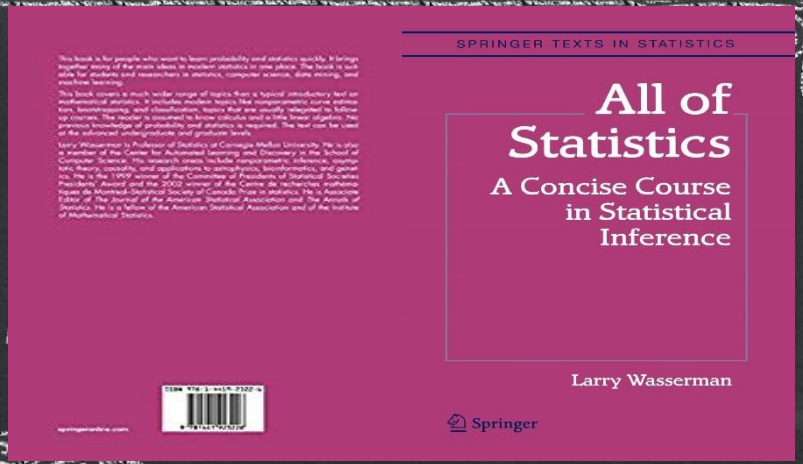


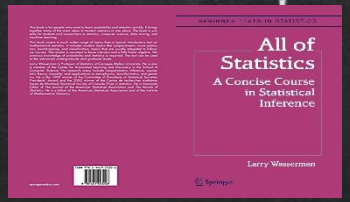
مقدمه کتاب All of Statistics

ارائه دهنده: پرهام پیشرو / ۴۰۱۱۳۰۹۰۱۲

استاد: دکتر وحید فکور

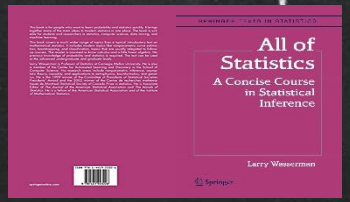
Taken literally, the title “All of Statistics” is an exaggeration. But in spirit, the title is apt, as the book does cover a much broader range of topics than a typical introductory book on mathematical statistics.

به معنای واقعی کلمه، عنوان "همه آمار" اغراق آمیز است. اما در باطن، عنوانی مناسب است، زیرا کتاب طیف گسترده تری از موضوعات را نسبت به یک کتاب مقدماتی معمولی در مورد آمار ریاضی پوشش می دهد.



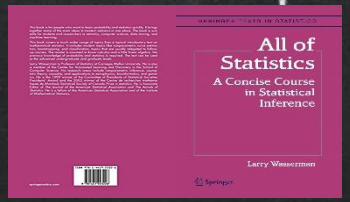
This book is for people who want to learn probability and statistics quickly. It is suitable for graduate or advanced undergraduate students in computer science, mathematics, statistics, and related disciplines. The book includes modern topics like nonparametric curve estimation, bootstrapping, and classification, topics that are usually relegated to follow-up courses. The reader is presumed to know calculus and little linear algebra. No previous knowledge of probability and statistics is required.

این کتاب برای افرادی است که می خواهند آمار و احتمالات را به سرعت یاد بگیرند. برای دانشجویان مقطع کارشناسی ارشد یا دانشجویان پیشرفته کارشناسی در رشته های علوم کامپیوتر، ریاضی، آمار و دیگر رشته های مرتبط مناسب است. این کتاب شامل موضوعات مدرنی مانند برآورد منحنی ناپارامتری، خودگردان سازی و طبقه بندی است، موضوعاتی که معمولاً به دوره های بعدی منتقل می شوند. فرض بر این است که خواننده حساب دیفرانسیل و انتگرال و کمی جبر خطی را می داند. هیچ دانش قبلی از آمار و احتمال مورد نیاز نیست.



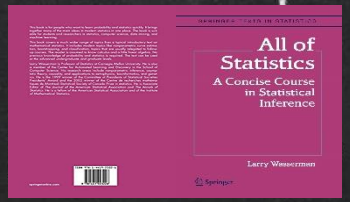
Statistics, data mining, and **machine learning** are all concerned with collecting and analyzing data. For some times, statistics research was conducted in statistics departments while data mining and machine learning research was conducted in computer science departments. Statisticians thought that computer scientists were reinventing the wheel. Computer scientist thought that statistical theory didn't apply to their problems.

آمار، داده کاوی و **یادگیری ماشین** همگی با جمع آوری و تجزیه و تحلیل داده ها سروکار دارند. برای مدتی، تحقیقات آماری در بخش های آماری انجام می شد در حالی که داده کاوی و یادگیری ماشین در بخش های علوم کامپیوتر انجام می شد. آماردانان فکر می کردند که متخصصان کامپیوتر در حال اختراع مجدد چرخ هستند. متخصصان کامپیوتر نیز فکر می کردند که نظریه آماری برای مشکلات آن ها کاربرد و راه حلی ندارد.



