گزارش سخنرانی «یادگیری ماشینی» دکتر طبیبیان در آکادمی دانایان

آکادمی دانایان (گروه مالی دانایان) روز سه‌شنبه مورخ ۱۴۰۲/۰۵/۳۱ از ساعت ۱۷:۳۰ الی ۲۰ میزبان نشستی با موضوع «یادگیری ماشینی (Machine Learning) و برخی کاربردهای آن در اقتصاد» بود که سخنران آن جناب آقای دکتر محمد طبیبیان، استاد محترم اقتصاد کشور، بودند. نشست به دو صورت حضوری و آنلاین برگزار شد. افرادی که علاقه‌مند به مشاهده‌ی فیلم این نشست هستند، از هفته‌ی آینده می‌توانند برای استفاده از فایل سخنرانی به سایت آکادمی مراجعه فرمایند.

دکتر طبیبیان در تعریف یادگیری ماشینی (ML) آن را راهی برای آموزش کامپیوترها دانست که چگونه از داده‌ها خود ظرفیت تصمیم‌گیری یا استنتاج بیابد. یادگیری ماشینی زیرمجموعه‌ی هوش مصنوعی (AI)است که با توسعه‌ی الگوریتم‌ها و مدل‌ها، کامپیوتر از داده‌ها یادگرفته و پیش‌بینی می‌کند یا تصمیم می‌گیرد. این شیوه مانند سیستم‌های یادگیری گذشته به شکل برنامه‌نوشتن کامپیوتری نیست، بلکه کامپیوتر الگوها و روابط داده‌ها را از طریق فرآیندهای ریاضی و محاسباتی خاصی فرا می‌گیرد؛ هوشمندی کامپیوتر از دو مأخذ حاصل می‌شود که یکدیگر را تقویت می‌کنند: یکی ساختاری است که الگوی ریاضی و محاسبات بهینه بر داده‌ها اعمال می‌کنند، و دیگری محتوای اطلاعاتی داده‌هاست.

سخنران آن‌گاه چند اصطلاح در یادگیری ماشینی را توضیح دادند: نخست این‌که feature در یادگیری ماشینی به همان معنای «متغیر» در اقتصادسنجی است؛ یعنی اصطلاحات «ویژگی» و «متغیر» اغلب به جای یکدیگر برای اشاره به متغیرهای ورودی مدل به کار می‌روند. به عنوان مثال، در ساخت مدلی برای اعطای وام مسکن، ویژگی‌ها شامل جنسیت سرپرست خانوار یا متقاضی، تعداد اعضای خانواده، مالکیت تعداد خودروی خانواده، نوع شغل، درآمد، سن و غیره است. توضیح دادند که در زمینه‌ی یادگیری ماشینی، از متغیرها اغلب برای اشاره به ستون‌هایی در مجموعه‌ی داده استفاده می‌شود؛ ویژگی‌ها می‌توانند عددی (به عنوان مثال، سن یا تعداد خودرو) یا مقوله‌ای (به عنوان مثال، نوع شغل یا جنسیت) باشند.

اصطلاح دیگر مفهوم بردارهای حامی است. این بردارها از محاسبات ماشین حاصل می‌شوند و در محاسبه‌ی معادله‌ی صفحه‌ای که داده‌ها را دو قسمت می‌کنند، یا منحنی و رابطه‌ای که داده‌ها را چند قسمت می‌کنند، به کار گرفته می‌شوند. در یادگیری ماشینی، تفسیر این بردارها، ساختار پاسخ را برای پژوهشگر مهیا می‌کند.

مفهوم دیگری که قابل توجه است و از اقتصاد وارد این رشته شده، مفهوم «ارزش شپلی» است. این مفهوم کمک می‌کند که ارزش هر کدام از ویژگی‌ها یا متغیرها را در مجموع محاسبات تعیین کنیم؛ یا ارزش هر ویژگی را برای تبیین هر کدام از مشاهدات محاسبه نماییم. ارزش شپلی نقش اساسی در توضیح‌پذیری نتایج دارد. این امر از الزاماتی است که امروزه از برون‌ریزهای هوش مصنوعی مطالبه می‌شود و اتحادیه‌ی اروپا خصوصا بر ضرورت توضیح پذیری این سیستم‌ها تاکید دارد. محاسبه‌ی ارزش شپلی در این مورد بسیار کارساز است.

