

مقدمه‌ای بر یادگیری ماشین با Python

راهنمای کارشناسان علوم داده

نویسندگان: آندریاس سی. مولرو و سارا گوئیدو

www.ketab.ir

مترجمان:

دکتر محمد محسن صدر | دکتر مصطفی اخوان صفار | مهندس حمید شیخ‌ویسی

(اعضای هیأت علمی دانشگاه پیام نور)

مقدمه‌ای بر یادگیری ماشین با Python

راهنمای کارشناسان علوم داده

مترجمان: دکتر محمدحسن صدر، دکتر مصطفی اخوان صفار، مهندس حمید شیخ‌ویسی

ناشر: انتشارات آتی‌نگر

ناشر همکار: انتشارات وینا

طراحی جلد و صفحه‌آرایی: همتا بیداریان

چاپ اول، ۱۴۰۰

شمارگان: ۲۰۰ نسخه

قیمت: ۱,۴۵۰,۰۰۰ ریال

شابک: ۹۷۸-۶۲۲-۷۵۷۱-۴۲-۴

ISBN: 978-622-7571-42-4

حق چاپ برای انتشارات آتی‌نگر محفوظ است.

نشانی دفتر فروش: خیابان جمال‌الده جنوبی، روبه‌روی کوچه رشتچی، پلاک ۱۴۴، واحد ۱

نمابر: ۶۶۵۶۵۳۳۷

تلفن: ۸-۶۶۵۶۵۳۳۶

www.ati-negar.com * info@ati-negar.com



سرشناسه: سی. مولر، آندریاس C. Müller, Andreas

مقدمه‌ای بر یادگیری ماشین با Python «راهنمای کارشناسان علوم داده»: نویسندگان: آندریاس سی. مولر و سارا گوئیدو؛

مترجمان: محمدحسن صدر، مصطفی اخوان صفار، حمید شیخ‌ویسی

تهران: آتی‌نگر، وینا ۱۴۰۰

۴۱۴ ص.: مصور، جدول، نمودار.

ISBN: 978-622-7571-42-4

فیبا.

یادداشت: عنوان اصلی کتاب: Introduction to machine learning with Python : a guide for data scientists, 2017.

یادداشت: کتابنامه به‌صورت زیرنویس.

موضوع: پایتون (زبان برنامه‌نویسی کامپیوتر) - Python (Computer program language)

موضوع: زبان‌های برنامه‌نویسی کامپیوتر - Programming languages (Electronic computers) - داده‌کاوی - Data mining

شناسه افزوده: گوئیدو، سارا، Sarah, Guido

شناسه افزوده: صدر، محمدحسن، ۱۳۵۷- مترجم

شناسه افزوده: اخوان صفار، مصطفی، ۱۳۶۱- مترجم

شناسه افزوده: شیخ‌ویسی، حمید، ۱۳۶۰- مترجم

رده‌بندی کنگره

رده‌بندی دیویی

شماره کتابشناسی ملی

QAY۶/۷۳

۰۰۵/۱۳۳

۸۷۲۲۹۸۳

فهرست مطالب

۹	پیشگفتار
۱۵	فصل ۱: مقدمه
۱۶	چرا یادگیری ماشین؟.....
۱۶	چه مشکلاتی را می‌توان به کمک یادگیری ماشین حل و فصل کرد؟.....
۲۰	کارها و داده‌های خود را بشناسید.....
۲۱	چرا Python؟.....
۲۱	Scikit-learn.....
۲۲	نصب scikit-learn.....
۲۳	کتابخانه‌ها و ابزارهای اساسی.....
۲۳	Jupyter Notebook.....
۲۳	NumPy.....
۲۴	SciPy.....
۲۵	matplotlib.....
۲۶	pandas.....
۲۸	mglearn.....
۲۸	مقایسه Python 2 با Python 3.....
۲۹	نسخه‌های استفاده شده در این کتاب.....
۳۰	اولین کاربرد: طبقه‌بندی گونه‌های گل زنبق.....
۳۱	کار با داده‌ها.....
۳۴	ارزیابی موفقیت: داده‌های آموزش و ارزیابی (آزمون).....
۳۶	ابتدا داده‌های خود را بررسی کنید.....
۳۸	اولین مدل خود را بسازید: k-نزدیک‌ترین همسایگان.....
۴۰	انجام پیش‌بینی‌ها.....
۴۱	ارزیابی مدل.....
۴۲	خلاصه و چشم‌انداز.....
۴۵	فصل ۲: یادگیری با نظارت
۴۵	طبقه‌بندی و رگرسیون.....