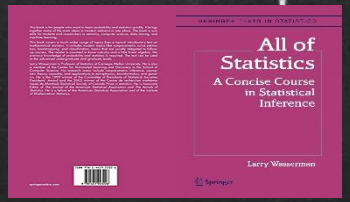
Things are changing. Statisticians now recognize that computer scientist are making novel contributions while computer scientist now recognize the generality of statistical theory and methodology. Clever data mining algorithms are more scalable than statisticians ever thought possible. Formal statistical theory is more pervasive than computer scientists had realized.

اوضاع در حال تغییر است. آماردانان اکنون انجام مشارکت های جدید متخصصان کامپیوتر را به رسمیت می شناسند و متخصصان کامپیوتر نیز اکنون کلیت نظریه و روش آماری را به رسمیت می شناسند. الگوریتم های هوشمند داده کاوی مقیاس پذیرتر از آن چیزی هستند که آماردانان تصور می کنند. نظریه آماری رسمی فراگیرتر (جامع تر) از آن چیزی است که متخصصان کامپیوتر تصور می کردند.



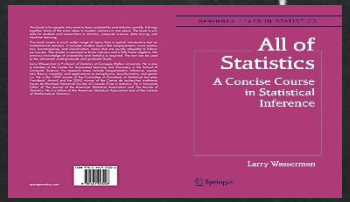
Students who analyze data, or who aspire to develop new methods for analyzing data, should be well grounded in basic probability and mathematical statistics. Using fancy tools like neural nets, boosting, and support vector machines without understanding basic statistics is like doing brain surgery before knowing how to use a band-aid.

دانشجویانی که داده ها را تجزیه و تحلیل می کنند، یا می خواهند روش های جدیدی برای تجزیه و تحلیل داده ها خلق کنند، باید به پایه و مبانی احتمالات و آمار ریاضی مسلط باشند. استفاده از ابزارهایی مانند شبکه های عصبی، تقویت کردن و ماشین بردار پشتیبان بدون درک پایه آمار مانند انجام عمل جراحی مغز قبل از دانستن نحوه استفاده از چسب زخم است.



But where can students learn basic probability and statistics quickly? Nowhere. At least, that was my conclusion when my computer science colleagues kept asking me: “Where Can I send my students to get a good understanding of modern statistics quickly?” The typical mathematical statistics course spends too much time on tedious and uninspiring topics (counting methods, two dimensional integrals, etc.) at the expense of covering modern concepts (bootstrapping, curve estimation, graphical models, etc.).

اما دانشجویان از کجا می توانند آمار و احتمالات مقدماتی را به سرعت یاد بگیرند؟ هیچ جا. حداقل، زمانی به این نتیجه رسیدم که همکاران متخصص کامپیوترم همیشه از من می پرسیدند: «دانشجویانم را کجا می توانم بفرستم تا به سرعت درک خوبی از آمارهای مدرن به دست آورند؟»دوره معمولی آمارِ ریاضی زمان زیادی را صرف موضوعات خسته کننده و غیر الهام بخش (روش های شمارش، انتگرال ها ی دو بعدی و غیره) به جای پوشش مفاهیم مدرن (بوت استرپ، برآورد منحنی، مدل های گرافیکی و غیره) می کند.

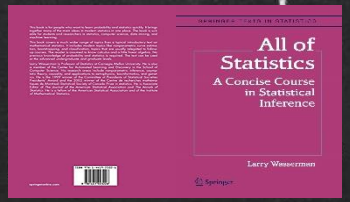


So I set out to redesign our undergraduate honors course on probability and mathematical statistics. This book arose from that course. Here is a summary of the main features of this book.

1. The book is suitable for graduate students in computer science and honors undergraduates in math, statistics, and computer science. It is also useful for students beginning graduate work in statistics who need to fill in their background on mathematical statistics.

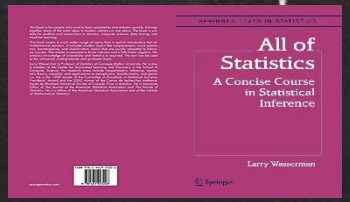
بنابراین من تصمیم گرفتم که دوره کارشناسی افتخاری خود را در مورد احتمالات و آمار ریاضی بازطراحی کنم. این کتاب برخواسته از همان دوره است. در اینجا خلاصه ای از ویژگی های اصلی این کتاب آورده شده است.

