



انتشارات، شماره ۶۹

مبانی یادگیری ماشین برای تحلیل پیشگویانه داده‌ها الگوریتم‌ها، نمونه مثال‌ها و موردهای مطالعاتی

نویسندگان:

جان کلهر، بریان مکنامی، ایفقه دل‌رسی

مترجمین:

امیرکیوان شفیع‌ی (عضو هیئت علمی دانشگاه بیرجند)

صادق پاسبان (عضو هیئت علمی دانشگاه بیرجند)

زمستان ۱۴۰۰



مبانی یادگیری ماشین برای تحلیل پیشگویانه داده‌ها الگوریتم‌ها، نمونه‌های مثال‌ها و مورد‌های مطالعاتی

مترجمین: امیرکیوان شفیعی، صادق یاسیان

سرشناسه: کلیهر، جان دی.، ۱۹۷۴-م. Kelleher, John D., 1974-

عنوان و نام پدیدآور: مبانی یادگیری ماشین برای تحلیل پیشگویانه داده‌ها الگوریتم‌ها، نمونه‌های مثال‌ها و مورد‌های مطالعاتی / نویسندگان: جان کلیهر، بریان مک‌نامی، ایفہ داری؛ مترجمین: امیرکیوان شفیعی، صادق یاسیان.

مشخصات نشر: بیرجند: دانشگاه بیرجند، انتشارات، ۱۴۰۰.

مشخصات ظاهری: ف، ۴۸۲ص: مصور.

شابک: ۹۷۸-۶۲۲-۶۵۸۲-۳۱-۵

وضعیت فهرست نویسی: فیبا

یادداشت: عنوان اصلی: Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies, 2020.

موضوع: فراگیری ماشینی Machine Learning

شناسه افزوده: مک‌نامی، بریان، ۱۹۷۸-م.

شناسه افزوده: MacNamee, Brian, 1978-

شناسه افزوده: داری، ایفہ، ۱۹۷۸-م.

شناسه افزوده: D'Arcy, Aoife, 1978-

شناسه افزوده: یاسیان، صادق، ۱۳۶۶-م.

شناسه افزوده: شفیعی، امیرکیوان، ۱۳۵۸-م.

شناسه افزوده: دانشگاه بیرجند، انتشارات

رده بندی کنگره: Q۳۲۵/۵

رده بندی دیویی: ۵۱۹/۲۸۷۰۲۸۵۶۳۱

شماره کتابشناسی ملی: ۸۵۲۱۷۷۹

اطلاعات رکورد کتابشناسی: فیبا

ناشر: دانشگاه بیرجند

تیراژ: ۵۰۰ جلد

چاپ اول: دی ماه ۱۴۰۰

امور فنی: انتشارات فکرپر

صفحه آرای: فائزه دستگردی

طراح جلد: نفیسه شکرانی

لیتوگرافی و چاپ: واصف

قیمت: ۱۱۰۰۰۰ ریال

شابک: ۹۷۸-۶۲۲-۶۵۸۲-۳۱-۵

آدرس: بیرجند- انتهای بلوار دانشگاه - پردیس شوکت آباد دانشگاه بیرجند - کتابخانه مرکزی و مرکز نشر (ساختمان دکتر محمدحسن گنجی) - اداره انتشارات دانشگاه بیرجند.

کدپستی: ۹۷۱۷۴۳۴۷۶۵ - تلفن: ۰۵۶-۳۱۰۲۲۶۵۰ / پست الکترونیکی: entesharat@birjand.ac.ir

فروشگاه اینترنتی کتاب دانشگاه بیرجند: <http://press.birjand.ac.ir>

«کلیه حقوق متعلق به انتشارات دانشگاه بیرجند است»