۴۷	تعمیم، بیش‌برازش و کم‌برازش
۴۹	رابطه پیچیدگی مدل با ابعاد مجموعه داده
۵۰	الگوریتم‌های یادگیری ماشین با نظارت
۵۱	چند نمونه مجموعه داده
۵۶	k-نزدیک‌ترین همسایگان
۶۶	مدل‌های خطی
۹۱	طبقه‌بندهای بیز ساده
۹۳	درخت‌های تصمیم
۱۰۷	گروه‌های درخت‌های تصمیم
۱۱۸	ماشین‌های بردار پشتیبانی هسته دار (کرنل) شده
۱۳۰	شبکه‌های عصبی (یادگیری ژرف)
۱۴۴	برآوردهای عدم قطعیت از طبقه‌بندها
۱۴۵	تابع تصمیم
۱۴۸	پیش‌بینی احتمالات
۱۵۱	عدم قطعیت در طبقه‌بندی چند کلاسی
۱۵۳	خلاصه و چشم‌انداز

فصل ۳: یادگیری بدون نظارت و پیش‌پردازش

۱۵۷	انواع یادگیری بدون نظارت
۱۵۷	چالش‌های یادگیری بدون نظارت
۱۵۸	پیش‌پردازش و مقیاس‌بندی
۱۵۹	انواع مختلف پیش‌پردازش
۱۶۰	اعمال تبدیلات داده‌ها
۱۶۳	مقیاس‌بندی داده‌های آموزش و ارزیابی به روش مشابه
۱۶۶	تأثیر پیش‌پردازش روی یادگیری با نظارت
۱۶۷	کاهش ابعاد، استخراج ویژگی و یادگیری منیفولد
۱۶۸	تحلیل مؤلفه اصلی (PCA)
۱۸۴	فاکتورگیری ماتریس نامنفی (NMF)
۱۹۱	یادگیری منیفولد به کمک t-SNE
۱۹۶	خوشه‌بندی
۱۹۶	خوشه‌بندی k-میانگین
۲۰۹	خوشه‌بندی تجمعی
۲۱۵	DBSCAN

۲۲۰	مقایسه و ارزیابی الگوریتم‌های خوشه‌بندی
۲۲۶	خلاصه روش‌های خوشه‌بندی
۲۲۷	خلاصه و چشم‌انداز

فصل ۴: نمایش داده‌ها و مهندسی ویژگی‌ها

۲۴۰	متغیرهای دسته‌ای
۲۴۱	کدگذاری one-hot (متغیرهای مجازی)
۲۴۷	اعداد می‌توانند دسته‌ای‌ها را کدگذاری کنند
۲۴۹	بسته‌بندی، گسسته‌سازی، مدل‌های خطی و درخت‌ها
۲۵۳	تعاملات و چندجمله‌ای‌ها
۲۶۲	تبدیلات غیرخطی تک‌متغیره
۲۶۶	انتخاب خودکار ویژگی
۲۶۶	آمار تک‌متغیره
۲۶۹	انتخاب ویژگی مبتنی بر مدل
۲۷۰	انتخاب ویژگی تکراری
۲۷۲	استفاده از دانش تخصصی
۲۸۱	خلاصه و چشم‌انداز

فصل ۵: ارزیابی و بهبود مدل

۲۸۴	اعتبارسنجی متقابل
۲۸۵	اعتبارسنجی متقابل در scikit-learn
۲۸۶	مزایای اعتبارسنجی متقابل
۲۸۷	اعتبارسنجی متقابل k -بخشی طبقه‌بندی شده و سایر استراتژی‌ها
۲۹۴	جستجوی مشبک
۲۹۴	جستجوی مشبک ساده
۲۹۵	خطر بیش‌برازش پارامترها و مجموعه اعتبارسنجی
۲۹۸	جستجوی مشبک با اعتبارسنجی متقابل
۳۱۱	معیارهای ارزیابی و امتیازدهی
۳۱۱	هدف نهایی را در ذهن داشته باشید
۳۱۲	شاخص‌های طبقه‌بندی دودویی
۳۳۵	معیارهایی برای طبقه‌بندی چندکلاسی
۳۳۸	شاخص‌های رگرسیون
۳۳۹	استفاده از شاخص‌های ارزیابی در انتخاب مدل

