بسمه‌تعالی

هوش‌مصنوعی قابل اعتماد **نیمسال دوم ۱۴۰۳-۱۴۰۲**

امتحان میان ترم **وقت آزمون: ۱۳۰ دقیقه**

توجه: استفاده از کتاب، جزوه، ساعت هوشمند، لپتاپ و هر گونه وسایل الکترونیکی در حین امتحان غیرمجاز است.

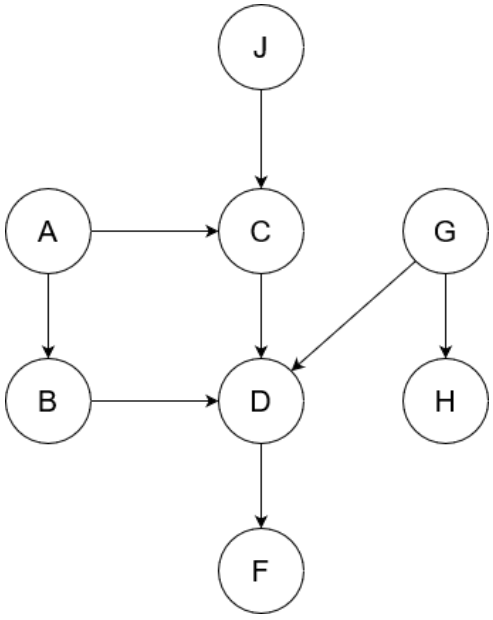
توجه: نمره امتحان از ۱۱۰ است و برای کامل شدن باید ۱۰۰ نمره کسب نمایید.

سوال ۱ سوالات پاسخ کوتاه (۲۴ نمره)

در هر یک از موارد زیر درست یا نادرست بودن آن را مشخص کنید و به صورت مختصر علت را توضیح دهید. (هر مورد ۴ نمره)

الف) افزایش robustness همواره با کاهش دقت مدل همراه است.

ب) فرض کنید $f(x)$ یک طبقه‌بند L-Lipschitz باشد. اگر یک پیش‌پردازش به قبل این طبقه‌بند اضافه کنیم به طوری که روی ورودی x یکسری پیش‌پردازش‌ها انجام دهیم و سپس آن را به طبقه‌بند دهیم، می‌توان مطمئن بود که همچنان این طبقه‌بند جدید Lipschitz است.



ج) در گراف بالا، $B \perp J \mid F$

د) در گراف بالا، $A \perp J$

ه) در گراف بالا، $P(F \mid do\ A = a) = P(F \mid A = a)$

و) Confidence که یکی از الزامات مورد نیاز برای اعتماد به مدل‌های زبانی است بدین معنا که هر ادعایی، آن مدل مطرح می‌کند باید بتوان به یک منبع دانش معتبر نسبت دهد.

سوال ۲ Certified Robustness (۱۵ نمره)

فرض کنید که یک Soft Classifier به اسم $F(x)$ داریم که L-Lipschitz است. ثابت کنید که اگر $\|x - \hat{x}\| \leq \frac{1}{2L}(P_a - P_b)$ باشد که در آن P_a و P_b به ترتیب احتمال اولین و دومین برچسب محتمل برای نمونه‌ی x هستند، آنگاه محتمل‌ترین برچسب برای x و $\hat{x}$ یکسان خواهد بود.

سوال ۳ (۱۰ نمره)

فرض کنید یک مدل black-box به شما داده شده است و شما می‌خواهید به ازای نمونه‌های داده شده، نمونه‌های خلاف واقع[1] آن‌ها را بدست آورید. با توجه به black-box بودن مدل، چه پیشنهادی برای انجام این کار دارید؟

سوال ۴ (۱۰ نمره)

تصور کنید یک مدل زبانی داریم که وظیفه ترجمه از زبان انگلیسی به زبان فارسی را برعهده دارد. حال می‌خواهیم ببینیم آیا در لایه‌های نهان این مدل اطلاعاتی در مورد طول جمله ورودی (کوتاه-بلند) انکود می‌شود یا خیر. روشی ارائه دهید که مشخص کند آیا این مدل زبانی، طول جمله ورودی را برای ما انکود می‌کند یا خیر؟

سوال ۵ (۱۶ نمره)

در مورد مدل β-VAE برای ایجاد بازنمایی از هم گسیخته[2] موارد زیر را به صورت کامل توضیح دهید. (هر مورد ۴ نمره)

الف) چرا پارامتر β در تابع خطای این مدل را نمی‌توان خیلی بزرگ در نظر گرفت؟

ب) چرا پارامتر β در تابع خطای این مدل را نمی‌توان خیلی کوچک (مثلا نزدیک ۱) در نظر گرفت؟

ج) اگر پارامتر β این مدل را مقداری کوچکتر از ۱ در نظر بگیریم چه مشکلی برای این مدل بوجود می آید؟

د) اگر ویژگی‌های یک مجموعه داده ذاتا مستقل از یکدیگر نباشند، آیا مدل β-VAE قادر به یافتن بازنمایی از هم گسیخته خواهد بود؟ چرا؟

---

[1] Counterfactual Example

[2] Disentangled representation

سوال ۶ روش‌های تفسیرپذیری (۱۸ نمره)

جدول زیر را با کلمات yes و no تکمیل کنید. (هر مورد ۱ نمره)

حداکثر در ۲ مورد می‌توانید یک توضیح مختصر برای جواب خود بنویسید. توضیح در موارد بیشتر، نمره منفی خواهد داشت.

	Explanation Method	Local	Black-box	Model agnostic
1	PDP			
2	Shapley value			
3	Counterfactual Explanation			
4	Grad-CAM			
5	LIME			
6	Feature visualization in CNNs			

سوال ۷ (۱۷ نمره)

الف) (۸ نمره) در رابطه با مقاله *GAT*، وجود فضای نهان از هم گسیخته در مدل مولد چه امکانی را برای ما فراهم می‌کند؟ نویسندگان مقاله از این امکان چگونه بهره بردند؟

ب) (۹ نمره) نحوه عملکرد روش Defense-GAN را شرح دهید.

موفق باشید.