فهرست مطالب

فصل ۱: یادگیری ماشین برای تحلیل پیشگویانه داده‌ها

۱-۱	تحلیل پیشگویانه داده‌ها	۲
۲-۱	یادگیری ماشین	۳
۳-۱	چگونگی عملکرد یادگیری ماشین	۷
۴-۱	اشتباهات یادگیری ماشین	۱۱
۵-۱	چرخه زندگی پروژه تحلیل پیشگویانه داده‌ها : CRISP-DM	۱۲
۶-۱	ابزارهای تحلیل پیشگویانه داده‌ها	۱۴
۷-۱	مسیر پیش رو	۱۶
۸-۱	تمرین	۱۷

فصل ۲: داده‌ها برای درک تصمیم‌گیری

۱-۲	تبدیل مسائل کسب و کار به راه حل‌های تحلیل	۱۹
۱-۱-۲	مورد مطالعاتی: تقلب در بیمه ماشین	۲۱
۲-۲	ارزیابی امکان‌سنجی	۲۲
۱-۲-۲	مورد مطالعاتی: تقلب بیمه ماشین	۲۳
۳-۲	طراحی جدول پایگاه تحلیل	۲۴
۱-۳-۲	مورد مطالعاتی: تقلب بیمه ماشین	۲۷
۴-۲	طراحی و پیاده‌سازی ویژگی‌ها	۲۹
۱-۴-۲	انواع مختلف داده‌ها	۳۰
۲-۴-۲	انواع مختلف ویژگی‌ها	۳۰
۳-۴-۲	مدیریت زمان	۳۲
۴-۴-۲	مسائل قانونی	۳۵
۵-۴-۲	پیاده‌سازی ویژگی‌ها	۳۷
۶-۴-۲	مورد مطالعاتی: تقلب بیمه ماشین	۳۷
۵-۲	خلاصه	۴۰
۶-۲	مطالعه بیشتر	۴۲
۷-۲	تمرین	۴۲

فصل ۳: اکتشاف داده

۱-۳	گزارش کیفیت داده‌ها	۴۸
۱-۱-۳	مورد مطالعاتی: تقلب بیمه ماشین	۴۹
۲-۳	آشنایی با داده‌ها	۵۱
۱-۲-۳	توزیع نرمال	۵۴
۲-۲-۳	مورد مطالعاتی: تقلب بیمه ماشین	۵۵
۳-۳	تشخیص مسائل کیفیت داده‌ها	۵۶
۱-۳-۳	مقادیر گم شده	۵۷
۲-۳-۳	کاردینالیتی نامنظم	۵۷

۵۸	۳-۳-۳ داده‌های پرت
۵۹	۳-۳-۳ مورد مطالعاتی: تقب بيمه ماشين
۵۹	۳-۳-۳ ۱-۴ مقادير گم شده
۵۹	۳-۳-۳ ۲-۴ کارديناليتي نامنظم
۶۰	۳-۳-۳ داده‌های پرت
۶۱	۳-۳-۳ ۴-۴ طرح کيفيت داده‌ها
۶۱	۳-۳-۳ ۴-۴ اداره کردن مسائل کيفيت داده‌ها
۶۱	۳-۳-۳ ۱-۴ اداره کردن مقدير کم تنده
۶۲	۳-۳-۳ ۲-۴ اداره کردن داده‌های پرت
۶۴	۳-۳-۳ مورد مطالعاتی: تقب بيمه ماشين
۶۵	۳-۳-۳ ۵-۳ اکتشاف داده‌های پيشرفته
۶۵	۳-۳-۳ ۱-۵ مصور سازي ارتباطات ميان ويژگي‌ها
۶۶	۳-۳-۳ ۱-۱-۵ مصور سازي جفت ويژگي‌های پيوسته
۶۸	۳-۳-۳ ۲-۱-۵ مصورسازي جفت ويژگي‌های قطعي
۷۱	۳-۳-۳ ۳-۱-۵ مصورسازي يك ويژگي قطعي و يك ويژگي پيوسته
۷۳	۳-۳-۳ ۲-۵ اندازه‌گيري کواريانس و همبستگي
۷۸	۳-۳-۳ ۶-۳ آماده‌سازي داده‌ها
۷۸	۳-۳-۳ ۱-۶ نرمال‌سازي
۸۰	۳-۳-۳ ۲-۶ پياله‌بندي
۸۳	۳-۳-۳ ۳-۶ نمونه‌برداري
۸۴	۳-۳-۳ ۷-۳ خلاصه
۸۵	۳-۳-۳ ۸-۳ مطالعه بيشتر
۸۶	۳-۳-۳ ۹-۳ تمرين