هر ویژگی (یا متغیر) دارای مقادیر متناظر برای هر نمونه در مجموعه‌ی داده است. در یک مجموعه‌ی داده، هر ردیف معمولاً یک نمونه یا مشاهده‌ی فردی را نشان می‌دهد (به عنوان مثال، یک وام)، و هر ستون نشان‌دهنده‌ی یک ویژگی (به عنوان مثال سن، درآمد،

یا جنسیت) است. مقادیرِ درون ستون‌ها، اندازه‌گیری‌ها (ارزش‌ها) را برای هر نمونه ارائه می‌کنند. در طول مرحله‌ی آموزش مدل، الگوریتم یاد می‌گیرد این مقادیر را با متغیر هدف (نتیجه‌ای که می‌خواهیم پیش‌بینی کنیم) مرتبط کند، و به مدل اجازه‌ی تعمیم و پیش‌بینی داده‌های جدید را بدهد.

دکتر طبیبیان اصطلاح SVM (Support Vector Machine) را هم توضیح داد: الگوریتم قدرتمند یادگیری ماشینی است که برای کارهای طبقه‌بندی، به ویژه برای طبقه‌بندی‌های دوتایی (باینری) استفاده می‌شود؛ جایی که هدف جداسازی نقاط داده به دو گروه مجزا بر اساس ویژگی‌های آنهاست. کار SVM یافتن ابرصفحه‌ای است که نقاط داده‌ی گروه‌های مختلف را به بهترین نحو از هم جدا می‌کند، به گونه‌ای که حاشیه‌ی بین گروه‌بندی‌ها را حداکثر کند. حاشیه‌ی مورد اشاره فاصله‌ی بین این صفحه و نزدیک‌ترین نقاط داده در هر گروه است. به این نزدیک‌ترین نقاط داده به ابرصفحه «بردارهای پشتیبان» Support Vectors می‌گویند که بیشترین تأثیر را در تعیین حاشیه و مرز تصمیم دارند. در فضای دو بعدی، ابرصفحه خطی است که نقاط داده دو گروه‌بندی را از هم جدا می‌کند. در فضاهای با ابعاد بیشتر، ابرصفحه به مرز تصمیم‌گیری خطی تبدیل می‌شود. گفتیم که «حاشیه» فاصله‌ی بین ابرصفحه و نزدیک‌ترین نقاط داده از هر گروه‌بندی است. هدف SVM یافتن ابرصفحه‌ای است که این حاشیه را به حداکثر می‌رساند، و به افزایش استحکام مدل در داده‌های جدید کمک می‌کند. در مواردی که داده‌ها به صورت خطی در فضای ورودی تفکیک‌پذیر نباشند، SVM از فضایی با ابعاد بالاتر استفاده می‌کند، جایی که داده‌ها به صورت خطی قابل‌جداسازی شوند.

سخنران اصطلاح کرنل یا هسته را هم توضیح داد: در یادگیری ماشینی، کرنل به تابعی اطلاق می‌شود که فاصله‌ی بین نقاط داده را محاسبه می‌کند. از کرنل‌ها معمولاً در تکنیک‌هایی مانند ماشین‌های بردار پشتیبان(SVM) و برخی الگوریتم‌های «کاهش ابعاد» استفاده می‌شود. در SVM ها، کرنل‌ها نقش بسزایی در یافتن مرزهای تصمیم‌گیری بهینه دارند. تعیین کرنل خطی یا غیر خطی الگوریتم‌های یادگیری را قادر می‌سازد تا در فضاهای با ابعاد بالا یا غیرخطی بدون محاسبه‌ی صریح مختصات نقاط، کار کنند و نتیجه به دست آورند.

