کوییز شماره دو- درس پردازش زبان طبیعی
تاریخ: ۱۴۰۲/۱/۲۱

سوال ۱

یک نشریه قصد دارد تا از میان گزارش های دریافتی اخبار مرتبط و غیر مرتبط را پیدا کند. برای این هدف نشریه وجود یا عدم وجود ۵ کلمه به خصوص technology, politics, business, sports و entertainment را بررسی می‌کند. جدول ۱ پایگاه داده‌ای متشکل از ۶ نمونه آموزش و یک نمونه تست را نشان می‌دهد که هر نمونه مربوط به یک گزارش خبری است.

سلول های با مقدار ۱ از جدول نماینده وجود کلمه مربوط به هر ستون در هر گزارش و ۰ نشان دهنده عدم وجود آن است.

No. Report	Entertainment	Sports	Business	Politics	Technology	Relevant/Irrelevant
1	1	1	0	0	0	YES
2	0	1	1	0	1	YES
3	0	0	0	1	1	NO
4	0	0	1	0	1	NO
5	1	0	1	1	0	YES
6	1	0	0	1	0	?

با استفاده از طبقه بندی Naive Bayes مشخص کنید که گزارش شماره ۶ باید به عنوان اخبار مرتبط برای درج در نشریه انتخاب شوند یا خیر. (از روش add-1 smoothing در محاسبات خود استفاده کنید) (۴ نمره)

پاسخ:

$$P(YES|F1, F2, F3, F4, F5)$$
$$=P(F1=1|YES)*P(F2=0|YES)*P(F3=0|YES)*P(F4=1|YES)*P(F5=0|YES)*P(YES)$$

$$P(NO|F1, F2, F3, F4, F5)$$
$$=P(F1=1|NO)*P(F2=0|NO)*P(F3=0|NO)*P(F4=1|NO)*P(F5=0|NO)*P(NO)$$

$$P(YES)=(3+1)/(5+2)=0.57$$
$$P(F1=1|YES)=(2+1)/(3+2)=0.6$$
$$P(F2=0|YES)=(1+1)/(3+2)=0.4$$

P(F3=0|YES)=(1+1)/(3+2)=0.4
P(F4=1|YES)=(1+1)/(3+2)=0.4
P(F5=0|YES)=(2+1)/(3+2)=0.6

P(NO)=(2+1)/(5+2)=0.43
P(F1=1|NO)=(0+1)/(2+2)=0.25
P(F2=0|NO)=(2+1)/(2+2)=0.75
P(F3=0|NO)=(1+1)/(2+2)=0.5
P(F4=1|NO)=(1+1)/(2+2)=0.5
P(F5=0|NO)=(2+1)/(2+2)=0.75

P(YES|F1, F2, F3, F4, F5)=0.57*0.6*0.4*0.4*0.4*0.6=0.013
P(NO|F1, F2, F3, F4, F5)=0.43*0.25*0.75*0.5*0.5*0.75=0.015

سوال ۲

روش های Naive Bayes و logistic Regression هر دو جز طبقه بند های محبوب هستند. با این وجود وابسته به ویژگی های داده در هر مسئله ممکن است عملکرد یکی از این دو روش از دیگری بهتر باشد. بر این اساس تاثیر هر یک از موارد زیر را بر عملکرد این دو طبقه بند بررسی کنید. سپس تعیین کنید در حالتی که اندازه دادگان کوچک، ابعاد بردار های ویژگی بالا و همبستگی بین ویژگی ها زیاد باشد کدام طبقه بند بهتر عمل میکند؟

۱. اندازه دادگان (تعداد نمونه ها) (۴ نمره)

پاسخ) از آنجایی که طبقه بند Naive Bayes بین ویژگی ها فرض استقلال در نظر می‌گیرد، زمانیکه تعداد سمپل ها نسبت به تعداد ویژگی ها کمتر است میتواند مدل robust تری نسبت به نفرین ابعاد بالا (curse of dimensionality) ارائه کند. در حالیکه مدل logistic regression به دلیل آنکه فرض های کمتری روی داده‌ها در نظر می‌گیرد، میتواند روابط پیچیده‌تری بین آن‌ها در صورتی که تعداد داده ها بیشتر باشد پیدا کند.