۱. کتاب برای دانشجویان کارشناسی ارشد علوم کامپیوتر و دانشجویان کارشناسی افتخاری ریاضی، آمار و علوم کامپیوتر مناسب است. همچنین برای دانشجویانی که شروع به انجام کارهای فارغ التحصیلی می کنند و نیاز به پر کردن پیشینه خودشان در آمار ریاضی دارند، مفید است.



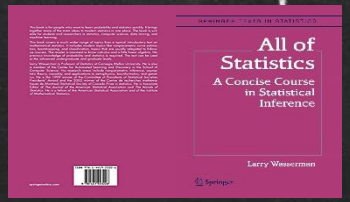
2. I cover advanced topics that are traditionally not taught in a first course. For example, nonparametric regressions, bootstrapping, density estimation, and graphical models.
3. I have omitted topics in probability that do not play a central role in statistical inference. For example, counting methods are virtually absent.
4. Whenever possible, I avoid tedious calculations in favor of emphasizing concepts.
5. I cover nonparametric inference before parametric inference.

۲. من موضوعات پیشرفته ای که معمولاً در دوره اول تدریس نمی شوند را پوشش می دهم. به عنوان مثال، رگرسیون ناپارامتری، بوت استرپینگ، برآورد چگالی و مدل های گرافیکی.
۳. من موضوعاتی در احتمال که نقش محوری در استنباط آماری نداشتند را حذف کردم. به عنوان مثال، روش های شمارش عملاً وجود ندارند.
۴. در صورت امکان، از محاسبات خسته کننده به نفع تاکید بر مفاهیم اجتناب می کنم.
۵. من استنباط ناپارامتری را قبل از استنباط پارامتری پوشش می دهم.



6. I abandon the usual “First Term = Probability” and “Second Term = Statistics” approach. Some students only take the first half and it would be a crime if they did not see any statistical theory. Furthermore, probability is more engaging when students can see it put to work in the context of statistics.An exception is the topic of stochastic processes which is included in the later material.

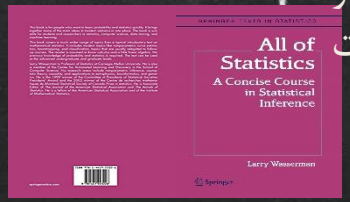
۶. من رویکرد معمول "ترم اول = احتمال" و "ترم دوم = آمار" را کنار می گذارم. برخی از دانشجویان فقط نیمه اول را می گذرانند و اگر هیچ نظریه آماری را نبینند، همانند این است که به خودشان ظلم کرده باشند. علاوه بر این، وقتی که دانشجویان عملکرد احتمال را در چارچوب آمار می بینند، احتمال جذاب تر می شود. یک استثنا، موضوع فرآیندهای تصادفی است که در مطالب بعدی گنجانده شده است.



7. The course moves very quickly and covers much material.My colleagues joke that I cover all of statistics in this course and hence the title.The course is demanding but I have worked hard to make the material as intuitive as possible so that the material is very understandable despite the fast pace.
8. Rigor and clarity are not synonymous. I have tried to strike a good balance. To avoid getting bogged down in uninteresting technical details, many results are stated without proof. The bibliographic references at the end of each chapter point the student to appropriate sources.

۷. دوره بسیار سریع حرکت می کند و مطالب زیادی را پوشش می دهد. همکاران من به شوخی می گویند که من تمام آمارهای لازم در این دوره و بنابراین عنوان کتاب را پوشش می دهم. دوره مشکل است اما من سخت کار کرده ام تا مطالب را در حد امکان شهودی کنم، تا با وجود سرعت بالا، مطالب بسیار قابل درک باشند.

۸. دقت و وضوح مترداف نیستند. من تلاش کرده ام که تعادل خوبی بین این دو برقرار کنم. برای جلو گیری از گرفتار شدن درجزئیات فنی غیر جالب، بسیاری از نتایج بدون اثبات بیان می شوند. قسمت کتاب شناسی در پایان هر فصل، دانشجو را به سمت منابع مناسب راهنمایی می کند.



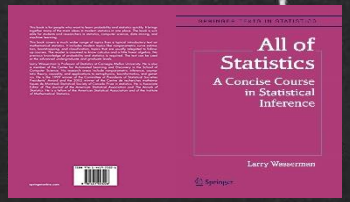
9. On my website are files with R code which students can use for doing all the computing. The website is:

http://www.stat.cmu.edu./~larry/all-of-statistics

However, the book is not tied to R and any computing language can be used.

۹. در وب سایت من فایل هایی به زبان R وجود دارد که دانشجویان می توانند برای انجام تمام محاسبات از آن استفاده کنند.

به هر حال، کتاب به R وابسته نیست و می توان از هر زبانی برای انجام محاسبات استفاده کرد.

