به نام خدا

دانشکده علوم و فنون نوین

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۲-۱۴۰۱

تمرین شماره ۱

تاریخ تحویل:
۱۴۰۱/۰۸/۲۰

---

۱. (۵٪) **[قانون بیز]** در یک سیستم بازشناسی ارقام گسسته فارسی ۰ تا ۹، پس از آموزش (Training) متوجه شده‌ایم ارقام هفت و هشت دارای سیگنال‌های نسبتاً شبیه به هم هستند به طوری‌که سیگنال آزمون (Test) رقم هفت به احتمال ۵۰٪ به عنوان عدد هفت و به احتمال ۵۰٪ به عنوان عدد هشت توسط سیستم تشخیص داده می‌شود. همچنین، سیگنال آزمون مربوط به رقم هشت، به احتمال ۳۰٪ به عنوان هفت و به احتمال ۷۰٪ به عنوان عدد هشت قابل تشخیص است. یک سیگنال ناشناس وارد سیستم می‌شود، احتمال تشخیص صحیح اعداد هفت و هشت به ترتیب چقدر است؟ فرض کنید سایر ارقام صحیح تشخیص داده می‌شوند و احتمال رخداد سیگنال ورودی برای همه اعداد یکسان باشد.

---

۲. (۵٪) **[یادگیری تقویتی]** همان‌گونه که می‌دانید یادگیری تقویتی (Reinforcement Learning) یکی از انواع روش‌های یادگیری ماشین است که در آن یادگیری به صورت آزمایش و خطا و در تعامل با محیط انجام می‌شود. به بررسی نمونه کاربردهای این روش در حوزه پردازش زبان و زبانشناسی پرداخته و نتیجه این بررسی را با بیان نمونه کاربردها و نحوه استفاده از یادگیری تقویتی در آن کاربرد، به همراه ذکر منابع ارائه کنید.

---

۳. (۱۰٪) **[احتمال]** فرض کنید تابع توزیع احتمال (pdf) متغیر تصادفی X به صورت زیر باشد:

$$f(x) = \begin{cases} a(1-x^2) & -1 < x < 1 \\ 0 & else \end{cases}$$

الف) مقدار a را بدست آورید و از آن در دو بخش زیر استفاده کنید.

ب) تابع توزیع تجمعی (CDF) متغیر تصادفی X را بدست آورید.

ج) امید ریاضی را برای متغیر تصادفی X محاسبه کنید.

---

۴. (۵٪) **[نظریه اطلاعات]** یکی از حالت‌های اطلاعات متقابل با نام اطلاعات متقابل نقطه‌ای (PMI: Pointwise Mutual Information) شناخته شده است که در پردازش متن کاربردهای زیادی دارد. این روش را بررسی کرده و نحوه استفاده از آن برای دو کاربرد در پردازش متن تشریح کنید.

h.veisi@ut.ac.ir

به نام خدا

دانشکده علوم و فنون نوین

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)
نیم‌سال اول ۱۴۰۲-۱۴۰۱

تمرین شماره ۱

تاریخ تحویل:
۱۴۰۱/۰۸/۲۰

---

۵. (۱۰٪) [**تخمین**] تخمین بیشینه شباهت (MLE) را برای پارامتر λ با فرض داشتن تعداد N نمونه از داده‌هایی با تابع توزیع $p(x) = e^{-\lambda}\frac{\lambda^x}{x!}$ حساب کنید.

---

۶. (۱۵٪) [**پیاده‌سازی- هم‌بستگی**] هدف این تمرین استفاده از ضریب هم‌بستگی خطی برای محاسبه شباهت بین چند متغیر تصادفی است. برای این کار، ابتدا اعداد ۰ تا ۹ فارسی را در یک نرم‌افزار مانند Adobe Audition با نرخ نمونه‌برداری 16KHz و تعداد بیت ۱۶ به صورت مونو (تک کاناله) با فرمت Wav ضبط کنید. طول همه سیگنال‌ها را برابر کنید، برای این کار سکوت ابتدا و انتهای سیگنال‌ها را کم یا زیاد کنید. سپس، یک برنامه بنویسید که پوشه حاوی این ۱۰ فایل را گرفته و میزان هم‌بستگی خطی (پیرسون) بین هر جفت از آنها را محاسبه کرده و خروجی را در یک ماتریس نمایش بدهد. چه تحلیلی از نتایج بدست آمده دارید؟

---

۷. (۵۰٪) [**پیاده‌سازی- انتروپی و توزیع حروف فارسی**] یک برنامه رایانه‌ای بنویسید که احتمال حروف فارسی را محاسبه کرده و از روی آن مقدار انتروپی را بدست آورید. برای این کار، به همراه این تمرین ۷۰ فایل با فرمت txt ارائه شده که هر کدام حاوی یکی از متون خبری خبرگزاری ایسنا می‌باشد. فایل‌ها در ۷ عنوان خبری (هر کدام ۱۰ فایل) که هرکدام در یک پوشه قرار دارند، دسته‌بندی شده‌اند. از این به بعد، به کل این متن (۷۰ فایل) پیکره متنی تمرین‌های درس زبان‌شناسی رایانشی (**زبرا**) می‌گوییم. در این تمرین، احتمال‌های حروف را با شمارش آنها در مجموعه زبرا بدست آورید.

الف) با فرض استفاده از مقدار یکسان احتمال برای همه حروف (عدم محاسبه احتمال از روی پیکره متنی)، انتروپی را محاسبه کنید. در این حالت فقط از حروف استاندارد فارسی، با در نظر گرفتن حرف "فاصله" و بدون آن، استفاده کنید و از کاراکترهای خاص و علائم سجاوندی صرف‌نظر کنید.

ب) انتروپی را برای همه حروف از جمله کاراکترهای خاص و علائم سجاوندی موجود در متن پیکره، با احتمال‌های محاسبه شده از روی پیکره بدست آورید.

ج) قسمت ب را فقط برای حروف استاندارد فارسی، با در نظر گرفتن حرف "فاصله" و بدون آن،

h.veisi@ut.ac.ir

بدست آورید. تحلیل خود را از مقایسه نتایج حاصل شده در این بخش و بخش الف بیان کنید.

د) متوسط طول کلمات فارسی در پیکره را محاسبه کنید و مقدار آن را گزارش کنید. با استفاده از مقدار حاصل و نتیجه قسمت ج، انتروپی را برای کلمات فارسی هم بدست آورید.

ه) هیستوگرام نرمال شده حروف فارسی را رسم کنید. در این نمودار، محور x بیانگر حروف باشد و محور y بیانگر احتمال آن حرف باشد. هیستوگرام به صورت مرتب شده از چپ به راست برای حروف با احتمال بزرگ به کوچک رسم کنید.

و) یک هیستوگرام برای طول کلمات (بر حسب تعداد حروف) رسم کنید. محور x بیانگر تعداد حروف باشد (مقادیر ۱، ۲، ۳، ...) و محور y بیانگر تعداد کلمات (نرمال شده برای تبدیل به احتمال) زبان فارسی با آن تعداد حرف باشد. با فرض گاوسی بودن توزیع طول کلمات، یک توزیع گاوسی به این هیستوگرام متناسب کنید. نمودار توزیع برازش شده را روی هیستوگرام رسم کنید و مقدار بدست آمده برای پارامترهای توزیع را بنویسید.

h.veisi@ut.ac.ir