کرنل‌های خطی برای جداکردن داده‌هایی استفاده می‌شود که به صورت خطی تفکیک‌شدنی هستند، به این معنی که نقاط داده از گروه‌بندی‌های مختلف را می‌توان با یک خط مستقیم (در دو بعد)، یک ابرصفحه (در بیش از دو بعد) یا یک مرز تصمیم‌گیری خطی جدا کرد. این کرنل‌ها برای مسائلی مناسب‌اند که جداکننده‌ی دو گروه بین ویژگی‌ها و گروه‌بندی‌ها، رابطه‌ای نزدیک به خطی است. ازکرنل‌های غیرخطی زمانی استفاده می‌شود که داده‌ها به صورت خطی در فضای ویژگی اصلی تفکیک‌پذیر نباشند. کرنل‌های غیرخطی به الگوریتم‌های یادگیری ماشینی اجازه می‌دهند تا داده‌ها را به فضایی با ابعاد بالاتر تبدیل کنند؛ جایی که به صورت خطی قابل جداسازی می‌شوند، و جداسازی مؤثر را ممکن می‌کنند.

هرچند که سخنران معتقد بود برای شروع به یادگیری ماشینی، مهارت‌های زیادی لازم نیست، اما در طول سخنرانی روشن شد که برای تسلط بر آن، به ترکیبی از مهارت‌ها و دانش‌ها نیاز می‌رود: به ریاضیات (نظریه‌ی مجموعه‌ها، جبر خطی، حساب دیفرانسیل و انتگرال، و الگوریتم‌های بهینه‌سازی) و آمار؛ به آشنایی با فنون رگرسیون، درخت تصمیم‌گیری، شبکه‌های عصبی و خوشه‌بندی؛ برای پیاده‌سازی و آزمایش الگوریتم‌های یادگیری به اکسل، پایتون و آر (R)؛ به توان درک انواع مختلف الگوریتم‌های یادگیری

ماشینی؛ به توان مدیریت و پردازش داده‌ها؛ به قدرت تشخیص ویژگی‌های مرتبط از داده‌ها برای بهبود عملکرد مدل‌های یادگیری ماشینی؛ به توان ارزیابی و اعتبارسنجی مدل؛ به داشتن تخصص در حوزه‌ای که در آن از یادگیری ماشینی استفاده می‌شود؛ ... . دکتر طبیبیان استفاده از یادگیری ماشینی بدون برنامه‌نویسی را تا حدودی امکان‌پذیر می‌داند. به لطف توسعه‌ی ابزارها و سامانه‌های کاربرپسند، بدون کد و یا کم‌کد، ساده‌سازی فرآیند ساخت مدل‌های یادگیری ماشینی ممکن شده که به کاربران اجازه می‌دهد بدون نوشتن کد، مدل‌های ML خود را ایجاد کنند. سامانه‌ی colab.research.google یکی از این سامانه‌هاست. در امر کدنویسی هم ابزاری که بسیار کارگشا است chatgpt است. البته وی تاکید می‌کند که گرچه این ابزارها میزان کدنویسی مورد نیاز را کاهش می‌دهند، اما نیاز به درک اصول یادگیری ماشینی هنوز برای انتخاب ویژگی‌های مناسب، پردازش داده‌ها و تفسیر مؤثر نتایج بسیار مهم است. دکتر طبیبیان نشان داد که سامانه‌ی رایگان Google Colab محیط مناسبی برای نوشتن و اجرای کد پایتون به ویژه جهت یادگیری ماشینی است. راه‌اندازی آن نیازی به نصب نرم‌افزار ندارد، دسترسی آسان مبتنی بر اَبر و با اتصال اینترنتی لازم دارد. این شیوه دسترسی رایگان به منابع محاسباتی دارد، و به‌راحتی روی گوگل درآیو ذخیره می‌شود. در واقع Colab ابزار آموزشی ساده‌ای به‌خصوص برای مبتدیانی است که مفاهیم یادگیری ماشینی را فرا می‌گیرند؛ و همچنین برای پژوهشگرانی است که امکان محاسباتی محدودی در اختیار دارند.

