به نام خدا

بازیابی پیشرفته اطلاعات

نیم‌سال اول ۰۳-۱۴۰۲

مدرس: دکتر حمید بیگی

دانشکده‌ی مهندسی کامپیوتر

کوییز سوم (۱۰۰ نمره) | زمان: ۲۵ دقیقه | ۴ آذر

سوال ۱: تخمین احتمال تولید کوئری (۴۰)

در نظر بگیرید که سه داکیومنت زیر (d1,d2,d3) را در Collection داریم. به کمک روش هموارسازی Jelinek-Mercer smoothing احتمال تولید کوئری را برای هر داکیومنت برای کوئری داده‌شده محاسبه کنید و در انتها سندها را بر اساس این احتمالات رتبه‌بندی کنید. λ را برابر ۱/۴ در نظر بگیرید.

d1: Shakespeare was a renowned playwright and wrote poets
d2: the famous playwright William wrote the tragedy of Hamlet
d3: Works of Shakespeare are celebrated worldwide
query: Shakespeare wrote Hamlet

تعداد کل کلمات در کلکسیون: N = ۲۳
تعداد تکرار Shakespeare: ۲
تعداد تکرار wrote: ۲
تعداد تکرار Hamlet: ۱
d_1:
تعداد کل کلمات: ۸
$P(Shakespeare|M_{d_1}) = \frac{1}{8}$
$P(wrote|M_{d_1}) = \frac{1}{8}$
$P(Hamlet|M_{d_1}) = 0$
$P(Shakespeare|M_c) = \frac{2}{23}$
$P(wrote|M_c) = \frac{2}{23}$
$P(Hamlet|M_c) = \frac{1}{23}$
$P(query|d_1) = (\frac{1}{4}\frac{1}{8} + \frac{3}{4}\frac{2}{23})(\frac{1}{4}\frac{1}{8} + \frac{3}{4}\frac{2}{23})(\frac{1}{4}0 + \frac{3}{4}\frac{1}{23})$
d_2:
تعداد کل کلمات: ۹
$P(Shakespeare|M_{d_2}) = 0$
$P(wrote|M_{d_2}) = \frac{1}{9}$
$P(Hamlet|M_{d_2}) = \frac{1}{9}$
$P(query|d_2) = (\frac{1}{4}0 + \frac{3}{4}\frac{2}{23})(\frac{1}{4}\frac{1}{9} + \frac{3}{4}\frac{2}{23})(\frac{1}{4}\frac{1}{9} + \frac{3}{4}\frac{1}{23})$
d_3: تعداد کل کلمات: ۶
$P(Shakespeare|M_{d_3}) = \frac{1}{6}$
$P(wrote|M_{d_3}) = 0$
$P(Hamlet|M_{d_3}) = 0$
$P(query|d_3) = (\frac{1}{4}\frac{1}{6} + \frac{3}{4}\frac{2}{23})(\frac{1}{4}0 + \frac{3}{4}\frac{2}{23})(\frac{1}{4}0 + \frac{3}{4}\frac{1}{23})$
بنابراین با توجه به مقادیر احتمالات محاسبه شده، رتبه‌بندی سندها به شرح زیر خواهد بود:
$d_2 > d_1 > d_3$

محاسبات بیشتر که برای حواب نیاز نیست و همان ۳ عبارت فوق برای جواب درست کافی است.
$P(query|d_1) \approx 0.00030$
$P(query|d_2) \approx 0.00036$
$P(query|d_3) \approx 0.00026$

سوال ۲: naive bayes (۴۰)

با توجه به جدول زیر به سوالات الف ـ د پاسخ دهید.
الف) یک طبقه‌بندی چندجمله‌ای با استفاده از روش naive bayes را روی داده‌های جدول تخمین بزنید.

ب) طبقه‌بندی به دست آمده را روی داده تست اعمال کنید.
ج) یک طبقه‌بندی برنولی با استفاده از روش naive bayes را روی داده‌های جدول تخمین بزنید.
د) طبقه‌بندی به دست آمده را روی داده تست اعمال کنید.

نکته: در جدول زیر داک ۱، داک ۲، داک ۳ و داک ۴ مربوط به داده آموزش است و داک ۵ مربوط به داده تست است.