فصل ۴: يادگيري بر پايه اطلاعات

۹۹	۴-۱ هدف
۱۰۲	۴-۲ مباني
۱۰۲	۴-۲-۱ درختان تصميم‌گيري
۱۰۵	۴-۲-۲ مدل آنتروبي شانون
۱۰۹	۴-۲-۳ بهره اطلاعات
۱۱۴	۴-۳ ۳-۳ رويکرد استاندارد: الگوريتم ID3
۱۱۷	۴-۳-۱ مثال واقعي: پيش‌بنی توزیع پوشش گیاهی
۱۲۴	۴-۴ ۴-۴ ضمايم و تغييرات
۱۲۵	۴-۴ ۱-۴ معيارهای انتخاب ويژگي و آلودگي جايگزين
۱۲۹	۴-۴ ۲-۴ بررسي ويژگي‌های توصيفي پيوسته
۱۳۲	۴-۴ ۳-۴ پيش‌بنی هدف‌های پيوسته
۱۳۶	۴-۴ ۴-۴ هرس کردن درخت

۲۲۸	۱-۳-۶ نمونه مثال
۲۳۰	۴-۶ ضمایم و تغییرات
۲۳۱	۱-۴-۶ هموارسازی
۲۳۴	۲-۴-۶ ویژگی های پیوسته: تابع چگالی احتمال
۲۴۵	۳-۴-۶ ویژگی های پیوسته: پیاله بندی
۲۴۸	۴-۴-۶ شبکه های بیزی
۲۵۳	۱-۴-۴-۶ ساخت شبکه های بیزی
۲۵۸	۲-۴-۴-۶ استفاده از شبکه بیزی برای پیش بینی
۲۵۸	۳-۴-۴-۶ پیش بینی با مقادیر ویژگی توصیفی گم شده
۲۶۲	۵-۶ خلاصه
۲۶۴	۶-۶ مطالعه بیشتر
۲۶۵	۷-۶ تمرین

فصل ۷: یادگیری مبتنی بر خطا

۲۶۹	۱-۷ هدف
۲۷۰	۲-۷ مبانی
۲۷۰	۱-۲-۷ رگرسیون خطی ساده
۲۷۲	۲-۲-۷ اندازه گیری خطا
۲۷۵	۳-۲-۷ سطوح خطا
۲۷۶	۳-۷ رویکرد استاندارد: رگرسیون خطی چند متغیره با گرادیان نزولی
۲۷۷	۱-۳-۷ رگرسیون خطی چند متغیره
۲۷۸	۲-۳-۷ گرادیان نزولی
۲۸۳	۳-۳-۷ انتخاب نرخ یادگیری و وزن های اولیه
۲۸۵	۴-۳-۷ نمونه مثال
۲۸۷	۴-۷ ضمایم و تغییرات
۲۸۷	۱-۴-۷ تفسیر مدل های رگرسیون خطی چند متغیره
۲۸۹	۲-۴-۷ مقداردهی نرخ یادگیری با استفاده از زوال وزن
۲۹۱	۳-۴-۷ پردازش ویژگی های توصیفی قطعی
۲۹۲	۴-۴-۷ پردازش ویژگی های هدف قطعی: رگرسیون لجستیک
۲۹۳	۱-۴-۴-۷ پیش بینی اهداف قطعی با استفاده از رگرسیون خطی
۲۹۷	۲-۴-۴-۷ رگرسیون لجستیک
۳۰۰	۳-۴-۴-۷ نمونه مثالی از رگرسیون لجستیک چند متغیره
۳۰۵	۵-۴-۷ مدل سازی رابطه های غیر خطی
۳۱۱	۶-۴-۷ رگرسیون لجستیک چند متغیره
۳۱۴	۷-۴-۷ ماشین های بردار پشتیبان
۳۱۹	۵-۷ خلاصه
۳۲۱	۶-۷ مطالعه بیشتر