دکتر طبیبیان اشاره‌ای کوتاه هم به استفاده از روش ناپارامتریک بوت استرپینگ در یادگیری ماشینی داشت. هنگامی که داده‌ها فراوان هستند و مشاهدات زیاد است، به‌ویژه برای تنظیم پارامترهای مدل، و انتخاب و ارزیابی مدل‌ها، و بالاخص الگوریتم‌های جنگل‌های تصادفی، بوت استرپینگ تکنیکی ساده برای تخمین توزیع و تکرار نمونه‌گیری‌های لازم برای شبیه‌سازی مجموعه‌های آموزشی و آزمایشی متعدد تلقی می‌شود. سخنران درک مفاهیم بوت استرپینگ و نحوه‌ی کاربرد آن در یادگیری ماشینی را برای ساخت مدل‌های قوی و دقیق بسیار ارزشمند می‌دانست.

دکتر طبیبیان تصریح کرد که طبقه‌بندی جنبه‌ی مهم و برجسته‌ای از یادگیری ماشینی است. این کار شامل تخصیص نقاط داده به گروه‌بندی‌ها یا دسته‌های از پیش‌تعریف‌شده بر اساس ویژگی‌های آنهاست. اما یادگیری ماشینی محدود به طبقه‌بندی نیست؛ ML از فنون رگرسیون‌های خطی و پیچیده‌تر، خوشه‌بندی (گروه‌بندی نقاط داده‌ی مشابه بر اساس ویژگی‌های آنها، بدون دسته‌بندی‌های از پیش‌تعریف‌شده)، تکنیک کاهش ابعاد (کاهش تعداد ویژگی‌ها در عین حفظ اطلاعات مرتبط)، شناسایی نقاط داده‌ی نادر (برای شناسایی تقلب، محصولات معیوب و امنیت شبکه)، پردازش زبان طبیعی، تحلیل سری‌های زمانی، ... استفاده می‌کند.

سخنران معتقد بود که بی‌شمار کاربرد برای ML می‌توان برشمرد، و در اقتصاد هم ML در زمینه‌های مختلفی چون پیش‌بینی‌ها از جمله پیش‌بینی شاخص‌های اقتصادیِ رشد تولید ناخالص داخلی، روندهای بازار سهام و نرخ تورم؛ تحلیل‌های مالی، به‌ویژه تحلیل مجموعه داده‌های مالی بزرگ برای شناسایی فرصت‌های سرمایه‌گذاری بالقوه؛ امتیازدهی اعتباری؛ تحلیل رفتار و ترجیحات مصرف‌کننده، تقسیم‌بندی بازارها؛ مدیریت ریسک؛ ... یادگیری ماشینی کاربرد دارد.

در مورد استفاده از یادگیری ماشینی برای امتیازدهی اعتباری در وام‌های مصرفی جهت بهبود دقت و بهینه‌سازی تصمیمات وام، دکتر طبیبیان با ذکر مثال توضیح داد که نخست مجموعه‌ی داده‌های تاریخی درخواست وام، از جمله ویژگی‌هایی مانند درآمد، سابقه‌ی

اشتغال، سابقه‌ی اعتباری، مبلغ وام، مدت وام و موارد دیگر جمع‌آوری و پیش‌پردازش می‌شود. آن‌گاه به ترتیب ویژگی‌های مرتبط تعریف می‌شود؛ مدل (الگوریتم‌های یادگیری ماشینی مناسب برای طبقه‌بندی) انتخاب می‌شود؛ با چندین الگوریتم آزمایش می‌شود تا بهترین الگوریتم برای مجموعه‌ی داده یافت شود؛ ... .

