فصل سیزده : دسته بندی متن

۱- الف) اگر در یک مدل دسته بندی دو کلاسه، مثل دسته-بندی اخبار به دو دسته‌ی اخبار ورزشی و سیاسی، اگر تعداد کلماتی که به عنوان ویژگی انتخاب می‌شوند بیشتر متعلق به یک دسته باشند چه تاثیری در نتیجه دسته‌بندی دارد؟ توضیح دهید.

پاسخ :

اگر بیشتر ویژگی ها را از کلاس اول انتخاب کنیم: این حالت مانند این است که مدل تنها برای تشخیص کلاس اول آموزش ببیند. در استفاده از این مدل برچسب دوم به درستی تشخیص داده نمی شود. چون مدل برای متون این دسته آموزشی ندیده است ،به هر متنی به غیر از متون کلاس اول،برچسب کلاس دوم را میزند.

ب) دلیل سادگی یادگیری مدل بیز ساده را بیان کنید؟

پاسخ :

چون در مدل بیز ساده تنها احتمال رخداد یک کلمه از سند درنظر گرفته می‌شود و احتمال کلمات متوالی در اسناد موجود در یک دسته در نظر گرفته نمی‌شود. به بیان دیگر در مدل بیز ساده برای سادگی کار، فرض می-شود کلمات یک سند از یکدیگر مستقل هستند در حالی که چنین فرضی اشتباه است و کلمات بهم وابسته اند.

۲-با اجرای الگوریتم بیز ساده بر روی داده‌های جدول شماره ۱ ، دسته‌ی دو سند بدون برچسب را تعیین کنید ؟ فرض کنید اندازه‌ی لغتنامه برابر با ۴ است. (نوشتن تمامی محاسبات الزامی است)

جدول شماره ۱

سنـد	بـرچسب کـلاس
AACBBAABABBAACCBB	U
AAAAAACBBBBBBBABA	U
CACCCABCABBCCABCC	W
CCCCABACBCCBACCAB	W
AAACCABBBBBBAABBA	?
CACBCDCCCABBACDAB	?

پـاسخ :

$P(U) = 1/2 , P(W) = 1/2$

$$P(A|U) = \frac{15+1}{15+4+15+4} = \frac{16}{38} \quad , P(B|U) = \frac{15+1}{15+4+15+4} = \frac{16}{38} \quad ,$$

$$P(C|U) = \frac{4+1}{15+4+15+4} = \frac{5}{38} \quad , P(D|U) = \frac{1}{15+4+15+4} = \frac{1}{38} \quad ,$$

$$P(A|W) = \frac{8+1}{8+18+8+4} = \frac{9}{38} \quad , P(B|W) = \frac{8+1}{8+18+8+4} = \frac{9}{38} \quad ,$$

$$P(C|W) = \frac{18+1}{8+18+8+4} = \frac{19}{38} \quad , P(D|W) = \frac{1}{8+18+8+4} = \frac{1}{38}$$

AAACCCBBBBBBAABBA = X

$$P(X|U) = \log P(U) + 7 * \log P(A|U) + 2 * \log P(C|U) + 8 * \log P(B|U)$$

$$= -0.3 + 7 * -0.38 + 2 * -0.88 + 8 * -0.38 = -7.76$$

$$P(X|W) = \log P(W) + 7 * \log P(A|W) + 2 * \log P(C|W) + 8 * \log P(B|W)$$

$= -0.3 + 7 * -0.62 + 2 * -0.3 + 8 * -0.62 = -10.2$

→ $X \in U$

CACBCDCCCABBACDAB = Y

$P(Y|U) = \log P(U) + 4 * \log P(A|U) + 7 * \log P(C|U) + 4 * \log P(B|U) + 2 * \log P(D|U)$

$= -0.3 + 4 * -0.38 + 7 * -0.88 + 4 * -0.37 + 2 * -1.57 = -12.6$

$P(Y|W) = \log P(W) + 4 * \log P(A|W) + 7 * \log P(C|W) + 4 * \log P(B|W) + 2 * \log P(D|W)$

$= -0.3 + 4 * -0.62 + 7 * -0.3 + 4 * -0.62 + 2 * -1.57 = -10.5$

→ $Y \in W$

فصل چهارده : دسته بندی برداری

۳-الف) با استفاده از الگوریتم دسته بندی Rocchio و با توجه به جدول شماره ۲ کلاس d_6, d_7 را مشخص کنید. (مقادیر جدول، مقادیر tf-idf هستند)

