به نام پروردگار

پردازش زبان طبیعی تمرین اول

موعد تحویل :

(alipoor)

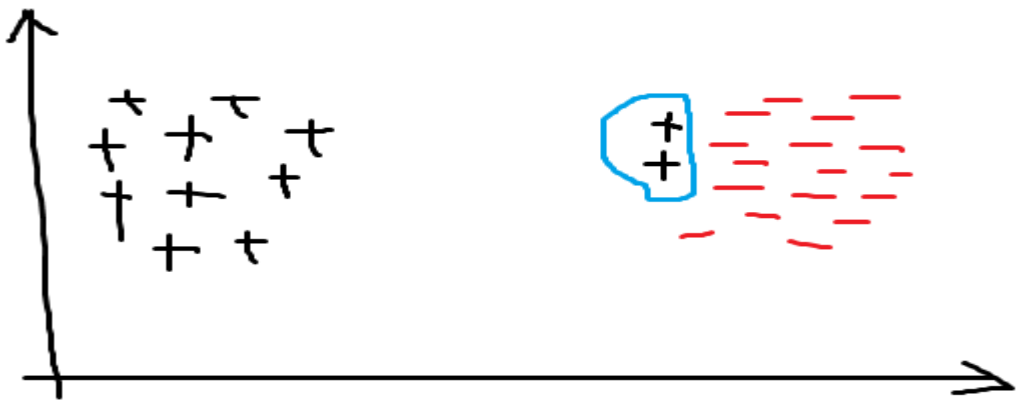
۱- به سوالات زیر پاسخ مناسب دهید

۱. می‌دانیم Naïve Bayes و Logistic regression هر دو جزء طبقه‌بندهای خطی هستند، هنگامی که دادگان آموزش ما کم هستند به نظر شما کدام طبقه‌بند بهتر عمل می‌کند؟ هنگامی که دادگان آموزش زیاد هستند چطور؟ چرا؟

۲. مدل‌سازی Naïve Bayes به این صورت است که سند را به صورت BOW مدل‌سازی می‌کند، مزایا و معایب این روش مدل‌سازی را عنوان کنید. اگر در یک تسک شرط Conditional Independence وجود داشته باشد(یعنی ویژگی‌ها به شرط برچسب مستقل از هم باشند) آیا Naïve Bayes می‌تواند یک طبقه‌بند خوب باشد؟

۳. می‌توانید یک مثال بزنید که Naïve Bayes باعث تخمین بیش از اندازه (Overestimate) می‌شود؟ به عبارت دیگر: $P(f1, f2|class) < P(f1|class) \times P(f2|class)$

۴. گفته شد که Naïve Bayes یک طبقه‌بند خطی است. به نظر شما در شکل زیر این طبقه‌بند دو آیتمی را که دور آن خط آبی کشیده شده است جزء کلاس مثبت درنظر می‌گیرد یا کلاس منفی؟ کمی توضیح دهید.



۵. راجع‌به مشکلات مربوط به نامتوازن بودن دیتاست (imbalanced dataset) و روش‌های مقابله با آن تحقیق کنید و چند مورد را ذکر کنید؟

۶. تفاوت development test set (devtest) را با test set بیان کنید؟

۲- فرض کنید که اسناد زیر داده شده‌اند:

سند	برچسب
fun, couple, love, love	Comedy
fast, furious, shoot	Action
couple, fly, fast, fun, fun	Comedy
furious, shoot, shoot, fun	Action
fly, fast, shoot, love	Action

در صورت استفاده از طبقه‌بند Naïve Bayes به همراه هموارسازی add-1 به نظر شما سند زیر چه برچسبی خواهد گرفت؟

fast, couple, shoot, fly

۳ – جدول زیر ماتریس آشفتگی را برای سه کلاس نشان می‌دهد دقت و فراخوانی را برای هرکدام از کلاس‌ها حساب کنید:

		برچسب واقعی		
		A	B	C
خروجی سیستم	A	۱۰	۱۲	۲
	B	۵	۷۰	۸۰
	C	۸	۹	۲۰۰

لطفا به قواعد حل تمرین که در CECM قرار داده شده است توجه کنید.