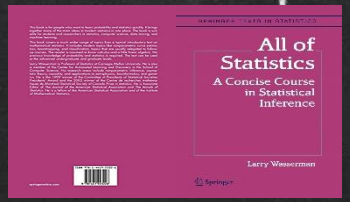


Part I of the text is concerned with probability theory, the formal language of uncertainty which is the basis of statistical inference. The basic problem that we study in probability is:

> Given a data generating process, what are the properties of the outcomes?

بخش اول متن به نظریه احتمالات مربوط می شود، زبان رسمی عدم قطعیت که پایه استنباط آماری است. مسئله اساسی که در احتمال مطالعه می کنیم این است:

با توجه به فرآیند تولید داده، خواص نتایج چیست؟

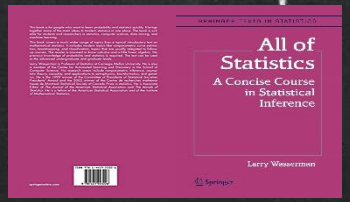


Part II is about statistical inference and its close cousins, data mining and machine learning. The basic problem of statistical inference is the inverse of probability:

> Given the outcomes, what can we say about the process that generated the data?

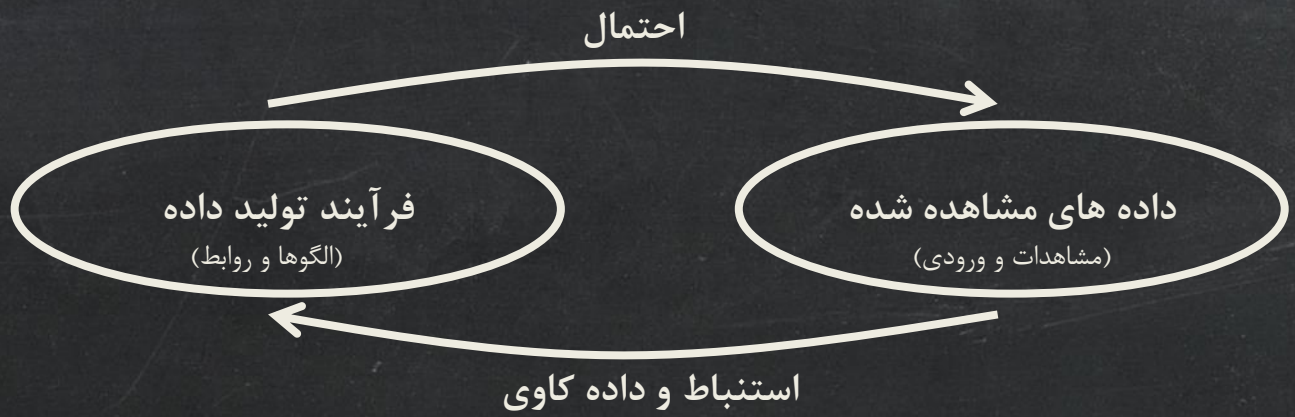
بخش دوم درباره استنباط آماری و پسرعموهای نزدیک آن، داده کاوی و یادگیری ماشین است.
مسئله اساسی در استنباط آماری، برعکس احتمال، این است:

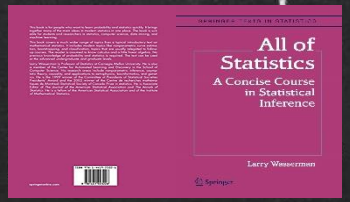
با توجه به نتایج، در مورد فرآیند تولید داده ها چه می توانیم بگوییم؟



These ideas are illustrated in Figure 1. Prediction, classification, clustering, and estimation are all special cases of statistical inference. Data analysis, machine learning and data mining are various names given to the practice of statistical inference, depending on the context.

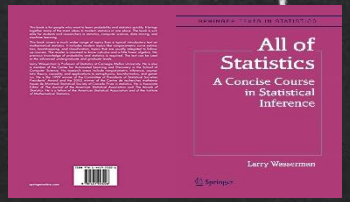
این ایده ها در شکل۱ (شکل زیر) نمایش داده شده اند. پیش بینی، طبقه بندی، خوشه بندی و برآورد همگی موارد خاصی از استنباط آماری هستند. تجزیه و تحلیل داده ها، یادگیری ماشین و داده کاوی نام های مختلفی هستند که با توجه به متن، به عمل استنباط آماری داده می شود.





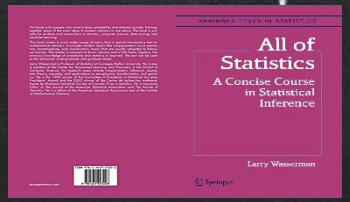
Part III applies the ideas from Part II to specific problems such as regression, graphical models, causation, density estimation, smoothing, classification, and simulation.Part III contains one more chapter on probability that covers stochastic processes including Markov chains.