جدول شماره ۲

	آموزشی			کلاس c		سند جدید	
کلمات	d_1	d_2	d_3	d_4	d_5	d_6	d_7
شنا	1	0.8	0.7	0	0.1	0.5	0.1
شیرجه	0.5	0.6	0.4	0.1	0	0.6	0
بازی	0.44	0	0	2.64	1.76	0.22	2.64
غرق	0	0.22	0	4.4	12.32	0	0.88

پاسخ :

این نوع دسته بندی بین هر دو کلاس با توجه به ابعاد فضا یک ابر صفحه تعیین می‌کند.

$$\sum_{i=1}^{M} w_i d_i = \theta$$

$$\vec{w} = \vec{\mu}(c_1) - \vec{\mu}(c_2)$$

$$\theta = 0.5 \times (|\vec{\mu}(c_1)|^2 - |\vec{\mu}(c_2)|^2)$$

بردار مرکز کلاس c : $\begin{pmatrix} 0.5 \\ 0.5 \\ 2.2 \\ 8.36 \end{pmatrix}$

بردار مرکز کلاس 1 : $\begin{pmatrix}1.25\\0.75\\0.15\\0.07\end{pmatrix}$

$$w = \begin{pmatrix}0.75\\0.25\\2.05\\8.29\end{pmatrix}$$

$$\theta = 73.07$$

$$wd1 - \theta < 0 . wd2 - \theta < 0 . wd3 - \theta < 0 . wd4 - \theta > 0 . wd5 - \theta > 0$$

بردارهایی که با جایگذاری در معادله ابرصفحه حاصل منفی می دهند(از ابر صفحه کوچکتر هستند)،متعلق به کلاس 1 هستندو بردارهایی که با جایگذاری در معادله ابرصفحه حاصل مثبت می دهند(از ابر صفحه کوچکتر هستند)،متعلق به کلاس c هستند .

$$wd6 - \theta < 0$$

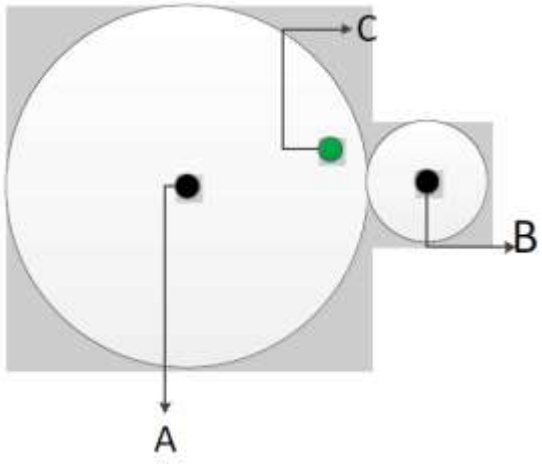
این‌مدل دسته بندی به سند ۶ برچسب 1 می دهد .

$$wd7 - \theta < 0$$

این‌مدل دسته بندی به سند ۷برچسب 1 می دهد .

ب) با یک مثال نشان دهید در الگوریتم دسته‌بندی Rocchio چگونه ممکن است یک سند به دسته ای تعلق گیرد که متفاوت از برچسب آموزشی آن است.

پاسخ:



در مثال بالا طبق برچسب های آموزشی، سند C به کلاس A متعلق است. اما فاصله‌ی این سند از مرکز کلاس B کمتر است و براساس الگوریتم Rocchio برچسب B می‌گیرد.

ج) با استفاده از الگوریتم k-نزدیکترین و داده‌های جدول شماره ۲تعیین کنید سندهای جدید در کدام دسته قرار میگیرد.

I. به ازای k = 1

II. به ازای k = 3

پاسخ :

$|d6 - d1| = 0.55$

$|d6 - d2| = 0.39$

$|d6 - d3| = 0.28$

$|d6 - d4| = 26.2$

$|d6 - d5| = 154.9$

$|d7 - d1| = 6.7$

$|d7 - d2| = 7.5$

$|d7 - d3| = 7.58$

$|d7 - d4| = 6.6$

$|d7 - d5| = 12.32$

نزدیکترین سند به سند6d ،سند3d و به سند d7 ،سند d4 است بنابراین برای k=1 ،d6 برچسب l و d7 برچسب c می‌گیرد.

برای k=3 ، d1,d2,d3 ،نزدیکترین ها به d6 هستند و همه متعلق به یک کلاس هستند بنابراین برچسب d6 هم l می‌شود .

برای k=3 ، d1,d2,d4 ،نزدیکترین ها به d7 هستند بنابراین برچسب d7 l می‌شود .

۴-با ذکر چند مثال عملکرد دو الگوریتم دسته بندی Rocchio و k-نزدیکترین را در کلاس‌های دو یا چند تکه مقایسه کنید.

