

یادگیری ماشین با پایتون

ریهان گوها

مترجمان:

دکتر کامبیز رهبر

(دانشیار دانشگاه آزاد اسلامی واحد تهران جنوب)

دکتر فاطمه طاهری

www.ketab.ir

سرشناسه : گوها، رهان، Guha, Rehan

عنوان و نام پدیدآور : یادگیری ماشین با پایتون / رهان گوها؛ مترجمان کامبیز رهبر، فاطمه طاهری

مشخصات نشر : تهران: آروین نگار، ۱۴۰۳

مشخصات ظاهری : ۲۹۸ص: مصور

شابک : ۹۷۸-۶۲۲-۵۷۵۶-۵۵-۷

وضعیت فهرست نویسی : فیبا

یادداشت : عنوان اصلی: Machine learning with python...

موضوع : پایتون (زبان برنامه نویسی کامپیوتر)، Python (Computer program language)

فرآگیری ماشینی، Machine learning

شناسه افزوده : رهبر، کامبیز، ۱۳۵۹، مترجم

شناسه افزوده : طاهری، فاطمه، ۱۳۶۳، مترجم

رده بندی کنگره : QA۷۶/۷۳

رده بندی دبویی : ۰۰۵/۱۳۳

شماره کتابشناسی ملی : ۹۶۸۶۶۵۵

این اثر مشمول قانون حمایت مؤلفان و مصنفان و هنرمندان مصوب ۱۳۴۸ است. هرکس تمام یا قسمتی از این اثر را بدون اجازه (ناشر) نشر یا پخش با عرضه کند مورد پیگرد قانونی قرار خواهد گرفت.



تلفن: ۶۶۴۱۸۵۱۲

همراه: ۰۹۳۹۱۲۴۱۶۱۹

عنوان کتاب	یادگیری ماشین با پایتون
مؤلف	رهان گوها
مترجمان	دکتر کامبیز رهبر، دکتر فاطمه طاهری
ناشر	انتشارات آروین نگار
مدیر فنی و ناظر چاپ	محررعلی قاصدی
شابک	۹۷۸-۶۲۲-۵۷۵۶-۵۵-۷
طراح جلد	علی ذوالفقاری
نوبت و سال چاپ	اول / ۱۴۰۳
تیراژ	۲۰ نسخه
قیمت	۳۰۰۰ تومان

مرکز پخش: فروشگاه اینترنتی کتاب آرتین www.artinbook.ir

فروشگاه اینترنتی گنبد نیلگون آسمان www.gnapub.ir

تلفن: ۰۹۱۲۴۱۶۱۹۰۹ - ۶۶۴۸۱۸۷۰

پست الکترونیکی arvinnegarpub@gmail.com

کلیه حقوق این کتاب برای آروین نگار محفوظ است.

پیشگفتار

پیشگفتار با «مقدمه‌ای بر یک کتاب، که معمولاً موضوع، دامنه یا اهداف آن را بیان می‌کند» تعریف می‌شود.

من تلاش می‌کنم در پیشگفتار در قالب پنج سوال به همه جنبه‌ها و زمینه‌هایی که می‌خواهم پوشش دهم پاسخ دهم.

چرا این کتاب را می‌نویسم؟

وقتی به فکر ورود به یادگیری ماشین افتادم، با چند مانع روبرو شدم. یکی از موانع اصلی نحوه شروع یادگیری ماشین بود. من به بسیاری از دوره‌های آنلاین، فیلم‌ها و کتاب‌ها مراجعه کردم، اما در برخی از آنها، مطالب به شدت ریاضی محور است و برخی دیگر فقط کاربردی و بدون توضیح ریاضی بودند. یکی دیگر از موانع اصلی، منابع مرجع بود که برای مبتدیان در نظر گرفته شده بود. آنها به ندرت حاوی ریاضیات است، و توضیحات آن برای بررسی بیشتر آن موضوع بسیار ساده بود. به طور مشابه، برخی از منابع بیش از حد پیچیده بودند، به نحوی که ممکن است شخصی به راحتی بین اطلاعات زیادی که وجود دارد، گم شود.

چه انگیزه‌ای مرا برای نوشتن این کتاب ترغیب کرد؟

از آنجایی که علاقه‌مند به یادگیری ماشین بودم، زمانی که سعی می‌کردم یادگیری ماشینی را به عنوان شغل خود انتخاب کنم، هیچ کتاب یا ویدیوی مناسبی پیدا نکردم که به کسی کمک کند به راحتی وارد حوزه یادگیری ماشین شود. چیزهایی که من در یک منبع خاص به دست نیاوردم، حجم مناسب ریاضی و کدنویسی با پروژه‌های عملی و توضیح اصول اولیه است. دقیقاً همین نکته مرا برانگیخت تا کتابی متعادل بنویسم که می‌تواند مورد استفاده خوانندگان مختلف قرار گیرد و از آن لذت ببرند.

