



مؤسسه فرهنگی هنری
دیبانگران تهران

به نام خدا

مباحث نظری علم داده و یادگیری ماشین

در پایتون

با پیاده سازی کامل در محیط پایتون

مؤلف:

آتیا قشقای

(مدرس حوزه علم داده و هوش مصنوعی)



هرگونه چاپ و تکثیر از محتویات این کتاب بدون اجازه کتبی ناشر ممنوع است. متخلفان به موجب قانون حمایت حقوق مؤلفان، مصنفان و هنرمندان تحت پیگرد قانونی قرار می گیرند.

◀ عنوان کتاب: **مباحث نظری علم داده و یادگیری ماشین در پایتون** با پیاده سازی کامل در محیط پایتون

◀ مولف: **آتیلا قشقایی**

◀ ناشر: **مؤسسه فرهنگی هنری دیباگران تهران**

◀ ویراستار: **مهديه مخبري**

◀ صفحه آرایي: **نازنین نصیری**

◀ طراح جلد: **داریوش فرسایي**

◀ نوبت چاپ: **اول**

◀ تاریخ نشر: **۱۴۰۳**

◀ چاپ و صحافی: **ثامن**

◀ تیراژ: **۱۰۰ جلد**

◀ قیمت: **۲۹۸۰۰۰۰ ریال**

◀ شابک: **۹۷۸-۶۲۲-۲۱۸-۸۵۸-۰**

نشانی واحد فروش: تهران، خیابان انقلاب، خیابان دانشگاه

-تقاطع شهدای ژاندارمری-پلاک ۱۵۸ ساختمان دانشگاه-

طبقه دوم-واحد ۴ تلفن ها: ۶۶۹۶۵۷۴۹-۲۲۰۸۵۱۱۱

فروشگاههای اینترنتی دیباگران تهران :

WWW.MFTBOOK.IR

www.dibagartehran.com

سرشناسه: قشقایی، آتیلا، ۱۳۵۸-
عنوان و نام پدیدآور: مباحث نظری علم داده و یادگیری ماشین در پایتون با پیاده سازی کامل در محیط پایتون / مولف: آتیلا قشقایی؛ ویراستار: مهديه مخبري.
مشخصات نشر: تهران: دیباگران تهران: ۱۴۰۳
مشخصات ظاهري: ۲۵۶ ص: جدول، نمودار.
شابک: ۹۷۸-۶۲۲-۲۱۸-۸۵۸-۰
وضعیت فهرست نویسی: فیبا
موضوع: پایتون (زبان برنامه نویسی کامپیوتر)
موضوع: Python (computer program language)
موضوع: فراگیری ماشینی
موضوع: Machine learning
موضوع: داده کاوی Data mining
رده بندی کنگره: ۳۲۵/۵ Q
رده بندی دیویی: ۰۰۶/۳۱
شماره کتابشناسی ملی: ۹۶۷۳۵۰۵

نشانی اینستاگرام دیبا **dibagaran_publishing** نشانی تلگرام: **@mftbook**

هر کتاب دیباگران، یک فرصت جدید علمی و شغلی.

هر گوشی همراه، یک فروشگاه کتاب دیباگران تهران.

از طریق سایتهای دیباگران، در هر جای ایران به کتابهای ما دسترسی دارید.

فهرست مطالب

۷	مقدمه ناشر
۱۰	فصل اول تجزیه و تحلیل داده‌ها
۱۰	عناصر داده‌های ساخت‌یافته
۱۱	داده‌های مستطیلی
۱۳	چارچوب‌ها و شاخص‌های داده
۱۳	ساختارهای داده غیر مستطیلی
۱۵	خلاصه در علم داده
۱۶	تخمین موقعیت مکانی
۱۷	میانگین
۱۷	میانگین پیراسته یا میانگین اصلاحی
۱۸	میانگین وزنی
۱۸	میانه (Median)
۲۰	میانه وزنی (Weighted Mean)
۲۱	مقادیر پرت (Outliers)
۲۱	تخمین تنوع و پراکندگی
۲۱	واریانس σ^2
۲۲	انحراف معیار (STD)
۲۲	تفاوت واریانس و انحراف معیار
۲۳	میانگین قدرمطلق انحراف (Mean absolute Deviation)
۲۴	میانه قدرمطلق انحراف (Median Absolute Deviation - MAD)
۲۴	تشخیص ناهنجاری (Anomaly Detection)
۲۵	ناهنجاری نقطه‌ای (Point Anomaly)
۲۵	ناهنجاری زمینه‌ای (Contextual Anomaly)
۲۶	روش‌های بصری ساده برای تشخیص ناهنجاری
۲۸	روش‌های تشخیص ناهنجاری با استفاده مدل‌ها و توابع
۳۱	Range
۳۲	چندک در علم داده (Quantile)
۳۳	IQR (Interquartile Range) یا فاصله بین چارکی
۳۶	نما (Mode)
۳۶	گشتاور در آمار
۳۷	نمودارهای میله‌ای
۳۸	نمودارهای دایره‌ای

