تفسیر مدل‌های یادگیری ماشین

پیشگفتار

سرعت واگذاری وظایف، به مدل‌های یادگیری ماشین در حال افزایش است. این مدل‌ها درای عملکرد بهتر، سرعت بالاتر، تکثيرپذیری راحتتر، نتایج سازگارتر هستند. اما این مدل‌ها پیش‌بینی‌های خود را توضیح نمی‌دهند که پذیرش و عیب‌یابی آنها را مشکل می‌کند. به عنوان مثال، یک مدل یادگیری ماشین، به سوالات زیر جواب نمی‌دهد:

* تغییرات یک ویژگی ورودی، چه تاثیری در پیش‌بینی خروجی دارد؟
* کدامیک از ویژگی‌های ورودی برای پیش‌بینی مهم‌تر هستند؟
* آیا ویژگی‌ها با یکدیگر تعامل دارند؟
* در پیش‌بینی یک نمونه تک، هریک از مقادیر ویژگی‌ها چه سهمی دارند؟

با تفسیر مدل‌های یادگیری ماشین می‌توان به این سوالات پاسخ داد.

فصل اول کتاب درباره مفاهیم تفسیرپذیری، مطالبی را ارائه می‌دهد. این فصل خلاصه‌ای از کتاب "یادگیری ماشین تفسیرپذیر" می‌باشد. در ادامه، روش‌های تفسیر متداول شرح داده می‌شود. در ابتدای هر روش، به صورت شهودی و تئوری معرفی می‌گردد و سپس با ذکر مثال‌های متنوع، درک مناسبی از روش مذکور ارائه شود. در این کتاب فقط بر روی داده‌های جدولی (داده‌هایی هستند که در جداول قرار می‌گیرند و هر ردیف نشان‌دهنده یک نمونه و هر ستون یک ویژگی است) متمرکز است. مثال‌ها در نرم‌افزار MATLAB پیاده سازی شده‌اند. برای درک مناسب تفاسیر، باید دانش کافی از مجموعه‌داده استفاده‌شده برای آموزش مدل یادگیری ماشین وجود داشته باشد. به همین دلیل در پیوست کتاب، کلیه مجموعه داده‌های بکاربرده‌شده در مثال‌ها توضیح داده شده است.

مقدمه

مجموعه داده‌ها

CreditRating\_Historical

3932 مشاهده، 7 ویژگی و 1 خروجی

|  |  |  |
| --- | --- | --- |
|  |  |  |
| ID |  |  |
| WC\_TA |  |  |
| RE\_TA |  |  |
| EBIT\_TA |  |  |
| MVE\_BVTD |  |  |
| S\_TA |  |  |
| Industry |  |  |
| Rating | طبقه‌ای | A, AA, AAA, B, BB, BBB, CCC |