خوانندگان کتاب چه کسانی هستند؟

این کتاب یک کتاب سنتی نیست که در بازار آنلاین یا آفلاین پیدا کنید. این یک کتاب پروژه است و دارای تمام مفاهیم اولیه و همچنین برخی از مفاهیم پیشرفته یادگیری ماشین است که با استفاده از مفاهیم ریاضی کافی اما نه فراگیر معرفی و توضیح داده شده است. با در نظر گرفتن همه موارد، این کتاب می‌تواند برای گروه‌های متعدد خوانندگان مانند علاقه‌مندان به یادگیری ماشین، متخصصان و دانش‌آموزانی که می‌خواهند طعم یادگیری ماشین را در واقعیت بچشند، مفید باشد. جدای از مزایای مشترک کتاب، می‌توان از این کتاب خاص به عنوان راهنمای مرجع نیز استفاده کرد و یا راه‌های حل مسائل را توسط متخصصان یادگیری ماشین مختلف شناخت.

چه چیزی در کتاب خواهید یافت؟

هدف اولیه این نوع کتاب‌ها ارائه مهارت‌های مناسب و زمینه لازم برای ساخت یک مدل یادگیری ماشین است و اینکه مفهوم پیش‌بینی چیست؟ همراه با ساخت یک مدل یادگیری ماشین، اهمیت پیش‌پردازش داده‌ها، ارزیابی مدل و تنظیم مجدد مدل برای به دست آوردن دقت بهتر از خروجی را نیز یاد خواهید گرفت.

این کتاب شامل مثال‌های واقعی است که به همه این تجربه را می‌دهد تا کل مسیر^۱ را برای یک پروژه یادگیری ماشین مدیریت کنند.

در طی پنج فصل کتاب، موارد زیر را خواهید آموخت:

- فصل ۱: مقدمه‌ای در مورد نحوه مدیریت داده‌ها، مبانی چارچوب O.S.E.M.N.، و توضیح اجزای آن ارائه می‌دهد.
- فصل ۲: این فصل به تکنیک‌های مورد استفاده برای کاوش داده‌ها، تجزیه و تحلیل و خرد کردن داده‌ها می‌پردازد.

- فصل ۳: از این فصل به بعد، مفاهیم یادگیری ماشین مانند خوشه‌بندی، رگرسیون و طبقه‌بندی را معرفی شده است.
- فصل ۴: این فصل بر روی مسئله طبقه‌بندی و استنباط از طریق یک نوع مجموعه داده خاص که ما از ویژگی‌های آن آگاه نیستیم، تمرکز دارد.
- فصل ۵: به تنظیم مجدد مدل و بهبود عملکرد مدل برای دستیابی به نتایج بهبود یافته می‌پردازد.

بعد از خواندن کتاب چه تأثیری خواهد داشت؟

این کتاب با روش نوشتن افزایشی نوشته شده است، در این روش مفاهیم به صورت تدریجی نوشته و جمع‌آوری می‌شوند.

اگر کسی کتاب را از ابتدا تا انتها بخواند، آنگاه قادر به انجام تجزیه و تحلیل داده‌ها، به دست آوردن بینش‌های مختلف از داده‌ها، درک مسئله تجاری و ایجاد یک مدل پیش‌بینی و در نهایت تنظیم مدل برای دستیابی به دقت مطلوب خواهد بود.

فهرست

۵	پیشگفتار.....
۵	چرا این کتاب را می‌نویسم؟.....
۵	چه انگیزه‌ای مرا برای نوشتن این کتاب ترغیب کرد؟.....
۶	خوانندگان کتاب چه کسانی هستند؟.....
۶	چه چیزی در کتاب خواهید یافت؟.....
۷	بعد از خواندن کتاب چه تأثیری خواهد داشت؟.....
۱۷	فصل ۱: جنایت بوستون.....
۱۷	مقدمه.....
۱۸	ساختار.....
۱۸	هدف.....
۱۸	داده چیست؟.....
۱۸	انواع داده‌ها.....
۱۹	بیا باید در مورد مجموعه داده بوستون صحبت کنیم.....
۱۹	فرهنگ لغت داده.....
۲۱	چارچوب O.S.E.M.N.....
۲۱	به دست آوردن داده چیست؟.....
۲۲	پاکسازی داده‌ها چیست؟.....
۲۲	یافتن انواع داده‌ها.....
۲۴	چگونه داده‌های از دست رفته را مدیریت کنیم؟.....
۲۷	با مقادیر تکراری چگونه رفتار کنیم؟.....