دکتر طبیبیان بر این خاصیت ML تأکید فراوان داشت که هر وقت یک مدل یادگیری ماشینی را در مورد موضوعی خاص آموزش بدهیم، اغلب می‌توانیم از آن برای پیش‌بینی یا طبقه‌بندی موارد جدید و دیده‌نشده بارها و بارها استفاده کنیم؛ وی تعمیم را یکی از مزایای کلیدی یادگیری ماشینی می‌دانست. یعنی، یک مدل یادگیری ماشینی آموزش‌دیده باید به‌خوبی به داده‌های جدیدی تعمیم داده شود که آموزش‌ندیده‌اند؛ باید بتوان الگوها را شناسایی کرد و پیش‌بینی‌های دقیقی برای موارد مشابه با موارد آموزش‌دیده انجام داد. وی روشن کرد که محدودیت‌های مدل، کیفیت داده‌ها، تفسیرپذیری، ملاحظات اخلاقی و مقرراتی، ضرورت بازآموزی، ... چالش‌های روبه‌روی رویکرد ML است.

طی جلسه، دکتر طبیبیان مراقب بود که حاضران یادگیری ماشینی را با مفاهیم اقتصادسنجی درهم‌نیامیزند. تصریح کرد که یادگیری ماشینی جایگزین مستقیمی هم برای اقتصادسنجی نیست. اقتصادسنجی شاخه‌ای از علم اقتصاد است که از روش‌های آماری برای تعیین کمیت و تبیین روابط اقتصادی و پیش‌بینی پدیده‌های اقتصادی استفاده می‌کند. یادگیری ماشینی حوزه‌ی وسیع‌تری است که کامپیوترها را قادر می‌سازد الگوهایی را از داده‌ها یاد بگیرند و در حوزه‌های بسیار متنوع پیش‌بینی کنند یا تصمیم بگیرند. مدل‌های اقتصادسنجی اغلب با هدف ارائه‌ی ضرایب تفسیرپذیر با معانی اقتصادی طراحی می‌شوند. مدل‌های یادگیری ماشینی می‌توانند پیچیده باشند و غالباً دقت پیش‌بینی را بر تفسیرپذیری اولویت می‌دهند. تکنیک‌های یادگیری ماشینی انعطاف‌پذیرتر از مدل‌های اقتصادسنجی‌اند و همیشه به فرضیات دقیق نیاز ندارند. به نظر می‌رسد که دکتر طبیبیان این دو رشته را مکمل یکدیگر می‌داند و معتقد است که گرچه فنون یادگیری ماشینی تحلیل اقتصادی را با مدیریت مجموعه‌های داده‌ی بزرگ و پیچیده و روابط غیرخطی بهبود بخشیده است، اما اقتصادسنجی به مثابه‌ی ابزاری حیاتی برای آزمایش نظریه‌ی اقتصادی، تحلیل سیاست‌ها و ارائه‌ی تفسیرهای اقتصادی به عمر خود ادامه خواهد داد.

غیر از ارائه‌ی مثال‌های اقتصادی و طراحی الگوریتم‌های ML برای آنها، دکتر طبیبیان علاقه‌مند بود که دو مقوله‌ی اقتصادی (ارزش شپلی و ضریب جینی) را توضیح داده و برای آنها مدل‌سازی ML کند. با توجه به کمبود وقت، ایشان فقط ارزش شپلی را توضیح دادند، و فرصت مدل‌سازی نیافتند. دکتر طبیبیان اشاره کرد که قبل از «نظریه‌ی شپلی،» در اقتصاد فقط «ارزش بازار» را داشتیم. «ارزش شپلی» مفهومی از نظریه‌ی بازی‌هاست که برای محاسبه‌ی آن از رویکرد مبتنی بر ریاضی جهت حل مسائل تخصیص در ترتیبات مشارکتی (بازیکنان برای دستیابی به اهداف مشترک با یکدیگر همکاری می‌کنند) استفاده می‌شود.

