تکلیف دوم درس شبکه های عصبی

وحید ملکی، پوریا دادستان

۱۴ آذر ۱۴۰۴

سوال ۱

توضیح دهید چرا شبکه‌های عصبی اغلب به عنوان "جعبه سیاه" شناخته می‌شوند و این مسئله چه چالش‌هایی برای قابلیت تفسیر و شفافیت مدل ایجاد می‌کند و راه‌های رفع این مشکل چیست؟

۰.۱ چرا شبکه‌های عصبی "جعبه سیاه" (Black Box) هستند؟

اصطلاح "جعبه سیاه" در هوش مصنوعی به مدل‌هایی اشاره دارد که فرآیند تصمیم‌گیری درونی آن‌ها برای انسان به سادگی قابل درک نیست. دلایل اصلی این ویژگی در شبکه‌های عصبی عبارتند از:

- **پیچیدگی ساختاری و ابعاد بالا:** شبکه‌های عصبی عمیق (Deep Neural Networks) اغلب شامل میلیون‌ها یا حتی میلیاردها پارامتر (وزن‌ها و بایاس‌ها) هستند. ردیابی اینکه چگونه یک ورودی خاص از طریق این اتصالات متعدد به یک خروجی منجر می‌شود، عملاً برای انسان غیرممکن است.
- **غیرخطی بودن:** استفاده از توابع فعال‌ساز غیرخطی (مانند ReLU، Sigmoid یا Tanh) در لایه‌های پنهان باعث می‌شود که رابطه بین ورودی و خروجی یک نگاشت ریاضی بسیار پیچیده باشد که نمی‌توان آن را با یک رابطه خطی ساده توضیح داد.
- **نمایش ویژگی‌های انتزاعی:** لایه‌های میانی شبکه، ویژگی‌هایی را استخراج می‌کنند که اغلب برای انسان معنای بصری یا منطقی مستقیمی ندارند (مثلاً ترکیبی از پیکسل‌ها که نه خط است و نه دایره، اما برای شبکه معنی‌دار است).

۰.۲ چالش‌های ناشی از عدم تفسیرپذیری

عدم شفافیت در عملکرد شبکه‌های عصبی چالش‌های جدی‌ای را به وجود می‌آورد، از جمله:

- **عدم اعتماد (Lack of Trust):** در حوزه‌های حساس مانند پزشکی (تشخیص سرطان) یا مالی (تایید وام)، کاربران و متخصصان نیاز دارند بدانند چرا مدل یک تصمیم خاص را گرفته است. بدون توضیح منطقی، اعتماد به سیستم دشوار است.
- **عیب‌یابی دشوار (Debugging Difficulty):** وقتی یک شبکه عصبی اشتباه می‌کند، تشخیص اینکه کدام لایه یا کدام نورون باعث خطا شده است بسیار دشوار است، که فرآیند بهبود مدل را کند می‌کند.

- **سوگیری و تبعیض (Bias and Fairness):** مدل‌ها ممکن است سوگیری‌های موجود در داده‌های آموزشی را یاد بگیرند (مثلاً تبعیض نژادی یا جنسیتی). در یک جعبه سیاه، کشف این سوگیری‌ها قبل از وقوع فاجعه سخت است.
- **مسائل قانونی و مقرراتی:** قوانینی مانند GDPR در اروپا بر "حق توضیح" (Right to Explanation) تاکید دارند. استفاده از مدل‌های جعبه سیاه در تصمیم‌گیری‌هایی که بر زندگی افراد تاثیر می‌گذارد، ممکن است با موانع قانونی روبرو شود.

۳. راه‌های رفع مشکل: هوش مصنوعی توضیح‌پذیر (XAI)

برای مقابله با این چالش‌ها، حوزه "هوش مصنوعی توضیح‌پذیر" (Explainable AI) راهکارهایی را ارائه می‌دهد:

- **روش‌های مستقل از مدل (Model-Agnostic Methods):** تکنیک‌هایی مانند LIME و SHAP سعی می‌کنند رفتار مدل پیچیده را با تقریب زدن آن به یک مدل ساده‌تر (مثل رگرسیون خطی) در همسایگی یک نمونه خاص توضیح دهند. این روش‌ها نشان می‌دهند کدام ویژگی‌های ورودی بیشترین تاثیر را بر تصمیم نهایی داشته‌اند.
- **بصری‌سازی (Visualization):** در شبکه‌های کانولوشنی (CNN)، می‌توان از روش‌هایی مانند *Saliency Maps* یا *Grad-CAM* استفاده کرد تا مشخص شود شبکه به کجای تصویر توجه کرده است (Heatmap).
- **مکانیزم توجه (Attention Mechanisms):** در مدل‌های پردازش زبان (مانند ترنسفورمرها)، مکانیزم توجه نشان می‌دهد که مدل هنگام تولید خروجی، روی کدام کلمات ورودی تمرکز کرده است که خود نوعی تفسیرپذیری ذاتی ایجاد می‌کند.
- **استفاده از مدل‌های تفسیرپذیر:** در مواردی که شفافیت اولویت بالاتری نسبت به دقت نهایی دارد، می‌توان به جای شبکه‌های عصبی عمیق از مدل‌های ساده‌تر مانند درخت تصمیم (Decision Trees) یا رگرسیون لجستیک استفاده کرد.