۳۹ نمودارهای خطی
۴۰ نمودارهای پراکندگی
۴۱ نمودارهای جعبه‌ای
۴۲ نمودارهای حرارتی
۴۳ نمودار هیستوگرام (Histogram) در پایتون
۴۴ نمودار هگزین Hexbin Plot
۴۶ همبستگی (Correlation)
۴۷ همبستگی خطی
۴۹ روش تشخیص نوع همبستگی
۵۱ احتمال در آمار
۵۱ مفاهیم کلیدی در احتمال
۵۲ قوانین اساسی احتمال
۵۳ قانون جمع احتمال برای رویدادهای مستقل
۵۴ قانون جمع احتمال برای رویدادهای غیرمستقل
۵۴ امید ریاضی
۵۵ خلاصه

فصل دوم داده‌ها و توزیع‌های نمونه‌گیری ۵۶

۵۶ نمونه‌گیری تصادفی و سوگیری نمونه
۵۶ سوگیری (Bias)
۵۸ انتخاب تصادفی
۵۹ اندازه در مقابل کیفیت
۶۰ چه زمانی به مقادیر انبوه داده نیاز است؟
۶۰ آشنایی با توزیع نمونه‌ای
۶۱ تفاوت توزیع نمونه با توزیع جمعیت
۶۲ خطای استاندارد (Standard Error)
۶۲ تفاوت خطای استاندارد و انحراف استاندارد
۶۳ محاسبه خطای استاندارد با یک نمونه یا چند نمونه
۶۵ بوت استرپ (Bootstrapping) چیست؟
۶۸ فاصله اطمینان (Confidence Interval - CI)
۷۲ نمودار چندک-چندک Q-Q-PLOT
۷۵ درجه آزادی (df) چیست؟
۷۷ چولگی (Skewness)
۷۷ کشیدگی (Kurtosis)
۷۸ توزیع نرمال
۸۴ توزیع یونیفرم (یکنواخت)

۸۶	توزیع برنولی
۸۸	توزیع چند جمله‌ای (مالتی نومینال)
۹۰	توزیع آماری نوع t - (t-Distribution)
۹۴	توزیع دو جمله‌ای
۹۵	تابع جرم احتمال PMF
۹۶	تابع توزیع تجمعی CDF
۹۷	تابع چگالی احتمال (PDF)
۹۹	توزیع نمایی (Exponential Distribution)
۱۰۴	توزیع پواسون
۱۰۷	توزیع خی دو (Chi-Square)

فصل سوم آزمایش‌های آماری و آزمون فرض آماری ۱۰۹

۱۱۱	آزمون آماری و آزمون فرض آماری
۱۱۲	آزمون‌های فرضیه
۱۱۳	فرضیه صفر
۱۱۳	نمونه‌گیری مجدد
۱۱۳	تست A/B
۱۱۷	تست جایگشت - جایجایی (Permutation Test)
۱۲۴	آزمون t
۱۲۸	آزمون خی دو (Chi-Square)
۱۳۲	آزمون فیشر - F
۱۳۷	آزمون پواسون
۱۴۱	یادگیری ماشین کلاسیک چیست؟
۱۴۱	انواع یادگیری ماشین
۱۴۱	۱. یادگیری تحت نظارت (Supervised Learning)
۱۴۲	۲. یادگیری بدون نظارت (Unsupervised Learning)
۱۴۳	۳. یادگیری تقویتی (Reinforcement Learning)
۱۴۴	آینده یادگیری ماشین
۱۴۶	زمینه‌های نوظهور
۱۴۶	نتیجه‌گیری

فصل چهارم رگرسیون و پیش‌بینی ۱۴۷

۱۴۹	رگرسیون خطی ساده (Simple Linear Regression)
۱۵۲	رگرسیون خطی چندگانه (Multiple Linear Regression)
۱۵۴	رگرسیون ریج (Ridge Regression)
۱۵۷	رگرسیون الاستیک نت (Elastic Net Regression)