بخش سوم ایده های قسمت دوم را برای مسائل خاص رگرسیون، مدل های گرافیکی، علیت، برآورد چگالی، هموارسازی، طبقه بندی و شبیه سازی استفاده می کند. بخش سوم شامل یک فصل دیگر در مورد احتمال است که فرآیندهای تصادفی شامل زنجیرهای مارکوف را پوشش می دهد.



I have drawn on other books in many places. Most chapters contain a section called Bibliographic Remarks which serves both to acknowledge my debt to other authors and to point readers to other useful references. I would especially like to mention the books by DeGroot and Schervish (2002) and Grimmett and Stirzaker (1982) from which I adapted many examples and exercises.

من در بسیاری از جاها از کتاب های دیگر استفاده کرده ام. بیشتر فصل ها شامل بخشی به نام تذکرات کتاب شناسی هستند که هم برای اظهار نام سایر نویسندگان و هم برای راهنمایی خوانندگان به سایر مراجع مفید است. من به صورت ویژه تمایل دارم که به کتاب های DeGroot و Schervish (۲۰۰۲) و همچنین Grimmett و Stirzaker (۱۹۸۲) اشاره کنم که مثال ها و تمرین های زیادی را از آن ها اقتباس کردم.

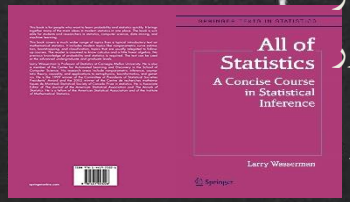


As one develops a book over several years it is easy to lose track of where presentation ideas and, especially, homework problems originated. Some I made up. Some I remembered from my education. Some I borrowed from other books. I hope I do not offend anyone if I have used a problem from their book and failed to give proper credit.As my colleague Mark Schervish wrote in his book (Schervish (1995)),

> "…the problems at the ends of each chapter have come from many sources. …These problems, in turn, came from various sources unknown to me … if I have used a problem without giving proper credit, please take it as a compliment."

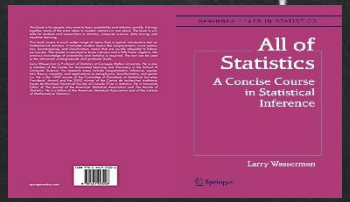
وقتی فردی کتابی را طی چندین سال توسعه می دهد، به راحتی می توان ایده های ارائه را پیدا و مشکلات نشات گرفته از تکالیف را مفقود کرد. بعضی از آن ها را درست کردم. بعضی از آن ها را از دوران تحصیلم به یاد آوردم.بعضی را از کتاب های دیگر به امانت گرفتم.امیدوارم اگر از مسئله ای در کتاب دیگران استفاده کرده ام و اعتبار (ذکر نام) مناسبی ارائه نداده ام، به کسی توهین نکنم. همانطور که همکار من Mark Schervish در کتاب خود (Schervish (1995) نوشت:

« مسائل پایان هر فصل ممکن است از کتاب هایی باشند که برای من ناشناخته است. »



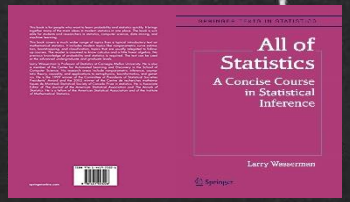
I am indebted to many people without whose help I could not have written this book. First and foremost, the many students who used earlier versions of this text and provided much feedback. In particular, Liz Prather and Jennifer Bakal read the book carefully. Rob Reeder valiantly read through the entire book in excruciating detail and gave me countless suggestions for improvements. Chris Genovese deserves special mention. He not only provided helpful ideas about intellectual content, but also spent many,many hours writing LaTeX code for the book.

من مدیون افراد زیادی هستم که بدون کمک آن ها نمی توانستم این کتاب را بنویسم؛اول و مهم تر از همه، خیلی از دانشجویانی که از نسخه های قبلی این متن استفاده و بازخوردهای زیادی ارائه کردند. به طور خاص، Liz Prather و Jennifer Bakal که کتاب را به دقت مطالعه کردند. Rob Reeder هم با شجاعت، تمام کتاب را با جزئیات طاقت فرسایش خواند و پیشنهاداتی بی شمار برای بهبود کتاب به من داد.