۲. ابعاد بردار های ویژگی نمونه ها (۴ نمره)

پاسخ) زمانیکه تعداد ویژگی ها در مقایسه با تعداد داده ها افزایش می‌یابد مدل Naive Bayes می‌تواند عملکرد بهتری در برخورد با پیچیدگی دادگان داشته باشد زیرا نسبت به نفرین ابعاد بالا robust تر است. در حالیکه مدل logistic regression روابط بین ویژگی ها و برچسب کلاس ها را مستقیما مدل می‌کند.

پس زمانیکه تعداد ویژگی ها افزایش پیدا می‌کند مدل پیچیده تر و تفسیر آن دشوارتر می‌شود. این امر می‌تواند زمانیکه همبستگی بین داده ها بالاست موجب بروز مشکل بیش برازش (overfitting) شده و عملکرد را تضعیف کند.

۳. همبستگی (correlation) بین ویژگی ها (۴ نمره)

پاسخ) از آنجایی که مدل Naive Bayes فرض استقلال بین ویژگی ها در نظر می‌گیرد، زمانیکه ویژگی‌ها واقعا مستقل باشند یا همبستگی بین آن ها کم باشد این مدل عملکرد خوبی دارد و بر عکس اگر ویژگی ها همبستگی بالایی داشته باشند این فرض صحیح نیست و عملکرد این مدل تضعیف می‌شود. در حالیکه مدل logistic regression می‌تواند روابط پیچیده بین ویژگی‌ها را بهتر مدل کند که باعث می‌شود اگر همبستگی بین ویژگی‌ها بالا باشد عملکرد نسبت به مدل Naive Bayes بهتر باشد.

با توجه به توضیحات بالا، در شرایطی که اندازه دادگان کوچک، ابعاد بردار های ویژگی بالا و همبستگی بین ویژگی ها زیاد باشد عملکرد مدل Naive Bayes بهتر است.

سوال 3

در یک مسئله binary classification با ۲ کلاس، ماتریس آشفتگی (confussion matrix) زیر برای ۱۰۰ نمونه تست بدست آمده است. (کلاس Present را کلاس مثبت در نظر بگیرید)

		Predicted	
		Present	Absent
Actual	Present	60	10
	Absent	20	10

مقادیر Micro Average و Macro Average را برای معیار F1-score با در نظر گرفتن هر دو کلاس محاسبه کنید. (۴ نمره)

پاسخ:

Macro Average F1-score:

Precision for Present = true positives for Present / (true positives for Present + false positives for Present) = 60 / (60 + 20) = 0.75

Precision for Absent = true positives for Absent / (true positives for Absent + false positives for Absent) = 10 / (10 + 10) = 0.5

Recall for Present = true positives for Present / (true positives for Present + false negatives for Present) = 60 / (60 + 10) = 0.8571

Recall for Absent = true positives for Absent / (true positives for Absent + false negatives for Absent) = 10 / (10 + 20) = 0.3333

F1-score for Present = 2 * (precision for Present * recall for Present) / (precision for Present + recall for Present) = 2 * (0.75 * 0.8571) / (0.75 + 0.8571) = 0.7990

F1-score for Absent = 2 * (precision for Absent * recall for Absent) / (precision for Absent + recall for Absent) = 2 * (0.5 * 0.3333) / (0.5 + 0.3333) = 0.4

Macro Average F1-score = (0.799+0.4)/2 = 0.6

Micro Average F1-score:

TP = 60 + 10 = 70

FP = 20 = 20

FN = 10 = 10

Micro-average precision = TP / (TP + FP) = 70 / (70 + 20) = 0.7778

Micro-average recall = TP / (TP + FN) = 70 / (70 + 10) = 0.875

Micro-average F1-score = 2 * (0.7778 * 0.875) / (0.7778 + 0.875) = 0.8235