۱۵۹.....	رگرسیون پولینومیال (Polynomial Regression)
۱۶۱.....	مدل‌های رگرسیون براساس فاکتور محبوبیت
۱۶۳.....	معیارهای بررسی عملکرد در مدل‌های رگرسیون
۱۶۳.....	کاربردهای متریک‌های مدل رگرسیون و انواع آن
۱۶۷.....	محبوبیت و کاربرد هر یک از متریک‌ها

فصل پنجم طبقه‌بندی و پیش‌بینی ۱۷۰

۱۷۰.....	رگرسیون لجستیک (Logistic Regression)
۱۷۴.....	الگوریتم K نزدیک‌ترین همسایه (K-Nearest Neighbors)
۱۸۱.....	درخت تصمیم DTR (Decision Tree Model)
۱۸۴.....	مدل جنگل تصمیم (Random Forest)
۱۸۸.....	ماشین بردار پشتیبان (SVM)
۱۹۲.....	مدل SVR برای مسائل رگرسیون
۱۹۳.....	معیارهای ارزیابی مدل‌های طبقه‌بندی در یادگیری ماشین
۱۹۸.....	محبوبیت و کاربرد هر یک از معیارهای طبقه‌بندی

فصل ششم الگوریتم‌های خوشه‌بندی ۲۰۰

۲۰۲.....	الگوریتم K-Means
۲۰۶.....	خوشه‌بندی سلسله‌مراتبی (Hierarchical Clustering)
۲۱۰.....	الگوریتم DBSCAN (خوشه‌بندی فضایی مبتنی بر تراکم یا نویز)
۲۱۵.....	الگوریتم OPTICS (شناسایی ساختار خوشه‌بندی با مرتب‌سازی نقاط)
۲۱۷.....	مدل پنهان مارکوف (HMM) چیست؟
۲۲۰.....	معیارهای ارزیابی دقت (Metrics) مدل‌های خوشه‌بندی

فصل هفتم مدل‌های کاهش بعد (Dimensionality Reduction) ۲۲۵

۲۲۸.....	مدل کاهش بعد PCA (تجزیه و تحلیل مؤلفه اصلی)
۲۳۱.....	مدل کاهش بعد ICA (تجزیه و تحلیل مؤلفه مستقل)
۲۳۸.....	مدل کاهش بعد RP (Random Projection)

فصل هشتم مدل‌های دسته‌بندی (Ensemble Methods) ۲۴۰

۲۴۵.....	مدل جنگل تصادفی (Random Forest) این بار برای دسته‌بندی
۲۴۹.....	مدل بوستینگ (Boosting)
۲۵۱.....	الگوریتم پایلینگ (Stacking)
۲۵۳.....	معیارهای متریک برای مدل‌های دسته‌بندی
۲۵۵.....	معیارهای متداول برای مدل‌های دسته‌بندی به ترتیب محبوبیت

خط‌مشی انتشارات مؤسسه فرهنگی هنری دیباگران تهران در عرصه کتاب‌هایی با کیفیت عالی است که تواند
خواسته‌های به روز جامعه فرهنگی و علمی کشور را تا حد امکان پوشش دهد.
هر کتاب دیباگران تهران، یک فرصت جدید شغلی و علمی

حمد و سپاس ایزد منان را که با الطاف بی‌کران خود این توفیق را به ما ارزانی داشت تا بتوانیم در راه ارتقای دانش عمومی و فرهنگی این مرز و بوم در زمینه چاپ و نشر کتب علمی و آموزشی گام‌هایی هرچند کوچک برداشته و در انجام رسالتی که بر عهده داریم، مؤثر واقع شویم.

گسترده‌گی علوم و سرعت توسعه روزافزون آن، شرایطی را به وجود آورده که هر روز شاهد تحولات اساسی چشمگیری در سطح جهان هستیم. این گسترش و توسعه، نیاز به منابع مختلف از جمله کتاب را به عنوان قدیمی‌ترین و راحت‌ترین راه دستیابی به اطلاعات و اطلاع‌رسانی، بیش از پیش برجسته نموده است.

در این راستا، واحد انتشارات مؤسسه فرهنگی هنری دیباگران تهران با همکاری اساتید، مؤلفان، مترجمان، متخصصان، پژوهشگران و محققان در زمینه‌های گوناگون و مورد نیاز جامعه تلاش نموده برای رفع کمبودها و نیازهای موجود، منابعی پُر بار، معتبر و با کیفیت مناسب در اختیار علاقمندان قرار دهد.

کتابی که در دست دارید تألیف "جناب آقای آتیلا قشقایی" است که با تلاش همکاران ما در نشر دیباگران تهران منتشر گشته و شایسته است از یکایک این گرامیان تشکر و قدردانی کنیم.