یادگیری ماشین امروزه به یکی از اجزای لاینفک بسیاری از برنامه‌های کاربردی تجاری و پروژه‌های پژوهشی تبدیل شده است و حوزه‌های کاربرد گسترده‌ای، از تشخیص و درمان پزشکی گرفته تا جستجوی دوستان و آشنایان در شبکه‌های تجاری، پیدا کرده است. بسیاری از افراد تصور می‌کنند که یادگیری ماشین فقط توسط شرکت‌های بزرگی که تیم‌های پژوهشی عریض و طولی دارند، قابل استفاده است. در این کتاب به شما نشان می‌دهیم که به راحتی می‌توانید خودتان راه‌حل‌های یادگیری ماشین را بسازید و با آن کار کنید. به کمک اطلاعات و دانشی که در این کتاب به دست می‌آورید، می‌توانید با سیستم شخصی که برای خود می‌سازید، در تویتر نظرسنجی کنید یا گرمایش جهانی را پیش‌بینی کنید. یادگیری ماشین کاربردهای بیشماری دارد و با توجه به حجم عظیم داده‌های امروز، دیگر نمی‌توان هیچ حد و مرزی برای آن قائل شد.

چه کسانی باید این کتاب را بخوانند

این کتاب برای همه کارشناسان و فعالان علاقه‌مند به یادگیری ماشین است که به دنبال پیاده‌سازی راه‌حل‌ها در مسائل یادگیری ماشین دنیای واقعی هستند. کتاب حاضر مقدماتی است و به همین دلیل نیازمند هیچ دانش و اطلاعات قبلی در مورد یادگیری ماشین یا هوش مصنوعی (AI) نیست. در این کتاب روی استفاده از Python و کتابخانه scikit-learn تمرکز کرده‌ایم و تمام مراحل و گام‌های لازم برای ابداع یک برنامه کاربردی یادگیری ماشین موفق را به تفصیل ذکر می‌کنیم. متدهایی که معرفی می‌کنیم، برای کارشناسان و پژوهشگران و همچنین کارشناسان امور داده که روی برنامه‌های کاربردی تجاری کار می‌کنند، مفید خواهد بود. حتی اگر کمی هم با Python و کتابخانه‌های Numpy و matplotlib آشنایی داشته باشید، بیشترین بهره را از این کتاب خواهید برد.

مؤلفان این کتاب آگاهانه تلاش کرده‌اند به جای تمرکز روی ریاضی، روی جنبه‌های عملی استفاده از الگوریتم‌های یادگیری ماشین تمرکز کنند. از آنجایی که ریاضیات (به ویژه نظریه احتمالات) مبنای ساخت یادگیری ماشین است، تحلیل الگوریتم‌ها را با تفصیل بررسی نکرده‌ایم. چنانچه به مبانی ریاضی الگوریتم‌های یادگیری ماشین علاقه‌مند هستید، توصیه می‌کنیم به کتاب *The Elements of Statistical Learning* (چاپ Springer) به قلم Trevor Hastie, Robert Tibshirani, and Jerome Friedman که به صورت رایگان در وبسایت نویسندگان (به نشانی <http://statweb.stanford.edu/>)

~tibs/ElemStatLearn/) قابل دسترس است، مراجعه کنید. در این کتاب در مورد جزئیات اولیه نحوه نوشتن الگوریتم‌های یادگیری ماشین هم توضیحی نمی‌دهیم و در عوض روی نحوه استفاده از انواع مدل‌هایی که در حال حاضر در scikit-learn و سایر کتابخانه‌ها پیاده‌سازی می‌شوند، تمرکز خواهیم کرد.

چرا تصمیم به نوشتن این کتاب گرفتیم

کتاب‌های بی‌شماری در زمینه یادگیری ماشین و هوش مصنوعی وجود دارد، اما جمعیت هدف تمام آن‌ها دانشجویان کارشناسی ارشد یا دکتری در علوم کامپیوتر است و به مفاهیم و مسائل ریاضیات پیشرفته پرداخته‌اند. در حالی که یادگیری ماشین به‌عنوان یک ابزار مصرفی در برنامه‌های کاربردی تجاری و پژوهشی مطرح شده و به همین دلیل منابع موجود با ذات سادگی و گستردگی این علم در تضاد است. امروزه، برای کاربست یادگیری ماشین دیگر لازم نیست مدرک دکترا داشته باشد. البته منابعی هم وجود دارند که بدون اینکه پای شما را به مباحث ریاضی پیشرفته باز کنند، به تمامی جنبه‌های مهم پیاده‌سازی یادگیری ماشین به صورت عملی و کامل می‌پردازند. امید داریم این کتاب بتواند به افرادی که می‌خواهند بدون اینکه چندین سال را صرف مطالعه جبر، جبر خطی و نظریه احتمالات کنند، از یادگیری ماشین بهره بگیرند، کمک کنند.

