به نام خدا

دانشکده سامانه‌های هوشمند

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)
نیم‌سال اول ۱۴۰۵-۱۴۰۴

تمرین شماره ۱

تاریخ تحویل:
۱۴۰۴/۰۷/۳۰

---

۱. (۱۰٪) [پژوهش – مفاهیم آموزش در یادگیری ماشین]

الف) انواع یادگیری را خیلی کوتاه تعریف و مقایسه کنید: (بانظارت، بی‌نظارت، نیمه‌نظارتی و تقویتی.) برای هرکدام یک مزیت و یک محدودیت ذکر کنید.

ب) یادگیریِ خودنظارتی (Self-supervised learning) را تعریف کنید و بگویید چرا برای پیش‌آموزش یک مدل زبانی بزرگ (LLM) حیاتی است. مثال‌هایی از کاربردهای آن در این زمینه ارائه دهید.

پ) یکی از روش‌های آموزش LLMها، روش SFT (Instruction Fine-Tuning/Supervised Fine-Tuning) است، این روش را توضح دهید. چرا برای تبدیل یک «مدل زبانِ عمومی» به «مدلِ دستورپذیر و کاربردی» مفید است؛ و اگر انجام نشود چه پیامدهایی دارد؟

ت) 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. (۱۰٪) [نظری – مفاهیم یادگیری ماشین]

الف) 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 برخی معیارهای اندازه‌گیری عملکرد مدل‌ها مانند دقت (Precision) و صحت (Accuracy) آشنا شدید، تعدادی از معیارهای دیگر 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.

۱. BLEU (Bilingual Evaluation Understudy)
۲. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
۳. METEOR (Metric for Evaluation of Translation with Explicit Ordering)
۴. ارزیابی انسان‌محور (Human Evaluation)

پ) تعریف مفاهیم بایاس–واریانس و تعمیم‌پذیری.

ت) نقشِ Cross-Validation را در انتخاب ظرفیتِ مناسب و پایشِ تعمیم‌پذیری توضیح دهید.

---

h.veisi@ut.ac.ir

به نام خدا

دانشکده سامانه‌های هوشمند

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)
نیم‌سال اول ۱۴۰۵-۱۴۰۴

تمرین شماره ۱

تاریخ تحویل:
۱۴۰۴/۰۷/۳۰

---

۳. (۱۰٪) [نظری- مروری بر آمار و احتمال]

الف) پنج مقاله داریم که برای هرکدام «عنوان» و «چکیده» جداگانه‌اش موجود است. عناوین به‌صورت تصادفی بین پنج مقاله پخش می‌شود و مستقل از آن چکیده‌ها هم تصادفی پخش می‌شوند.

۱ احتمال این‌که دقیقاً k مقاله هم عنوان درست خود را بگیرند و هم چکیده درست خود را، چقدر است؟

۲ احتمال این‌که هیچ مقاله‌ای نه عنوان درست و نه چکیده درست را بگیرد، چقدر است؟

ب) 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؟

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

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{x^2}{2\sigma^2}}$$

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

ت) در پردازش زبان طبیعی، کوواریانس اغلب برای اندازه گیری رابطه بین ویژگی‌های 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 می‌دهد، 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:

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

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.

۱ نشان دهید که میانگین نمونه‌ای برآوردگر بدون اریب بیشینه درست‌نمایی (MLE) برای λ برابر است با

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

h.veisi@ut.ac.ir

به نام خدا

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)
نیم‌سال اول ۱۴۰۵-۱۴۰۴

تمرین شماره ۱

تاریخ تحویل:
۱۴۰۴/۰۷/۳۰

۲ 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.

---

۵. (۱۰٪) [نظری- محاسبه آنتروپی]

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 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.
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]

در کتابخانهٔ دانشکده یک رباتِ کنجکاو داریم که پیام‌های کوتاه را می‌خواند. شما باید با چند خط پایتون به او کمک کنید بفهمد کدام پیام‌ها به درس «روش های یادگیری ماشین در پردازش زبان طبیعی» مربوطند یا نه. وظیفه ما این هست که یک کلاس به نام TinyText بسازید که:

- سازنده آن text و یک label (پیش‌فرض: «نامشخص») را بگیرد و ذخیره کند.
- متد show_info داشته باشد که تعداد واژه‌ها را حساب کند و ۱۰ حرفِ اول متن را چاپ کند.
- متد set_label داشته باشد که برچسب را به‌روزرسانی کند.

یک تابع به نام is_related(text) بنویسید که اگر متن شامل واژه «ماشین» یا «متن» بود مقدار True برگرداند، وگرنه False.
یک لیست از ۳ تا ۵ جمله کوتاه بسازید (بعضی مرتبط با درس، بعضی نامرتبط). سپس از روی هر جمله یک شیء از کلاس TinyText بسازید و آن‌ها را در یک لیست به نام samples بریزید.
با یک حلقه for روی samples حرکت کنید و برای هر نمونه:

- show_info را صدا بزنید.

h.veisi@ut.ac.ir

- با استفاده از شرط و تابع is_related چاپ کنید «مرتبط با درس» یا «نامرتبط».
- یک شمارنده نگه دارید که تعداد پیام‌های «مرتبط» را بشمارد.
- اگر یکی از جمله‌ها شامل واژه‌هایی مثل «تبلیغ» یا «تخفیف» بود، برچسب همان نمونه را با set_label به «هرزنامه» تغییر دهید و دوباره show_info آن را چاپ کنید تا به‌روزرسانی دیده شود.

در پایان، تعداد کل پیام‌های مرتبط را چاپ کنید.

---

۷. (۴۰٪) [پیاده‌سازی- پیش‌پردازش و محاسبه آنتروپی]

هدف از این پرسش محاسبه احتمال حروف فارسی و مقدار آنتروپی آن‌ها است. برای این کار، از پیکره اشعار فارسی که در لینک زیر در دسترس است، استفاده کنید:

https://github.com/amnghd/Persian_poems_corpus

الف) با فرض استفاده از مقدار یکسان احتمال برای همه حروف (عدم محاسبه احتمال از روی پیکره متنی)، آنتروپی 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) آنتروپی 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، 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**، آنتروپی 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