با نظرات خود مشوق و راهنمای ما باشید

با ارائه نظرات و پیشنهادات و خواسته‌های خود، به ما کمک کنید تا بهتر و دقیق‌تر در جهت رفع نیازهای علمی و آموزشی کشورمان قدم برداریم. برای رساندن پیام‌هایتان به ما از رسانه‌های دیباگران تهران شامل سایتهای فروشگاهی و صفحه اینستاگرام و شماره‌های تماس که در صفحه شناسنامه کتاب آمده استفاده نمایید.

مدیر انتشارات

مؤسسه فرهنگی هنری دیباگران تهران
dibagaran@mftplus.com

بنام حضرت دوست

تقدیم به روان پاک پدرم

در ابتدا باید تشکر کنم از **مادر عزیزم** که مشوق من برای نگارش این کتاب بود، کاری که یکسال از وقت من را گرفت ولی خروجی آن باعث می‌شود دانشجویان حوزه علم داده و یادگیری ماشین پس از اتمام دوره‌های لازم در این گرایش‌ها بتوانند با تکمیل کردن دانششان در این علوم، آماده رویارویی با مسائل واقعی در این حوزه‌ها گردند و اینگونه نباشد که فقط دانش ابزاری در این زمینه را یدک بکشند و درکی از مسئله و رویکرد حل آن نداشته باشند.

مطالبی که در این کتاب بیان شده است مسلماً خارج از وقت و حوصله دوره‌های تخصصی این گرایش‌ها می‌باشند چون آنقدر در این حوزه‌ها ابزار فراوانند که وقتی برای بیان علوم پایه آن نمی‌ماند و عموماً کسب این دانش به عهده دانشجو گذارده می‌شود و یا حتی انتظار می‌رود که این دانش را دانشجو از قبل همراه داشته باشد.

مخاطب این کتاب

مخاطب این کتاب دانشجویان پایتون و علاقه‌مندان علم داده می‌باشند که اولاً به زبان پایتون آشنا بوده و همچنین با کتابخانه‌های مورد نیاز آنالیز داده و یادگیری ماشین در پایتون مانند `Numpy`، `Pandas`، `Sklearn`، `Matplotlib` آشنایی حداقلی را داشته باشند و در این مرحله می‌خواهند آن فضاهای خالی علمی میان این ابزارها تا نقطه انجام یک پروژه واقعی را پر کنند همان بخش‌هایی که در هیچ کلاس استاندارد علم داده‌ای به آنها پرداخته نخواهد شد و آن را پوشش نخواهد داد این کتاب تمام مفاهیم مهم و کاربردی علم آمار برای علم داده را همراه با نحوه پیاده‌سازی آنها در زبان پایتون را به شما عرضه می‌دارد و شما را قادر می‌سازد درکی درست از مفاهیم و اصطلاحات علم آمار برای علم داده بدست آورید، همچنین نگاهی کاربردی به انواع الگوریتم‌های یادگیری ماشین دارد و آنها را از بُعد حل مسئله مورد بررسی و طبقه‌بندی قرار می‌دهد.

در این کتاب برعکس اکثر کتاب‌های این حوزه که چند سکویی می‌باشند و هر قسمت از یک مسئله را با یک نرم‌افزار یا زبانی خاص بیان می‌کنند، تک سکویی بوده و فقط از پایتون برای پیاده‌سازی تمامی این مفاهیم استفاده می‌نماید که البته من به‌شخصه این را بزرگترین نقطه قوت این کتاب می‌دانم.

در نهایت این کتاب را باید به‌عنوان یک مرجع نگاه کرد نه صرفاً یک کتاب آموزشی و اکیداً توصیه می‌شود اگر به این علوم علاقه‌مندید آن را کنار دست خود داشته باشید.

همراه کتاب فایل‌ها و دیتاست‌های مربوطه نیز وجود دارد که می‌توانید از آدرس مخزن عمومی اینجانب در گیت‌هاب دانلود نمایید.

<https://github.com/atila1358/DS-ML>

در پایان تشکر می‌کنم از جناب مهندس فرسای مدیریت انتشارات دیباگران تهران و تیم سخت‌کوش ایشان که با وجود شرایط بسیار بد در حوزه نشر هنوز هم به جهت اعتلای دانش این سرزمین تلاش می‌کنند و امیدوارم در این مسیر همچون گذشته پرفروغ باشند.

Site: www.poyeshmashin.ir

Email: atila.gh@gmail.com

با طلب خیر

آتिला قشقایی

۱۴۰۳