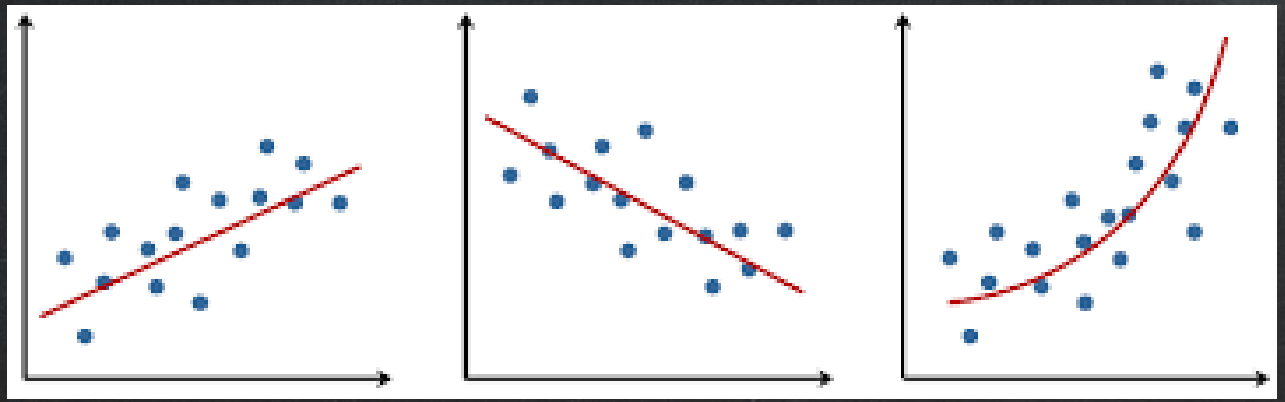
Chris Genovese deserves special mention. He not only provided helpful ideas about intellectual content, but also spent many, many hours writing LaTeX code for the book. The best aspects of the book's layout are due to his hard work; any stylistic deficiencies are due to my lack of expertise. David Hand,Sam Roweis, and David Scott read the book very carefully and made numerous suggestions that greatly improved the book. John Kimmel has been supportive and helpful throughout the writing process. Finally, my wife Isabella Verdinelli has been an invaluable source of love, support, and inspiration.

Chris Genovese شایسته ی یک نام بردن ویژه و مخصوص است؛ او نه تنها ایده های مفیدی در مورد محتوای ذهنی ارائه کرد،بلکه ساعات خیلی خیلی زیادی را صرف نوشتن کدلاتک در کتاب کرد. بهترین جنبه های صفحه آرایی کتاب مدیون سخت کوشی اوست؛ هر گونه کم و کاستی در سبک به دلیل عدم تخصص من است. David Hand، Sam roweis و David Scott کتاب را با دقت مطالعه کردند و پیشنهادات متعددی ارائه کردند که کتاب را بسیار بهبود بخشید. John Kimmel در طول فرآیند نگاش حامی و کمک کننده بوده است. در نهایت، همسرم Isabella Verdinelli منبع ارزشمندی از عشق، حمایت و الهام بوده است.



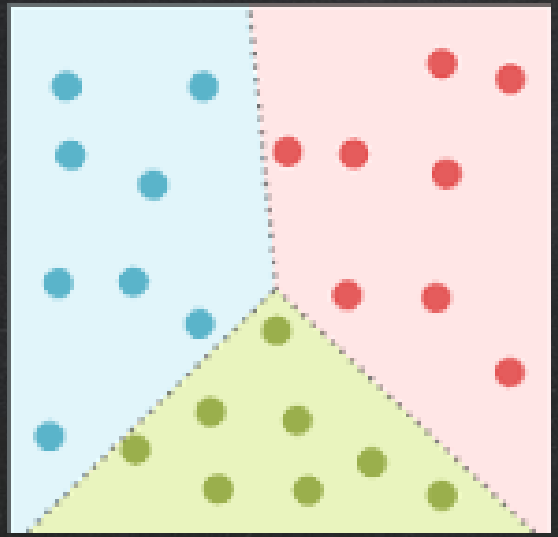
Statistics/Data Mining Dictionary

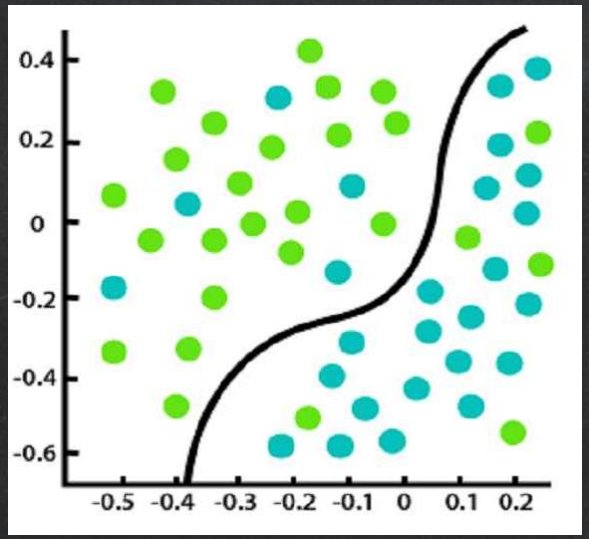
Statistics	Computer Science	Meaning
estimation برآورد	learning یادگیری	Using data to estimate an unknown quantity استفاده از داده جهت برآورد (تخمین) کمیت های ناشناخته



