بسمه تعالی

پروژه آخر درس انتقال داده - دانشگاه صنعتی نوشیروانی بابل - مبحث تئوری اطلاعات

مدرس: کاظمی‌تبار

در این پروژه قصد داریم جنبه‌ای از تابع اطلاعات متقابل را که برای سنجش میزان وابستگی ویژگی‌ها به متغیر هدف استفاده می‌شود مورد مطالعه قرار دهیم. به خاطر می‌آوریم که

$$\mathrm{I}(X;Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X, Y)$$

که در آن

$$\mathrm{H}(X, Y) = -\sum\sum P(x, y) \log_2 [P(x, y)]$$

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log \mathrm{P}(x_i)$$

در پروژه حاضر ما با مجموعه داده بازماندگان کشتی تایتانیک سر و کار داریم. در فایل اکسلی که در اختیار شما قرار داده شده اطلاعاتی از مسافران همچون نام، جنسیت، سن، تعداد همراهان از نوع والدین و فرزندان، تعداد همراهان از نوع برادروخواهریا همسر، طبقه خاص بلیطی که فروخته شده و نیز بندری که مسافر از آن وارد کشتی شده درج شده است. به علاوه جلوی نام هر مسافر درج شده که آیا زنده مانده یا خیر. شما با ستون نام مسافر کاری ندارید. شما می‌خواهید بدانید چه عواملی در نجات یافتن مسافران نقش بیشتری داشته‌اند. مثلا آیا این درست که گفته شود جنسیت مسافر در احتمال نجات او موثر بوده؟ یا اینکه آیا طبقه بلیط یا نوع بندری که مسافر از آن سوار شده به نوعی باعث تبعیض بین مسافران هنگام نجات آنها شده یا خیر. برای پاسخ به این

سوالات شما کافی است اطلاعات متقابل بین هر عامل را با متغیر هدف (نجات یافتن یا نیافتن) بیابید. به عبارت دیگر شما باید برای هر کدام از عوامل سن، جنسیت، طبقه، بندر و تعداد همراهان (برای هر کدام از دو نوع ذکر شده) یک عدد به عنوان اطلاعات متقابل آن عامل با متغیر نجات یافتن پیدا می‌کنید. سپس این اعداد را به طور نزولی مرتب می‌کنید و می‌بینید که کدامیک از عوامل تاثیر بیشتری در نجات یافتن افراد داشته است. تحویلی این تمرین برنامه ای است که می‌نویسید به علاوه جدول نزولی تاثیر عوامل. برای نوشتن برنامه هم می‌توانید پیاده سازی خودتان از توابع آنتروپی و اطلاعات متقابل را داشته باشید و یا از کتابخانه‌های موجود استفاده کنید .به جای نوشتن برنامه، اکسل هم قابل قبول است

نکاتی که قبل از اعمال تابع آنتروپی خوب است به آن دقت کنید:

۱. بدیهی است که منظور از x عامل مورد مطالعه و منظور از y متغیر هدف (نجات یافتگی) است. شما باید نهایتا ۶ مقدار اطلاعات متقابل برای ۶ عامل مورد مطالعه محاسبه کنید.

۲. اگر عامل مورد مطالعه شما از نوع عددی باشد لازم است آن را بازه بندی کنید. به عنوان مثال سن افراد را در بازه‌های زیر دوسال، سه تا سیزده سال، چهارده تا بیست و چهار سال، بیست و پنج تا پنجاه سال و پنجاه و یک سال به بالا در نظر بگیرید.

۳. برای محاسبه p(x) باید تعداد رخدادهای x مورد نظر را در کل ستون مربوطه بشمارید و بر تعداد کل تقسیم نمایید. مثلا احتمال مرد بودن مساوی است با تعداد کل مردان تقسیم بر تعداد کل مسافران.

۴. برای محاسبه p(x,y) لازم است از فرمول احتمال شرطی استفاده کنید. یعنی $P(x,y) = P(x|y).P(y)$

برای محاسبه احتمال شرطی هم کافی است فضای احتمال را تغییر دهید. به عنوان مثال احتمال مرد بودن به شرط زنده ماندن مساوی است با تعداد مردانی که نجات یافتند تقسیم بر تعداد کل نجات یافتگان .

۵. اگر یک جا سن مسافری نوشته نشده بود، می‌توانید سن را برابر میانه سن کل مسافران قرار دهید

Variable	Definition	Key
Survival	Survival	0 = No, 1 = Yes
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Gender	Male or Female	
Age	Age in years	
Sibsp	# of siblings / spouses aboard the Titanic	
Parch	# of parents / children aboard the Titanic	
Embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Variable Notes

pclass: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.