فصل ۸: ارزیابی

۳۲۹	۱-۸ هدف
۳۳۰	۳-۸ میانی
۳۳۱	۳-۸ رویکرد استاندارد: نرخ دسته‌بندی نادرست روی یک مجموعه آزمایشی نگهداری شده
۳۳۵	۴-۸ ضمایم و تغییرات
۳۳۵	۱-۴-۸ طراحی آزمایش‌های ارزیابی
۳۳۶	۱-۴-۸ نمونه‌برداری نگهداری شده
۳۳۸	۲-۴-۸ اعتبارسنجی متقابل k لایه‌ای
۳۴۰	۳-۴-۸ اعتبارسنجی متقابل ترک یک
۳۴۱	۴-۴-۸ راه‌اندازی
۳۴۲	۵-۴-۸ نمونه‌برداری خارج از زمان
۳۴۳	۲-۴-۸ معیارهای کارایی: اهداف قطعی
۳۴۳	۱-۴-۸ معیارهای کارایی مبتنی بر ماتریس درهم ریختگی
۳۴۳	۲-۴-۸ دقت، یادآوری و معیار F_1
۳۴۵	۳-۴-۸ میانگین دقت کلاس
۳۴۷	۴-۴-۸ اندازه‌گیری سود و زیان
۳۵۰	۳-۴-۸ معیارهای کارایی: امتیاز پیش‌بینی
۳۵۲	۱-۴-۸ منحنی مشخصه عامل گیرنده
۳۵۷	۲-۴-۸ آمار کلموگروف-اسمیرنوف
۳۶۰	۳-۴-۸ اندازه‌گیری سود و ترفیع
۳۶۴	۴-۴-۸ معیارهای کارایی: اهداف چندجمله‌ای
۳۶۷	۵-۴-۸ معیارهای کارایی: اهداف پیوسته
۳۶۷	۱-۴-۸ معیارهای اصلی خطا
۳۷۰	۲-۴-۸ اندازه‌گیری‌های مستقل از دامنه خطا
۳۷۰	۶-۴-۸ ارزیابی مدل‌ها پس از استقرار
۳۷۱	۱-۴-۸ نظارت بر تغییرات معیارهای کارایی
۳۷۲	۲-۴-۸ نظارت تغییرات توزیع خروجی مدل
۳۷۴	۳-۴-۸ نظارت بر تغییرات توزیع ویژگی‌های توصیفی
۳۷۵	۴-۴-۸ آزمایش‌های مقایسه‌ای با استفاده از یک گروه کنترل
۳۷۷	۵-۸ خلاصه
۳۷۷	۶-۸ مطالعه بیشتر
۳۷۸	۷-۸ تمرین

فصل ۹: مورد مطالعاتی: ریزش مشتری

۳۸۵	۱-۹ درک کسب و کار
۳۸۸	۲-۹ درک داده‌ها

پیشگفتار

هدف ما از نوشتن این کتاب، ارائه متنی قابل دسترس و مقدماتی برای مبانی یادگیری ماشین و روش‌هایی است که یادگیری ماشین در عمل برای حل مسائل تحلیلی داده‌های پیشگو در کسب و کار، علوم و دیگر زمینه‌های سازمانی استفاده می‌شود. به این ترتیب، این کتاب از موضوعات استاندارد که در کتاب‌های یادگیری ماشین پوشش داده می‌شود فراتر رفته و چرخه زندگی پروژه‌های تحلیل پیشگو، آماده‌سازی داده‌ها، طراحی ویژگی و استقرار مدل را نیز پوشش می‌دهد.