علت تعریف و استفاده از ارزش شپلی به عنوان مثالی در زمینه‌ی یادگیری ماشینی توسط دکتر طبیبیان احتمالاً به زمینه‌ی تفسیرپذیری مدل و اهمیت ویژگی برمی‌گردد. در یادگیری ماشینی، درک مشارکت ویژگی‌ها یا متغیرهای مختلف در پیش‌بینی‌های یک مدل برای تفسیرپذیری و اعتماد بسیار مهم است. مقدار ارزش شپلی را می‌توان برای نسبت‌دادن «ارزش» یا تأثیر هر ویژگی در یک پیش‌بینی اعمال کرد. این به توضیح این‌که چرا هر مدل، پیش‌بینی خاصی انجام می‌دهد، و چگونه ویژگی‌های مختلف برای رسیدن

به نتیجه با یکدیگر در تعامل‌اند، کمک می‌کند. دکتر طبیبیان از این مثال برای آموزش یادگیری ماشینی استفاده کرد، چون مقدار شپلی با رویکرد «مقادیر» شپلی در یادگیری ماشین تطبیق داده شده است؛ جایی که از آن برای توضیح پیش‌بینی مدل‌های پیچیده مانند شبکه‌های عصبی، جنگل‌های تصادفی و غیره استفاده می‌شود. با محاسبه‌ی مقادیر شپلی برای ویژگی‌ها، محققان و متخصصان می‌توانند در مورد این‌که کدام ویژگی‌ها بیشترین تأثیر را بر پیش‌بینی‌ها دارند و این‌که چگونه تعاملات آنها بر نتایج تأثیر می‌گذارند، تصویر روشنی به دست آورند.

لوید شپلی، ریاضیدان و اقتصاددان برنده‌ی جایزه‌ی نوبل امریکایی، مفهوم ارزش Shapley را ارائه کرده که روشی برای توزیع عادلانه‌ی ارزش کل یا بازده تولیدشده در یک بازی مشارکتی است. «ارزش شپلی» فراتر از اقتصاد، در زمینه‌های علوم سیاسی، مذاکره و تخصیص منابع کاربرد دارد. ارزش شپلی راهی برای تخصیص کل ارزش ایجادشده با همکاری بین بازیکنان به شیوه‌ای منصفانه فراهم می‌کند. به بازیکنان برای مشارکت‌شان پاداش داده می‌شود، اما نقش‌های متفاوتی را که بازی می‌کنند و ترتیب پیوستن آنها به ائتلاف‌ها را نیز در نظر می‌گیرد.

بخش قابل‌ملاحظه‌ای از سخنرانی بیش از دو ساعته‌ی دکتر طبیبیان به ارائه‌ی مثال و عملاً طراحی مدل ML و نشان‌دادن نتایج محاسباتی تخصیص یافت. حاضران در جلسه در شگفت شدند که سخنران بر تمام جزئیات ریاضی و آماری محاسبات، نگاشت مدل، کدنویسی پایتون و آر، استفاده از اکسل در طبقه‌بندی، استفاده از سامانه‌های ساده‌تر کاربرپسند، و کارکرد ابزار Solver اکسل برای تعیین مقادیر بهینه‌ی متغیرهای تصمیم‌ساز مسلط بود. برای جوانان حاضر در جلسه بسیار جالب بود که استاد محمد طبیبیان، با چندین کتاب و صدها مقاله در حوزه‌های تاریخ نظریه‌های اقتصادی، معرفت‌شناسی اقتصاد، فلسفه‌ی اقتصاد، روش علم، اقتصاد ایران، اقتصاد خرد، اقتصاد کلان، اقتصاد سیاسی ... و نیز با تجربه‌ی مشارکت در طراحی دو برنامه‌ی توسعه‌ی اقتصادی موفق در ایران، در سن ۷۵ سالگی خود را موظف می‌داند که در رشته‌هایی چون اقتصادسنجی، یادگیری ماشینی، تحقیق در عملیات، اقتصاد ریاضی، هوش مصنوعی، مدلسازی ریاضی، نظریه‌ی اطلاعات، علوم شناختی، ... خود را به‌روز نگاه دارد.