- ۳۰..... کاوش داده چیست؟
- ۳۲..... تجزیه و تحلیل ویژگی تک متغیره
- ۳۲..... توزیع آماری چیست؟
- ۳۵..... تجزیه و تحلیل ویژگی‌های چند متغیره
- ۴۷..... فصل ۲: گزارش شادی جهانی
- ۴۷..... مقدمه
- ۴۷..... ساختار
- ۴۸..... هدف
- ۴۸..... بیا یاد در مورد مجموعه داده صحبت کنیم
- ۴۸..... پیش نیازها
- ۴۸..... یک جمع‌بندی سریع
- ۵۱..... کاوش داده‌ها
- ۵۱..... انواع مختلف داده
- ۵۱..... داده‌های بازنمایی شده در مقیاس اسمی (دسته‌بندی شده)
- ۵۲..... داده‌های بازنمایی شده در مقیاس ترتیبی
- ۵۲..... اندازه‌گیری داده‌ها
- ۵۲..... متغیرهای گسسته
- ۵۲..... متغیرهای پیوسته
- ۵۳..... مقیاس فاصله
- ۵۳..... مقیاس دایره‌ای
- ۵۳..... مقیاس نسبت
- ۵۳..... توزیع‌ها
- ۵۶..... توزیع نرمال
- ۵۸..... «قاعده تجربی» در آمار چیست؟
- ۵۹..... نمودار توزیع
- ۶۱..... تخمین چگالی کرنل
- ۶۲..... نمودار جعبه‌ای
- ۶۶..... توزیع چوله
- ۶۷..... چوله به چپ (انحراف منفی)

۶۸	"سلامتی (امید به زندگی)" چیست؟
۶۸	نمودار توزیع
۷۰	برآورد چگالی کرنل
۷۲	نمودار جعبه
۷۳	چوله به راست (انحراف مثبت)
۷۳	نمودار توزیع
۷۴	برآورد چگالی کرنل
۷۹	نمودار جعبه
۸۰	سایر توزیع‌ها
۸۳	فصل ۳: گونه‌های زنبق
۸۳	مقدمه
۸۴	ساختار
۸۴	هدف
۸۴	مقدمه‌ای بر یادگیری ماشین
۸۵	یادگیری ماشین چیست؟
۸۶	انواع یادگیری ماشینی
۸۶	یادگیری نظارت شده
۸۶	یادگیری بدون نظارت
۸۶	یادگیری تقویتی
۸۷	خط لوله یادگیری ماشین
۸۸	درک مجموعه داده
۸۹	تجزیه و تحلیل داده‌های اکتشافی
۸۹	بارگذاری مجموعه داده
۹۰	Scikit-Learn
۹۲	Seaborn
۹۳	تحلیل داده‌ها
۹۳	نمودار جفتی
۹۴	مشاهدات
۹۵	همبستگی

۹۹	چگونه الگوریتم‌های ML را انتخاب کنیم؟
۹۹	طبقه‌بندی
۱۰۰	K - نزدیکترین همسایگان (K-nn)
۱۰۱	چگونه کار می‌کند؟
۱۰۲	آموزش مدل
۱۰۳	ارزیابی
۱۰۳	ماتریس سردرگمی چیست؟
۱۰۵	تنظیم مدل
۱۰۷	خوشه‌بندی
۱۰۸	K-means
۱۰۸	چگونه کار می‌کند؟
۱۱۰	آموزش مدل
۱۱۵	ارزیابی و تنظیم مدل
۱۱۵	ایترسی با معیار مجموع مربعات درون خوشه ای
۱۱۵	روش آرنج با استفاده از ایترسی
۱۱۸	امتیاز سیلوئت ^{۳۰}
۱۲۱	رگرسیون
۱۲۴	رگرسیون خطی
۱۲۴	چگونه کار می‌کند؟
۱۲۵	آموزش مدل
۱۳۰	ارزیابی
۱۳۱	میانگین خطای مطلق
۱۳۱	خطای میانگین مربعات
۱۳۱	ریشه میانگین مربعات خطا ^{۴۲}
۱۳۲	امتیاز R^2 یا ضریب تعیین
۱۳۹	فصل ۴: تشخیص تقلب در کارت اعتباری
۱۳۹	مقدمه
۱۴۰	ساختار
۱۴۰	هدف

