به نام خدا

دانشکده سامانه‌های هوشمند

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۴-۱۴۰۳

تمرین شماره ۱

تاریخ تحویل:
۱۴۰۳/۰۷/۳۰

---

۱. (۱۰٪) **[پژوهش – مراحل آموزش یک مدل زبانی بزرگ]** در درس با چندین روش آموزش مختلف شامل یادگیری با نظارت، بدون نظارت، نیمه نظارتی و یادگیری تقویتی آشنا شدید. در این بخش قصد داریم استفاده‌ی بعضی از این روش‌ها را در مدل‌های زبانی بزرگ بررسی کنیم.

I. روش‌های یادگیری ذکر شده را در بسیار کوتاه توضیح دهید و مقایسه کنید.

II. مفهوم Self-supervised learning را توضیح دهید و به اهمیت آن در آموزش مدل‌های زبانی بزرگ بپردازید. مثال‌هایی از کاربردهای آن در این زمینه ارائه دهید.

III. یکی از مراحل آموزش مدل‌های زبانی بزرگ، Reinforcement Learning from Human Feedback (RLHF) است. این مرحله را به زبان ساده توضیح دهید و به مزایا و معایب آن اشاره کنید. در چه مواردی استفاده از این روش ترجیح داده می‌شود؟ همچنین تأثیر بازخورد انسانی بر بهبود مدل را مورد بحث قرار دهید.

---

۲. (۱۵٪) **[نظری – مفاهیم یادگیری ماشین]** هدف این بخش مروری مفاهیم ابتدایی یادگیری ماشین می‌باشد.

آ) فرض کنید یک مساله‌ی پنج کلاسه داریم که مدلی را با استفاده از یک الگوریتم Supervised Learning را برای آن آموزش داده‌ایم. آیا می‌توان بدون آموزش دوباره هیچ پارامتری، این مدل را برای مساله‌ی ۷ کلاسه نیز به کار برد؟ توضیح دهید که چه شرایطی نیاز است تا مدل عملکرد مناسبی داشته باشد در صورت نیاز شکلی از نحوه‌ی جدا سازی رسم کنید. (راهنمایی: به مدل‌های clustering فکر کنید).

ب) در درس با معیارهای اندازه‌گیری عملکرد مدل‌ها مانند دقت و صحت آشنا شدید،

I. آیا این معیارها همیشه بهترین روش برای اندازه‌گیری عملکرد مدل‌های طبقه‌بندی هستند؟ مثالی برای رد یا قبول این موضوع ارائه بدهید.

II. تعدادی از معیارهایی که برای بررسی مدل‌های مولد زبانی استفاده می‌شوند شامل موارد زیر می‌باشند. به هر یک از این معیارها به‌صورت جداگانه پرداخته و توضیح دهید که چگونه کار می‌کنند و چه نقاط قوت و ضعفی دارند.

أ) BLEU (Bilingual Evaluation Understudy)

ب) ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ت) METEOR (Metric for Evaluation of Translation with Explicit Ordering)

ث) Perplexity

ج) انسان‌محور (Human Evaluation)

---

h.veisi@ut.ac.ir

به نام خدا

دانشکده سامانه‌های هوشمند

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۴-۱۴۰۳

تمرین شماره ۱

تاریخ تحویل:
۱۴۰۳/۰۷/۳۰

---

۳. (۱۵٪) **[نظری- مروری بر آمار و احتمال]** هدف این بخش مروری بر آمار و احتمال می‌باشد.

الف) پنج مهمان در یک مهمانی کت‌ها و کلاه‌های خود را روی تخت می‌اندازند. بعد از مهمانی، هر کدام به صورت تصادفی یک کت و به صورت تصادفی و مستقل یک کلاه انتخاب می‌کنند. احتمال اینکه همه آن‌ها کت و کلاه صحیح خود را دریافت کنند چقدر است؟

ب) متغیر تصادفی پیوسته X دارای تابع توزیع تجمعی زیر است:

$$F_x(x) = \begin{cases} 1 - \frac{1}{x^3} & for\ x \geq 1 \\ 0 & OW \end{cases}$$

۱. احتمال $P\ (-1 < X < 2)$ را محاسبه کنید.
۲. تابع چگالی احتمال $F_x(x)$ را پیدا کنید.

