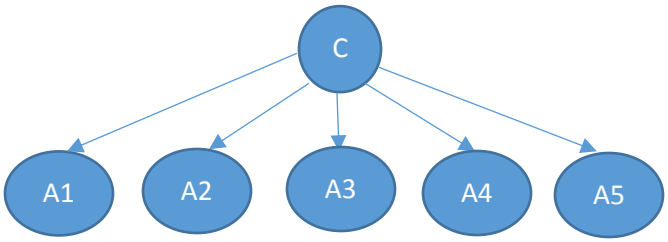
Statistical Learning: Homework 2
Due March 20th (Esfand 30th), 23:59 pm
Naïve Bayes Classifier

۱- داده‌های آموزش در فایلی به نام trainingData.txt به همراه این تمرین موجود می‌باشد. تعداد این داده‌ها ۲۰۰ عدد است. هر ردیف در این فایل بیانگر یک نمونه داده است. اولین ستون در این فایل متغیر کلاس C و دومین تا ششمین ستون ویژگی‌های A1, A2, A3, A4, A5 هستند. داده‌های آزمون نیز در فایلی به نام testingData.txt قرار داده شده است. تعداد این داده‌ها ۱۰۰ عدد است که فرمت آن همانند داده‌های آزمایش می‌باشد. شبکه naïve Bayes زیر را در این تمرین در نظر بگیرید:



توزیع‌های احتمال شرطی در شبکه naïve Bayes را با استفاده از تخمین درست‌نمایی بیشینه (MLE) بر روی داده‌های آموزش بدست آورید. بر مبنای شبکه بیزی فوق می‌توان نمونه‌ها را با استفاده کردن از قاعده بیز دسته‌بندی کرده، احتمالات شرطی به شکل $P\ (C\ |A_1, A_2, A_3, A_4, A_5)$ را بدست آورده و برچسب کلاس برای داده‌ها را پیش‌بینی کرد. در این قسمت می‌بایست پارامترهای $P(C)$ و $P(A_{i\in\{1,\dots,5\}}|C)$ را بدست آورده و درصد خطای دسته‌بندی را بر روی مجموعه داده آزمون اعلام نمایید. (توجه: ویژگی‌های موجود در این مساله از جنس گسسته هستند)

۲-

- **شرح مساله**

بر پایه داده‌های خرید پیشین یک فروشگاه اینترنتی، مطلوب است مدلی آموزش داده شود که احتمال پس فرستادن یک خرید مشخص را بر پایه داده‌های خرید جدید فروشگاه پیش‌بینی نماید. داده‌های آموزشی شامل داده‌های خرید، ارسال و ویژگی‌های مختلف محصول و خریدار می‌باشد. کلاس هر داده آموزشی شامل 1 (پس فرستادن کالا) و 0 (نگهداری کالا) می‌باشد.

- **کار موردنظر**

در این تمرین داده‌های واقعی مربوط به یک سال فروشگاه اینترنتی در اختیارتان قرار می‌گیرد. (تقریبا شامل ۴۸۱۰۰۰ سفارش است.) با استفاده از این داده‌ها مدلی هوشمند، با استفاده از روش *Naïve Bayse*، طراحی کنید که قابلیت پیش-بینی بازگشت محصول یا نگهداری آن را داشته باشد. کلاس مربوط به این داده شامل 0 (نگهداری محصول) و ۱ (پس فرستادن محصول) می‌باشد(ستون آخر دیتاست). برای خریدهای یک ماه (حدود ۵۰۰۰۰ سفارش) باید کلاس مربوط به محصول را پیش‌بینی نمایید. بدین منظور برای هر سفارش لازم است پیش‌بینی صورت گیرد که مقدار آن بین بازه [0,1] می‌باشد. هر چه مقدار پیش‌بینی بالاتر باشد احتمال بازگشت بیشتر است.

- **ارزیابی**

خطای ارزیابی در این مدل باید با استفاده از رابطه زیر محاسبه شود:

$$E = \sum_{i} |return\ shipment - prediction|$$

prediction یک عدد حقیقی بین صفر تا یک می‌باشد و return shipment عدد قطعی صفر یا یک می‌باشد.

- **چالش‌ها**

۱. برخی از ویژگی‌های این داده شامل اندازه و رنگ بصورت رشته است، برخی دیگر از ویژگی‌ها مانند تاریخ سفارش بصورت تاریخ است. بدین منظور برای باز کردن فایل بهتر است از دستوراتی مانند textscan و یا textread در متلب استفاده کنید و یا از زبان برنامه‌نویسی دیگر مثل پایتون استفاده نمایید.

۲. در این دیتاست ۱۴ ویژگی وجود دارد. (ویژگی‌ها، نوع آنها و توصیف آنها در یک فایل جداگانه در اختیار شما قرار می‌گیرد.) برخی از ویژگی‌ها مثل قیمت و تاریخ بصورت پیوسته است و برخی از ویژگی‌ها گسسته است. برای راحتی کار می‌توانید ویژگی‌های پیوسته را گسسته کنید. بدین منظور باید با مطالعه دقیق دیتا عمل گسسته‌سازی را به صورتی انجام دهید که منطقی و باعث کاهش دقت مدل نشود. به عنوان مثال بازه قیمت محصولات را می-توانید به چند بازه تقسیم کنید و به هر بازه یک عدد نسبت دهید. همینطور می‌توانید ویژگی تاریخ را بدین صورت گسسته نمایید.

اگر داده‌ها را گسسته نکنیم و از نایو بیز برای ویژگیهای پیوسته با فرض توزیع نرمال استفاده کنیم، نتایج چقدر با یکدیگر متفاوت خواهند بود؟ نتایج را با یکدیگر مورد مقایسه قرار دهید.

کلیه مراحل پیاده‌سازی شامل خواندن دیتاست، گسسته‌سازی ویژگی‌های پیوسته و نحوه ارزیابی باید در گزارش ذکر شود.

موفق باشید