۱۴۰	بیاید داده‌ها را درک کنیم
۱۴۱	منبع مجموعه داده:
۱۴۱	توضیحات ستون
۱۴۲	مجموعه داده نامتعادل چیست؟
۱۴۴	شناخت ویژگی‌ها
۱۴۸	پیش نیازها
۱۴۸	نرمال‌سازی مقیاس‌بندی ویژگی
۱۵۰	مقیاس‌بندی حداقل/حداکثر ویژگی (مقیاس مجدد)
۱۵۱	نرمال‌سازی میانگین
۱۵۱	استانداردسازی (نرمال‌سازی امتیاز Z-score)
۱۵۱	تجزیه و تحلیل مؤلفه اصلی
۱۵۲	اعتبار سنجی متقابل
۱۵۴	تحلیل داده‌ها
۱۶۶	مقیاس بندی
۱۶۶	مقیاس کننده استاندارد
۱۶۹	مقیاس بندی مقاوم
۱۷۰	ترانسفورماتور قدرت
۱۷۱	ترانسفورماتور کوانتیل
۱۷۶	تقسیم مجموعه داده
۱۷۹	مدیریت عدم تعادل
۱۸۰	نمونه فرعی چیست؟
۱۸۰	نمونه‌گیری مجدد چیست؟
۱۸۳	نمونه‌گیری کم تصادفی
۱۸۷	نمونه‌برداری تصادفی بیش از حد
۱۸۹	پیوندهای توکم برای نمونه‌برداری کم
۱۹۳	تکنیک نمونه‌برداری بیش از حد از اقلیت مصنوعی
۱۹۶	تجزیه و تحلیل مجدد داده‌ها
۱۹۶	ویژگی مثبت (V1..., V28)
۱۹۷	داده‌های آموزشی اصلی

۱۹۷	نمونه‌گیری بیش از حد تصادفی
۱۹۸	نمونه‌گیری بیش از حد تصادفی
۱۹۹	نمونه‌برداری کم پیوند تومک
۱۹۹	SMOTE
۲۰۰	ویژگی منفی (V1.., V28)
۲۰۱	ویژگی‌های مقیاس شده
۲۰۲	ویژگی Time
۲۰۳	ویژگی Amount
۲۰۵	مدل‌سازی
۲۰۵	الگوریتم‌های یادگیری ماشین
۲۰۶	رگرسیون لجستیک
۲۰۶	چرا از رگرسیون لجستیک (مقایسه با رگرسیون خطی) استفاده کنیم؟
۲۰۶	تابع لاجیک چیست؟
۲۰۷	مشتق
۲۰۹	ساختار درخت تصمیم
۲۰۹	الگوریتم
۲۱۰	ریاضیات پشت ASM
۲۱۱	ماشین بردار پشتیبان (SVM)
۲۱۱	چگونه کار می‌کند؟
	ریاضیات پایه در پشت طبقه‌بندی‌کننده حاشیه سخت خطی (SVM خطی) برای موارد
۲۱۴	کاملاً قابل تفکیک:
۲۱۵	آموزش مدل
۲۱۵	آماده‌سازی داده‌ها
۲۱۶	تله متریک
۲۱۷	آموزش و ارزشیابی
۲۳۳	فصل ۵: بیماری قلبی UCI
۲۳۳	مقدمه
۲۳۴	ساختار
۲۳۴	هدف

۲۳۴	پیش نیازها
۲۳۴	چرا یادگیری ماشین در پزشکی بسیار مهم است؟
۲۳۶	هوش مصنوعی قابل توضیح چیست؟
۲۳۷	منظم سازی
۲۳۷	LASSO (حدافل انقباض مطلق و ایراتور انتخاب) ^۲
۲۳۸	رگرسیون ستیغی ^۷
۲۳۹	آزمایش فرضیه
۲۴۱	روش گروهی
۲۴۲	کیسه ای
۲۴۲	تقویت
۲۴۶	تحلیل داده ها
۲۴۷	همبستگی
۲۴۹	ویژگی «ca»
۲۵۲	ویژگی «oldpeak»
۲۵۴	ویژگی «thalach»
۲۵۶	ویژگی «cp»
۲۵۸	تقسیم مجموعه داده
۲۵۹	مدلسازی یادگیری ماشین
۲۵۹	الگوریتم های یادگیری ماشین
۲۵۹	طبقه بندی کننده نایو بیز ^{۳۱ ۳۲}
۲۶۱	جنگل تصادفی
۲۶۳	آموزش مدل
۲۶۴	آموزش و ارزشیابی پایه
۲۶۴	نایو بیز
۲۶۶	روش مجموعه (دسته بندی)
۲۶۹	روش گروهی (تقویت کننده)
۲۷۱	جنگل تصادفی
۲۷۴	انتخاب ویژگی و جستجوی شبکه
۲۷۴	انتخاب ویژگی

۲۷۵	خط لوله
۲۷۸	جستجوی شبکه
۲۸۳	هوش مصنوعی قابل توضیح
۲۸۳	اهمیت ویژگی
۲۸۵	تجسم مدل
۲۸۹	طرح وابستگی جزئی
۲۹۱	SHAP (توضیحات افزودنی شپ)
۲۹۷	منابع