کتاب در یک نگاه

ترتیب مطالب کتاب حاضر به این صورت است:

- ✓ **فصل ۱** مفاهیم اساسی یادگیری ماشین و کاربردهای آن را معرفی کرده و نحوه ارائه مطالب در کل کتاب را شرح می‌دهد؛
- ✓ **فصل‌های ۲ و ۳** الگوریتم‌های واقعی یادگیری ماشین را که بیشترین کاربرد را در عمل دارند، شرح داده و مزایا و معایب آن‌ها را مورد بحث قرار می‌دهد؛
- ✓ **فصل ۴** در مورد اهمیت نحوه نمایش داده‌هایی که توسط یادگیری ماشین پردازش می‌شود و جنبه‌هایی از داده‌ها را که باید مورد توجه قرار دهیم، بحث می‌کنیم؛
- ✓ **فصل ۵** متدهای پیشرفته برای ارزیابی مدل و تنظیم پارامتر را با تمرکز خاص روی اعتبارسنجی متقابل و جستجوی مشبک، تحت پوشش قرار می‌دهد؛
- ✓ **فصل ۶** مفهوم خطوط لوله در زنجیر کردن مدل‌ها و نحوه کوچک‌سازی جریان کاری را توضیح می‌دهیم؛
- ✓ **فصل ۷** نحوه اعمال متدهای شرح داده شده در فصل‌های قبل در داده‌های متنی را نشان داده و چند تکنیک پردازش خاص متن را معرفی می‌کند؛

✓ در فصل ۸ جمع‌بندی کامل و جامعی را ارائه داده و منابعی را برای مطالعه موضوعات پیشرفته‌تر معرفی می‌کنیم.

در فصل‌های ۲ و ۳ الگوریتم‌های واقعی را ارائه داده‌ایم، اما ممکن است درک تمام آن‌ها برای مبتدی‌ها لازم هم نباشد. اگر به دلیلی نیاز فوری به ساخت یک سیستم یادگیری ماشین دارید، پیشنهاد می‌کنیم ابتدا به سراغ فصل ۱ و سپس چند سرفصل اول فصل ۲ بروید که تمامی مفاهیم کلیدی را معرفی می‌کند. سپس به سراغ بخش "خلاصه و چشم‌انداز" فصل ۲ بروید که در آن لیستی از تمامی مدل‌های نظارت‌شده مورد بحث را گنجانده‌ایم. مدلی را انتخاب کنید که بیشترین تناسب را با نیازهای شما داشته و به سراغ بخشی بروید که با تفصیل و جزئیات به آن پرداخته شده است. در نهایت هم می‌توانید با استفاده از تکنیک‌های فصل ۵، مدل خود را ارزیابی و تنظیم کنید.

منابع آنلاین

توصیه می‌کنیم در حین مطالعه این کتاب، برای مستندسازی بهتر کلاس‌ها و توابع و مثال‌های بیشتر حتماً به وب‌سایت `scikit-learn` (<http://scikit-learn.org>) مراجعه کنید. به‌علاوه، یک دوره آموزشی ویدیویی توسط آندریاس مولر تحت عنوان "Advanced Machine Learning with `scikit-learn`" به‌عنوان مکمل این کتاب در نشانی http://bit.ly/advanced_machine_learning_scikit-learn گنجانده شده است.

نشانگان قراردادی در کتاب

در این کتاب از قراردادهای چایی زیر استفاده شده است:

ایتالیک

نشانه اصطلاحات جدید، URL‌ها، آدرس‌های ایمیل، نام فایل و پسوند فایل‌ها است.

Calibri (Body)

برای تهیه لیست برنامه‌ها و همچنین در داخل پاراگراف‌ها برای ارجاع به بخش‌های برنامه مثل نام یک متغیر یا تابع، پایگاه‌های داده، نوع داده، متغیرهای محیطی، جملات و کلیدواژه‌ها، فرمان‌ها، ماژول‌ها و نام پکیج‌ها استفاده شده است.

Calibri (Body)

برای اشاره به فرمان‌ها یا هر متنی که باید توسط کاربر تایپ شود، استفاده شده است.