آیدی داکیومنت	کلمات موجود در داکیومنت	آیا مهم است؟
داک ۱	باران تند	بله
داک ۲	برف تند شد	بله
داک ۳	حیوان آمد	خیر
داک ۴	آمد پلنگ تند	خیر
داک ۵	تند تند آمد	؟

سوال ۳: kNN vs SVM (۱۰)

روش k-Nearest Neighbors (kNN) با روش Support Vector Machines (SVMs) از را از نظر مرزهای تصمیم‌گیری (decision boundaries) مقایسه کنید و تفاوت‌ها را ذکر نمایید.

kNN و SVM دو روش کاملاً متفاوت برای تعیین مرزهای تصمیم در طبقه‌بندی داده‌ها هستند.
kNN یک روش غیرپارامتری است که به طور صریح مرزهای تصمیم را یاد نمی‌گیرد. بلکه، طبقه‌بندی‌ها را بر اساس مقیاس شباهت بین موارد آزمایشی جدید و داده‌های آموزش‌دیده انجام می‌دهد. مرزهای تصمیم به طور ضمنی بر اساس همسایگی‌های محلی اطراف هر نقطه داده پدید می‌آیند. بنابراین، مرزهای تصمیم kNN وابسته به داده و بسیار پیچیده هستند و می‌توانند با شکل داده‌های آموزش تطبیق پیدا کنند.
برعکس، SVMها مرزهای تصمیم صریح را به شکل هایپرپلین‌های حاشیه حداکثر یاد می‌گیرند. SVMها با هدف به‌دست‌آوردن یک جداسازی سراسری و واحد بین کلاس‌ها که حاشیه یا فاصله بین هایپرپلین و نزدیک‌ترین نقاط آموزشی (بردارهای پشتیبان) را به حداکثر می‌رساند، بهینه می‌شوند. بنابراین، مرزهای تصمیم SVM تابع‌های نرم خطی یا غیرخطی صاف و مستقل از نقاط داده فردی هستند.
پس می‌توان گفت که، مرزهای تصمیم kNN ضمنی و وابسته به داده هستند، در حالی که مرزهای تصمیم SVM تابع‌های صاف صریح و مستقل از توزیع داده هستند. SVMها به دنبال مرزهای تصمیم ساده و مقاوم هستند، در حالی که kNN می‌تواند مرزهای بسیار پیچیده را یاد بگیرد.

با توجه به جدول زیر به سوالات الف تا د پاسخ دهید.

الف) یک طبقه‌بندی چندجمله‌ای با استفاده از روش naive bayes را روی داده‌های جدول تخمین بزنید.

ب) طبقه‌بندی به دست آمده را روی داده تست اعمال کنید.

ج) یک طبقه‌بندی برنولی با استفاده از روش naive bayes را روی داده‌های جدول تخمین بزنید.

د) طبقه‌بندی به دست آمده را روی داده تست اعمال کنید.

نکته: در جدول زیر داک ۱، داک ۲، داک ۳ و داک ۴ مربوط به داده آموزش است و داک ۵ مربوط به داده تست است.

آیدی داکیومنت	کلمات موجود در داکیومنت	آیا مهم است؟
داک ۱	(باران تند)	بله
داک ۲	(برق تند شد)	بله
داک ۳	(حیوان آمد)	خیر
داک ۴	آمد پلنگ تند	خیر
داک ۵	تند تند آمد	؟

$P(C) = \frac{1}{2}$

$P(\bar{C}) = \frac{1}{2}$

$P(\text{باران} \mid C) = \frac{1+1}{5+7} = \frac{2}{12}$

$P(\text{تند} \mid C) = \frac{2+1}{5+7} = \frac{3}{12}$

$P(\text{شد} \mid C) = \frac{1+1}{5+7} = \frac{2}{12}$

$P(\text{برق} \mid C) = \frac{1+1}{5+7} = \frac{2}{12}$

$P(\text{جوان} \mid C) = P(\text{پلنگ} \mid C) = \frac{0+1}{5+7} = \frac{1}{12} = P(\text{آمد} \mid C)$

$P(\text{باران} \mid \bar{C}) = \frac{0+1}{5+7} = \frac{1}{12}$

$P(\text{تند} \mid \bar{C}) = \frac{1+1}{5+7} = \frac{2}{12}$

$P(\text{شد} \mid \bar{C}) = \frac{0+1}{5+7} = \frac{1}{12}$

$P(\text{برق} \mid \bar{C}) = \frac{0+1}{5+7} = \frac{1}{12}$

$P(\text{جوان} \mid \bar{C}) = \frac{1+1}{5+7} = \frac{2}{12}$

$P(\text{پلنگ} \mid \bar{C}) = \frac{1+1}{5+7} = \frac{2}{12}$

$P(\text{آمد} \mid \bar{C}) = \frac{2+1}{5+7} = \frac{3}{12}$

$P(C) \times P(\text{باران} \mid C)^2 \times P(\text{تند} \mid C) \times P(\text{شد} \mid C) \times P(\text{برق} \mid C) \times P(\text{جوان} \mid C) \times P(\text{پلنگ} \mid C) \times P(\text{آمد} \mid C) = \frac{9}{12 \times 12 \times 2}$

$P(\bar{C}) \times P(\text{باران} \mid \bar{C})^2 \times P(\text{تند} \mid \bar{C}) \times P(\text{شد} \mid \bar{C}) \times P(\text{برق} \mid \bar{C}) \times P(\text{جوان} \mid \bar{C}) \times P(\text{پلنگ} \mid \bar{C}) \times P(\text{آمد} \mid \bar{C}) = \frac{12}{12 \times 12 \times 2}$

مهم نیست خیر