به نام خدا

درس پردازش زبان طبیعی

تکلیف برنامه‌نویسی: مدل مخفی مارکوف گسسته

دانشگاه تهران
گروه علوم کامپیوتر

هدف از این تمرین پیاده‌سازی مدل مخفی مارکوف با بهره‌گیری از زبان پایتون و آموزش و ارزیابی آن بر روی دنباله‌ای از حروف الفبای انگلیسی است. در این تمرین، مجموعه نمادهای مشاهده، مجموعه حالات، احتمال انتقال حالات، احتمال اولیه حالت‌ها و تابع توزیع مشاهدات در هر حالت به شما داده شده است. همچنین، پنج مجموعه مستقل از دادگان جهت آموزش پنج مدل مخفی مارکوف جداگانه فراهم شده است. پس از آموزش پنج مدل به‌کمک الگوریتم باوم-ولش که مبتنی بر الگوریتم امید بیشینه است و در کلاس مفصل بحث شده است، آن‌ها را روی دو مجموعه از دادگان آزمون به‌کمک الگوریتم ویتربی ارزیابی می‌کنید. شما باید به‌ازای هر نمونه از مجموعه‌های دادگان آزمون، تعیین کنید که کدام مدل احتمال مشاهدات را بیشینه می‌کند.

در پیوست این تمرین، یک فایل زیپ وجود دارد که محتوای آن به شرح زیر است:

model_init.txt

این فایل حاوی احتمال اولیه حالت‌ها، احتمال انتقال حالات و تابع توزیع مشاهدات در هر حالت است. فرمت کلی فایل به صورت زیر است:

```
initial: 6
0.2   0.1   0.2   0.2   0.2   0.1

transition: 6
0.3   0.3   0.1   0.1   0.1   0.1
0.1   0.3   0.3   0.1   0.1   0.1
0.1   0.1   0.3   0.3   0.1   0.1
0.1   0.1   0.1   0.3   0.3   0.1
0.1   0.1   0.1   0.1   0.3   0.3
0.3   0.1   0.1   0.1   0.1   0.3

observation: 6
0.2   0.2   0.1   0.1   0.1   0.1
0.2   0.2   0.2   0.2   0.1   0.1
0.2   0.2   0.2   0.2   0.2   0.2
0.2   0.2   0.2   0.2   0.2   0.2
0.1   0.1   0.2   0.2   0.2   0.2
0.1   0.1   0.1   0.1   0.2   0.2
```

واژه‌های initial، transition و observation بیان می‌کنند که داده‌هایی که در ادامه قرار دارند، بیانگر کدام یک از احتمال اولیه حالات، احتمال انتقال حالات و تابع توزیع مشاهدات هستند. عدد نوشته شده در مقابل initial بیانگر تعداد حالات است و در خط بعد، احتمال حالات که با فاصله از هم جدا شده‌اند، قرار دارد. عدد مقابل transition نیز بیانگر تعداد حالات است و اگر تعداد حالات N باشد، N خط در ادامه خواهیم داشت که عنصر jام از خط iام نمایانگر aij (احتمال انتقال از حالت i به j) است. در نهایت، پس از کلیدواژه observation تعداد نمادهای مشاهده ذکر می‌شود و اگر این مقدار برابر با M باشد، M خط در ادامه خواهیم داشت که عنصر jام از خط iام نمایانگر bij (احتمال مشاهده نماد iام در حالت jام) است.

seq_model_01~05.txt:
این فایل‌ها حاوی داده‌های آموزش برای هر یک از مدل‌ها هستند. هر خط بیانگر که داده آموزشی است. نمادهای مشاهده نیز همگی از حروف بزرگ الفبای انگلیسی هستند. به‌طور مثال، اگر تعداد نمادهای مشاهده ۴ باشد، مجموعه نمادها به صورت $V = \{A, B, C, D\}$ است.

modellist.txt:
در هر خط این فایل، نام فایل‌هایی که پارامترهای مدل‌های ساخته شده در آن‌ها قرار می‌گیرد، آورده شده است.

testing_data1~2.txt:
داده‌های آزمون در این فایل‌ها قرار دارند.

testing_answer.txt:
هر خط از این فایل، نام مدلی است که احتمال مشاهدات را برای داده موجود در خط متناظر در فایل testing_data1.txt بیشینه می‌کند.

ساختار فایل‌های کد
فولدر کدهای شما شامل دو فایل train.py و test.py خواهد بود که دستورات مربوط به آموزش و ارزیابی مدل را دربردارند. همچنین، یک فایل hmm.py خواهید داشت که تمامی کدهای مربوط به پیاده‌سازی مدل مخفی مارکوف (از جمله الگوریتم‌های باوم-ولش و ویتربی جهت آموزش و ارزیابی مدل) در آن قرار دارند.

ورودی
ورودی به صورت آرگومان در هنگام اجرا به کد شما داده می‌شود. برای مثال،
python train.py 100 ../model_init.txt ../seq_model_01.txt ../output/model_01.txt
جهت آموزش مدل اول با ۱۰۰ تکرار و
python test.py ../modellist.txt ../testing_data1.txt ../output/result1.txt
جهت ارزیابی مدل‌ها بر روی مجموعه آزمون اول به‌کار می‌روند.

خروجی
خروجی شما، فایل‌های حاوی پارامترهای مدل با همان فرمت فایل model_init.txt و با اسامی ذکر شده در فایل modellist.txt، و نیز فایل‌های result1~2.txt به ازای هر فایل آزمون است. هر خط از فایل result حاوی نام مدلی که بیشترین احتمال را برای دنباله مشاهدات خط متناظر در فایل testing_data و مقدار آن احتمال است که با فاصله از هم جدا شده اند. به‌طور مثال، model_01.txt 2.019640e-34 یک خط از فایل result می‌تواند باشد.
همچنین، از شما انتظار می‌رود که یک فایل گزارش حاوی توضیحات از مراحل کار و تحلیل نتایج به‌دست آمده نیز ارسال کنید. لطفاً در این فایل، دقت پیش‌بینی برای مجموعه دادگان فایل testing_data1.txt را نیز محاسبه کنید (یادآوری: برچسب‌ها در فایل testing_answer.txt موجود است.)