پ) فرض کنید یک مدل یادگیری ماشین برای شناسایی احساسات در متون، ۷۰٪ مواقع احساس مثبت و ۳۰٪ مواقع احساس منفی را پیش‌بینی می‌کند. فرض کنید این مدل ۵٪ از متن‌های مثبت را به درستی شناسایی می‌کند و ۲٪ از متن‌های منفی را نیز به درستی شناسایی می‌کند. با توجه به اینکه مدل یک متن را به عنوان مثبت شناسایی کرده است، احتمال اینکه این متن در واقع مثبت باشد چقدر است؟

ت) متغیر تصادفی X یک متغیر تصادفی پیوسته است که تابع چگالی احتمال (PDF) آن به صورت زیر داده شده است:

$$f_x(\mathrm{x}) = \frac{1}{2}\mathrm{e}^{-|\mathrm{x}|}$$

اگر $\mathrm{Y} = \mathrm{X}^2$ توزیع تجمعی Y (CDF) را پیدا کنید.

ت) در پردازش زبان طبیعی، کوواریانس اغلب برای اندازه گیری رابطه بین ویژگی های مختلف استخراج شده از داده‌های متنی استفاده می شود. با توجه به دو بردار ویژگی X و Y، که فرکانس دو کلمه مختلف را در چندین سند نشان می دهد، توضیح دهید که یک کوواریانس مثبت، منفی و صفر بین X و Y به چه معنی خواهد بود.

---

۴. (۱۵٪) **[نظری- محاسبه MLE و MAP]**

الف) توزیع پارتو به صورت زیر تعریف می‌شود:

$$f(x|x_0, \theta) = \theta x_0^{\theta} x^{-\theta-1}, \quad x \geq x_0,\ \theta > 1$$

h.veisi@ut.ac.ir

به نام خدا

دانشکده سامانه‌های هوشمند

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۴-۱۴۰۳

تمرین شماره ۱

تاریخ تحویل:
۱۴۰۳/۰۷/۳۰

فرض کنید $x_0 > 0$ و $X_1, X_2, \dots, X_n$ نمونه‌های i.i.d باشند. MLE برای θ محاسبه کنید.

ب) توزیع پواسون به صورت زیر تعریف می‌شود:

$$P(X|\lambda) = \frac{\lambda^X e^{-\lambda}}{X!}$$

این توزیع دارای $E(X) = \lambda$ می‌باشد.

حال فرض کنید n نقطه i.i.d از توزیع پواسون $P(\lambda)$ به صورت $D = \{X_1, \dots, X_n\}$ داریم.

I. نشان دهید که میانگین نمونه‌ای

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

برآوردگر بدون اریب بیشینه درست‌نمایی λ (MLE) است.

II. حال فرض کنید به روش بیزی عمل کنیم و یک توزیع پیشین برای λ در نظر بگیریم. با فرض اینکه λ دارای توزیع گاما با پارامترهای (α, β) باشد، تابع چگالی احتمال آن به صورت زیر است:

$$p(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

که در آن

$$\Gamma(\alpha) = (\alpha - 1)!$$

است (در اینجا فرض می‌کنیم که α یک عدد صحیح مثبت است). توزیع پسین λ را محاسبه کنید.

III. یک بیان تحلیلی برای برآورد پسین بیشینه λ (MAP) تحت توزیع پیشین $\text{Gamma}(\alpha, \beta)$ دست آورید.

---

۵. (۵٪) **[نظری- محاسبه آنتروپی]** هدف از این بخش یادگیری مفهوم آنتروپی می‌باشد.

الف) اثبات کنید اطلاعات به‌دست‌آمده از مشاهده ترکیب N رویداد مستقل، که احتمالات آن‌ها به‌صورت p_i برای $i = 1, \dots, N$ تعریف شده است، با با مجموع اطلاعات به‌دست‌آمده از مشاهده هر یک از این رویدادها به‌صورت جداگانه برابر است.

ب) برای هر بخش زیر مقدار آنتروپی را محاسبه کنید:

۱. مقادیر پیکسل در یک تصویر که مقادیر خاکستری ممکن آن‌ها تمامی اعداد صحیح از ۰ تا ۲۵۵ با احتمال یکنواخت هستند.
۲. انسان‌ها بر اساس اینکه آیا پستاندار هستند یا نیستند، طبقه‌بندی شده‌اند.
۳. دسته‌بندی کلمات در یک مجموعه متنی که برخی از کلمات با احتمال‌های $\frac{1}{4}$، $\frac{1}{4}$ و $\frac{1}{2}$ رخ می‌دهد.

h.veisi@ut.ac.ir

۴. جمعیتی از افراد که بر اساس اینکه آیا سن آن‌ها بیشتر از سن میانه جمعیت است یا خیر، طبقه‌بندی شده‌اند.