Statistics/Data Mining Dictionary

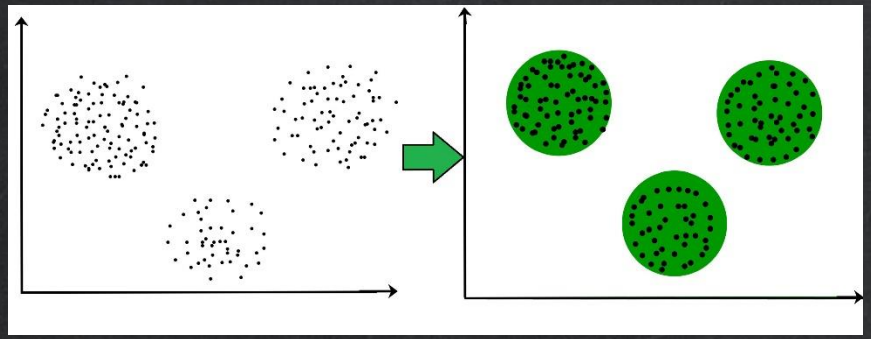
Statistics	Computer Science	Meaning
classification طبقه بندی	supervised learning یادگیری با نظارت	predicting a discrete Y from X پیش بینی یک Y مجزا از روی X

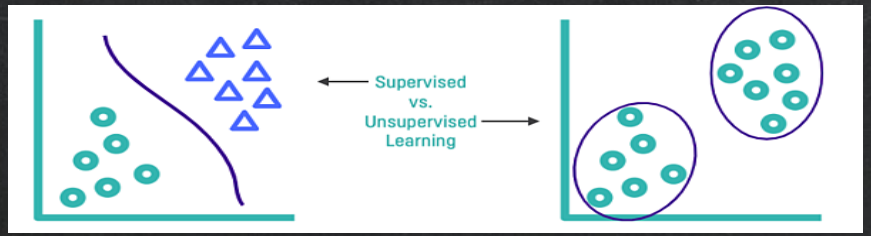


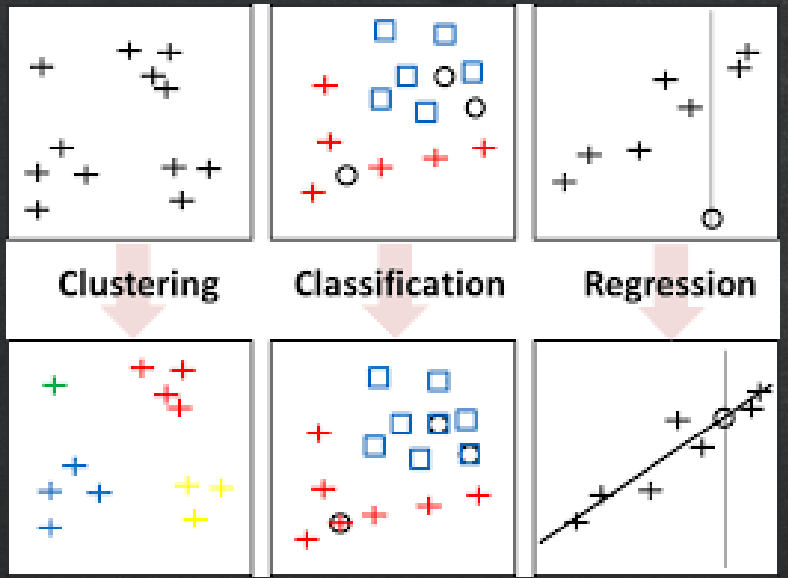


Statistics/Data Mining Dictionary

Statistics	Computer Science	Meaning
clustering خوشه بندی	unsupervised learning یادگیری بدون نظارت	putting data into groups قرار دادن داده ها در گروه های مختلف







Statistics/Data Mining Dictionary

Statistics	Computer Science	Meaning
data داده	training sample نمونه ی آموزشی (تمرینی)	$(X_1, Y_1), \ldots, (X_n, Y_n)$
covariate متغیر تصادفی	feature بردار ویژگی	the X_i's