این کتاب برای استفاده در واحدهای درسی یادگیری ماشین، داده‌کاوی، تحلیل داده‌ها یا هوش مصنوعی در دوره‌های کارشناسی و کارشناسی ارشد علوم کامپیوتر، علوم طبیعی و اجتماعی، مهندسی و کسب و کار در نظر گرفته شده است. این حقیقت که این کتاب با موردهای مطالعاتی، کاربرد یادگیری ماشین را در زمینه صنعت تحلیل داده‌ها شرح می‌دهد نیز آنرا متنی مناسب برای متخصصانی کرده است که به دنبال مقدمه‌ای به عنوان کتاب درسی دوره‌های آموزش صنعتی در این حوزه‌ها می‌گردند.

طراحی کتاب با توجه به سال‌های متمادی تجارب تدریس یادگیری ماشین انجام شده و موضوعات کتاب در کلاس درس توسعه یافته و آزمایش شده است. در نوشتن این کتاب اصول راهنمای زیر را بکار برده‌ایم تا مطالب قابل دسترس باشند:

۱- توضیح مهم‌ترین و معروف‌ترین الگوریتم‌ها به‌جای بررسی اجمالی تمام مطالب یادگیری ماشین. به عنوان معلم اعتقاد داریم که دادن دانش عمیق به دانشجویان از مفاهیم اساسی بر پایه یک زمینه، باعث می‌شود پایه‌ای محکم ایجاد شود تا خود دانشجویان بتوانند آن موضوع را کاوش کنند. این تمرکز دقیق‌تر به ما امکان می‌دهد تا زمان بیشتری را صرف معرفی، توضیح، تشریح و زمینه‌سازی الگوریتم‌هایی کنیم که در این موضوع و کاربردهای آن اساسی می‌باشند.

۲- توضیح غیر رسمی آنچه الگوریتم، سعی در انجام آن دارد قبل از توصیف رسمی چگونگی کارکرد آن. تدارک این مقدمه غیر رسمی برای هر موضوع، به دانشجویان، پایه محکمی جهت آمادگی برای آموختن مطالب فنی بیشتر می‌دهد. تجربه ما در تدریس این مطالب به مخاطبانی متشکل از دانشجویان کارشناسی، کارشناسی ارشد و حرفه‌ای‌ها نشان داده است که این مقدمات غیر رسمی، دانشجویان را قادر به دسترسی ساده‌تر به موضوع می‌کند.

۳- تدارک نمونه کارهای کامل. در این کتاب کارهای کاملی ارائه کرده‌ایم زیرا این امر خواننده را قادر می‌سازد جزئیات دقیق آن‌ها را درک کند.

ساختار کتاب

در هنگام آموزش یک موضوع فنی، نشان دادن کاربرد مفاهیم توضیح داده شده در مسائل زندگی واقعی مهم است. به همین دلیل، یادگیری ماشین را در زمینه تحلیل پیشگویانه داده‌ها عرضه کرده‌ایم که یک کاربرد صنعتی مهم و در حال رشد برای یادگیری ماشین است. پیوند میان یادگیری ماشین و تحلیل داده‌ها در تمام فصول کتاب برقرار است. در فصل ۱، یادگیری ماشین را معرفی کرده‌ایم و نقش آنرا در طول عمر پروژه تحلیل داده‌های استاندارد توضیح می‌دهیم. در فصل ۲ چارچوبی را برای طراحی و ساخت راه حل تحلیل پیشگویانه براساس یادگیری ماشین که نیاز کسب و کار برآورده می‌کند فراهم می‌کنیم. در تمام الگوریتم‌های یادگیری ماشین فرض می‌شود که یک مجموعه داده برای آموزش وجود دارد و در فصل ۳ توضیح می‌دهیم که قبل از آنکه از مجموعه داده در ساخت مدل پیش‌بینی بهره ببریم چطور آنرا طراحی کنیم، بسازیم و کنترل کیفیت نماییم.