---

۶. (۵٪) **[پیاده‌سازی- مروری بر پایتون]** اگر با پایتون آشنایی ندارید، با حل این سوال تمام مفاهیمی که نیاز به دانستن دارید خواهید و فهمید و فقط لازم است هر مورد جدیدی را در اینترنت جست‌وجو کنید. با استفاده از پایتون، یک برنامه بنویسید که شامل یک کلاس به نام Student باشد. این کلاس باید شامل موارد زیر باشد:

- یک تابع سازنده که نام و سن دانش‌آموز را به عنوان پارامتر دریافت کند.
- یک تابع برای نمایش اطلاعات دانش‌آموز.
- یک تابعی که سن دانش‌آموز را به‌روز کند.

سپس یک لیست از دانش‌آموزان ایجاد کنید و با استفاده از حلقه for اطلاعات هر دانش‌آموز را نمایش دهید.

همچنین یک دیکشنری ایجاد کنید که نام دانش‌آموزان به عنوان کلید و سن آن‌ها به عنوان مقدار باشد. برنامه شما باید بررسی کند که آیا سن هر دانش‌آموز بیشتر از ۱۸ است یا نه و نتیجه را با استفاده از یک شرط چاپ کند.

---

۷. (۳۵٪) **[پیاده‌سازی- پیش‌پردازش و محاسبه آنتروپی]**

پیش‌پردازش یکی از مراحل کلیدی در پردازش زبان طبیعی است که به هدف آماده‌سازی داده‌ها برای تجزیه و تحلیل و یادگیری ماشین انجام می‌شود. مراحل متداول شامل موارد زیر است:

۱. حذف نویز: شامل حذف کاراکترهای غیرفضای سفید (مانند تب‌ها، خط‌های جدید و کاراکترهای غیر ضروری) که می‌تواند به بهبود کیفیت داده‌ها کمک کند.

۲. حذف عبارات خاص: مانند "[removed]" که برای جلوگیری از اختلال در تحلیل متن ضروری است.

۳. کدگذاری و جایگزینی: حذف یا جایگزینی کدهای کاراکتر به معادل‌های ASCII آنها، که می‌تواند در تجزیه و تحلیل متون مختلف مفید باشد.

۴. واحدسازی (توکنایزیشن): تبدیل متن به توکن‌های فردی، که به عنوان ورودی برای مدل‌های یادگیری ماشین استفاده می‌شود.

در این پرسش قصد داریم شما را با این فرایند آشنا کنیم.

h.veisi@ut.ac.ir

به نام خدا

دانشکده سامانه‌های هوشمند

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۴-۱۴۰۳

تمرین شماره ۱

تاریخ تحویل:
۱۴۰۳/۰۷/۳۰

مجموعه داده Gutenberg: این مجموعه داده شامل متونی از نویسندگانی چون جین آستن و ویلیام شکسپیر است. اطلاعات بیشتر درباره پروژه Gutenberg در این لینک موجود است.

از این مجموعه داده یک کتاب انتخاب کنید و بخش‌های زیر را انجام دهید. مراحلی که نیاز به انجام دارید شامل موارد زیر می‌باشد:

I. پیش‌پردازش:

۱- همه کاراکترهای غیر Space از جمله new lineها، tab و غیره را با space جایگزین کنید.

۲- عبارات "[deleted]" یا "[removed]" را حذف کنید.

۳- کدهای کاراکتر را با معادل‌های ASCII آن‌ها جایگزین کنید.

۴- همه URLها یعنی توکن‌هایی که با http یا www شروع می‌شوند را حذف کنید.

۵- در صورت به وجود آمدن فضاهای تکراری بین توکن‌ها آن‌ها را حذف کنید.

- اکنون هر توکن با یک فضای واحد (یک space) جدا می‌شود.

II. مراحل زیر را با استفاده از کتابخانه اعمال کنید:

- ریشه‌یابی (lemmatization): توکن‌ها را با ریشه‌ی آن‌ها جایگزین کنید؛ مثلاً words به word تبدیل شود.
- تقسیم‌بندی جملات: بین هر جمله یک خط جدید (New Line) اضافه کنید. برای این، از spaCy's sentencizer برای تقسیم جملات استفاده خواهیم کرد.

III. محاسبه آنتروپی:

- با فرض برابری احتمال برای تمامی حروف (بدون محاسبه احتمال بر اساس پیکره متنی)، آنتروپی را محاسبه نمایید. در اینجا فقط از حروف استاندارد انگلیسی استفاده کنید،
- متوسط طول کلمات انگلیسی در پیکره را محاسبه و گزارش کنید. با استفاده از این مقدار، آنتروپی را برای کلمات انگلیسی نیز به‌دست آورید.

حال مراحل بالا را برای چند کتاب دیگر انجام دهید و نتایج را مقایسه کنید. (در صورت طولانی بودن فرایندها، متن کتاب‌ها را محدود کنید برای مثلا ۱۰۰۰۰ کاراکتر نگه دارید.)

h.veisi@ut.ac.ir