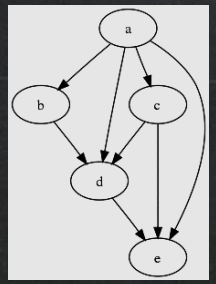
	Number	Male	Female	M/T	F/T	M/F	Ave_Age	Prev_Ed	Ave_Exp	Sum_Exp	Ave_Fear	Hint	Ans	Plague	Time	Turn	Victory
2	4	2	2	0/5	0/5	0	26/75	16	0/5	2	62/5	6	2	0	17	1	87
3	5	3	2	0/6	0/4	0/2	31/2	16	1/4	7	77	7	4	0	19	2	82
4	7	7	0	1	0	1	18/71	12	2/57	18	80/71	10	3	0	21	3	81
5	7	4	3	0/57	0/43	0/14	21/86	16	2/14	15	45	6	2	0	17	1	102
6	5	3	2	0/6	0/4	0/2	26/8	21	3/6	18	66	3	0	0	21	3	111
7	4	1	3	0/25	0/75	-0/5	22/25	12	0	0	28/75	6	1	0	21	3	70
8	6	4	2	0/67	0/33	0/33	31/83	16	0/5	3	62/5	4	0	0	19	2	91
9	7	4	3	0/57	0/43	0/14	22/29	14	0/43	3	62/86	5	3	0	21	3	88
10	8	5	3	0/625	0/375	0/25	17/5	12	0/25	2	64/38	5	1	0	19	2	80
11	5	2	3	0/4	0/6	-0/2	24/4	16	6	30	81	6	1	0	21	3	91
12	4	4	0	1	0	1	29/5	21	2	8	82/25	4	1	0	17	1	102
13	4	3	1	0/75	0/25	0/5	15/75	12	1/75	7	59/25	9	4	0	19	2	108
14	4	4	0	1	0	1	20/75	12	0	0	47/75	6	0	0	21	3	61
15	5	2	3	0/4	0/6	-0/2	19/4	12	0/2	1	79/6	7	3	0	17	1	84
16	5	3	2	0/6	0/4	0/2	29/8	16	2/4	12	36/8	5	2	0	19	2	84
17	5	3	2	0/6	0/4	0/2	24	14	0	0	71/8	11	3	0	21	3	81

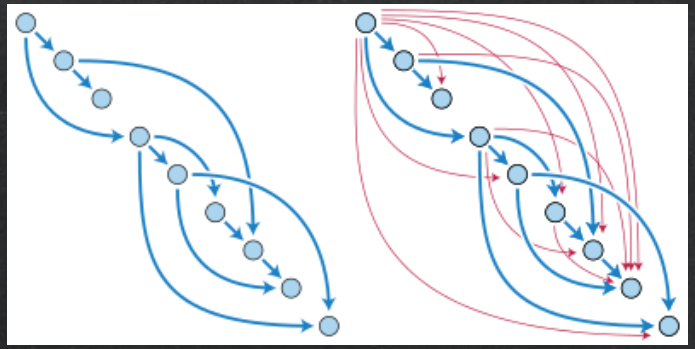
Statistics/Data Mining Dictionary

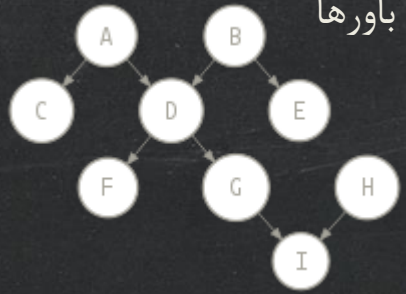
Statistics	Computer Science	Meaning
classifier طبقه بندی کننده	hypothesis فرض، فرضیه	a map from covariates to outcomes نقشه ای از متغیرها به نتایج
hypothesis فرض، فرضیه	-----	subset of a parameter space Θ زیرمجموعه ای از فضای پارامتر Θ
confidence interval فاصله اطمینان	-----	interval that contains an unknown quantity with given frequency بازه ای که حاوی یک کمیت ناشناخته با درصد اطمینان مشخصی است

Statistics/Data Mining Dictionary

Statistics	Computer Science	Meaning
directed acyclic graph گراف بدون دورِ جهت دار	Bayes net شبکه بیزی	multivariate distribution with given conditional independence relations توزیع چندمتغیره با روابط استقلال شرطی داده شده
Bayesian inference استنباط بیزی	Bayesian inference استنباط بیزی	statistical methods for using data to update beliefs روش های آماری برای استفاده از داده ها جهت به روز رسانی باورها







Statistics/Data Mining Dictionary

Statistics	Computer Science	Meaning
Frequentist inference استنباط فراوانی گرا	-----	statistical methods with guaranteed frequency behavior روش های آماری با رفتار فرکانسی تضمین شده
large deviation bounds کران انحراف بزرگ	PAC learning یادگیری احتمالا تقریبا درست	uniform bounds on probability of errors کران های یکنواخت احتمال خطاها

با تشکر از توجه شما