پاسخ :

کلاس های چند تکه کلاس هایی هستند که داده های آن ها به چند شکل مختلف ظاهر می شوند، اگر بتوانیم این داده ها را نمایش دهیم در یک ناحیه متمرکز نیستند ودر چندقسمت مختلف از صفحه مختصات دیده می شوند . برای چنین داده هایی الگوریتم Rocchio مناسب نیست به این دلیل که این الگوریتم تنها با مرکز داده ها کار می کند و به دلیل وضعیت این کلاس ها مرکز داده ها نماینده خوبی برای

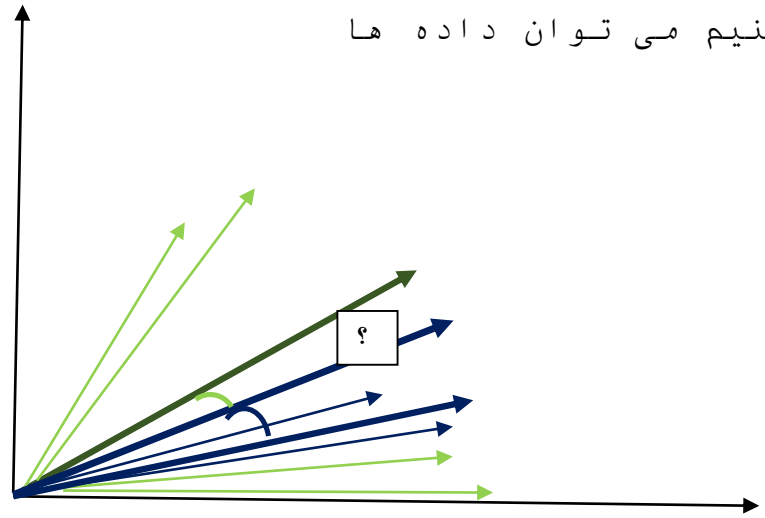
تمام اعضای کلاس نیست. در این حالت ممکن است که مرکز داده های یک کلاس در مجاورت داده های کلاس‌های دیگر قرار بگیرند.

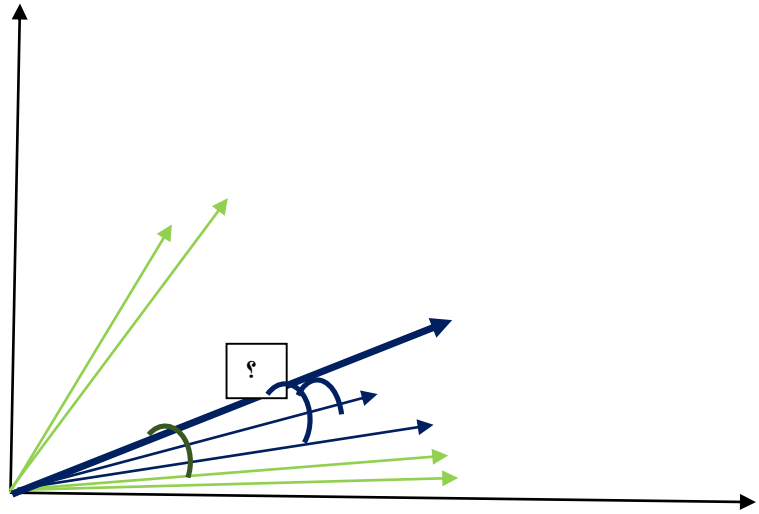
اما در الگوریتم k نزدیکترین به این دلیل که بر اساس نزدیکترین داده ها به هر داده آزمایشی تصمیم گرفته می-شود برای کلاس های چند تکه عملکرد بهتری دارد.

مثال :

در شکلاعضای کلاس قرمز در دو قسمت مختلف پراکنده هستند و مرکز آن ها به شکلی که در تصویر می بینند به دست آمده است .

اگر از الگوریتم Rocchio استفاده کنیم ، یک قسمت از داده ها که به مرکز کلاس آبی نزدیکتر هستند به اشتباه برچسب می خورند. اما اگر از الگوریتم k نزدیکترین با مقدار ۳ استفاده کنیم می توان داده ها





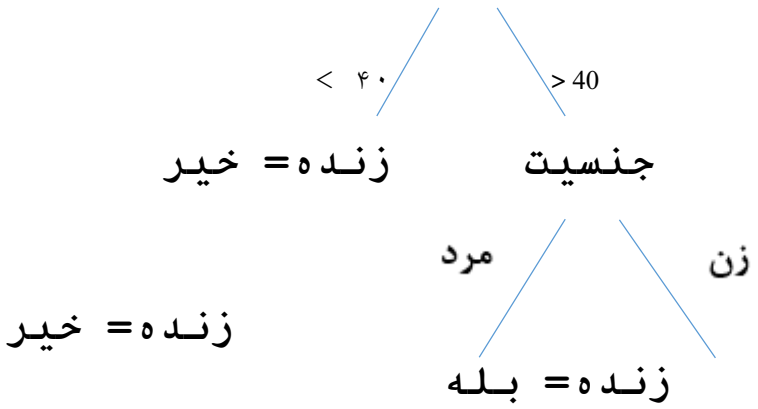
فصل پانزده : یادگیری ماشین روی اسناد

۵-درخت تصمیم مناسب برای اطلاعات جدول شماره ۳ را بسازید.

جدول شماره ۳

زنده	موقعیت کابین	نوع بلیت	جنسیت	سن	شماره مسافر
خیر	نیمه شمالی	ویژه	مرد	< ۴۰	1
خیر	نیمه جنوبی	عادی	مرد	<40	2
خیر	نیمه جنوبی	عادی	زن	>40	3
بله	نیمه شمالی	ویژه	مرد	>40	4
خیر	نیمه شمالی	عادی	زن	<40	5
بله	نیمه شمالی	عادی	مرد	>40	6

پاسخ :



فصل شانزده : خوشه‌بندی

۶- الف) مثالی ارائه دهید که در آن الگوریتم k-means خوشه های نامتوازن تولید کند که یک خوشه خیلی کوچک و یک خوشه خیلی بزرگ باشد.

پاسخ :

اگر نقاط مرکزی اولیه به صورتی انتخاب شوند که تعدادی از آنها از میان داده های خارج از محدودهی اکثر داده ها (outliers) انتخاب شوند، ممکن است باعث شود تقارن خوشهبندی بهم بریزد .در این حالت دادههای خارج از محدوده که تعداد کمی دارند برای خود خوشه جداگانهایی ایجاد میکنند و این باعث میشود که یک خوشه بزرگ و خوشهی دیگر بسیار کوچک شود.

ب) دو مورد از شرطهای خاتمه الگوریتم k-means به صورت زیر است:

- مرکز خوشه ها تغیر نکند.
- برچسب ها تغیر نکند.

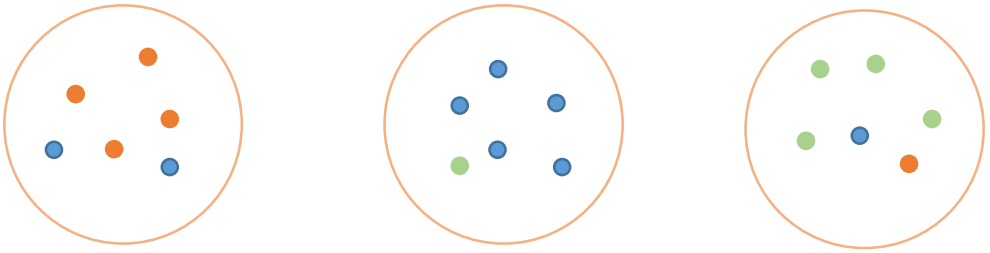
ایا این دو شرط یکدیگر را تضمین میکنند؟ توضیح سدهید.

پاسخ :

بله. اگر برچسبها تغییر نکنند مراکز هم تغیر نمی-کنند و اگر مراکز تغیر نکنند یعنی برچسبها تغیر نکرده اند.

۷- الف) در شکل زیر خروجی حاصل از یک الگوریتم خوشه-بندی آمده است. برای بررسی میزان عملکرد این الگوریتم معیارهای خواسته شده را محاسبه کنید.

- Precision
- Recall
- Rand index



خوشه ۳

خوشه ۲ خوشه ۱

پاسخ :

$$TP + FP = \binom{6}{2} + \binom{6}{2} + \binom{6}{2} = 45$$

$$TP = \binom{4}{2} + \binom{2}{2} + \binom{5}{2} + \binom{4}{2} = 23 \quad \Rightarrow FP = 45 - 23 = 22$$

$$TN = \binom{4}{2} + \binom{8}{2} + \binom{11}{2} = 89$$

$$TP + TN + FP + FN = \binom{18}{2} = 153$$

$$FN = 153 - 23 - 22 - 89 = 19$$

=>

Precision = TP/TP+FP = 23 / 45

Recall = TP/TP+FN = 23 / 44

Rand index = TP+TN/TP+FP+FN+TN = 112 / 153

ب) تفاوت معیار accuracy , Rand index در چیست؟

پاسخ:

RI به اسم خوشه حساس نیست ولی accuracy به اسم خوشه حساس است.