Introduction to Machine Learning

Jiangjiao Xu

University of Exeter, Penryn Campus, UK

March 2020

UNIVERSITY OF EXETER | DOCTORAL COLLEGE

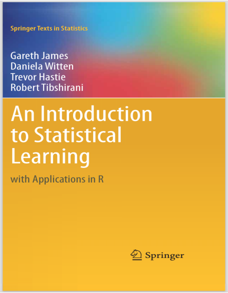
Researcher Development

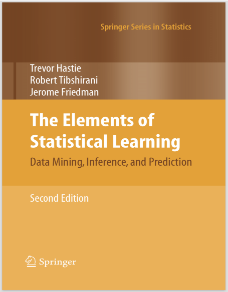
Workshop learning outcomes

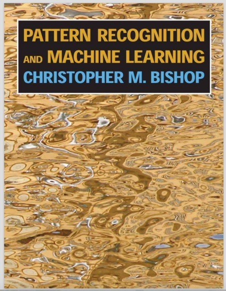
- Understand the key concepts and terminology used in the field of machine learning
- Build predictive models for clustering and classification problems
- Apply machine learning algorithms in Python/R to a variety of real-world datasets
- Recognise practical issues in data-driven modelling

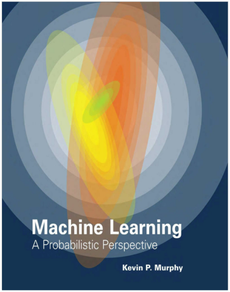
Note: All workshop material can be found here:
https://exeter-data-analytics.github.io/

Recommended reading









Overview

- Why are we here?
- What is machine learning?
- Types of machine learning methods
- Statistics vs Machine Learning
- Terminology
- A bird's-eye view of machine learning

Why are we here?

The infamous **Big Data**!

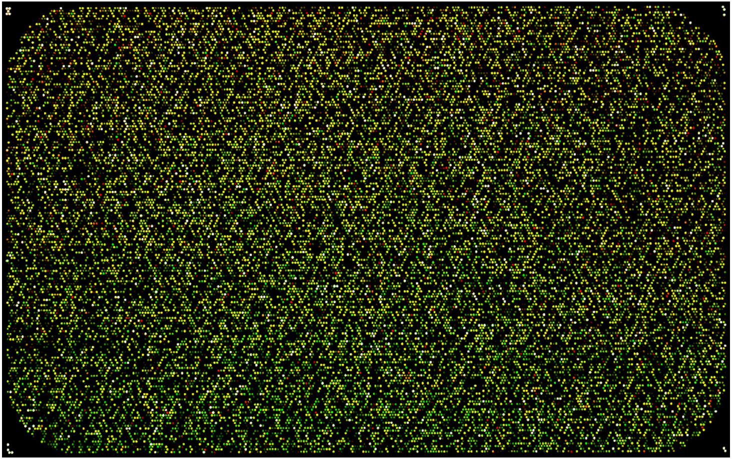
class	BrdIndx	Area	Round	Bright	Compact	ShpIndx	Mean_G	Mean_R	Mean_NIR	SD_G	SD_R	SD_NIR	LW	GLCM1	Rect	GLCM2	Dens	Assym	NDVI	BordLngth	GLCM3
car	1.27	91	0.97	231.38	1.39	1.47	207.92	241.74	244.48	21.41	20.4	18.69	2.19	0.48	0.87	6.23	1.6	0.74	-0.08	56	4219.69
concrete	2.36	241	1.56	216.15	2.46	2.51	187.85	229.39	231.2	6.57	6.97	7.02	1.52	0.45	0.63	6.35	1.55	0.69	-0.1	156	3682.08
concrete	2.12	266	1.47	232.18	2.07	2.21	206.54	244.22	245.79	6.16	4.93	5.53	1.14	0.52	0.78	6.19	1.88	0.19	-0.08	144	2943.35
concrete	2.42	399	1.28	230.4	2.49	2.73	204.6	243.27	243.32	5.76	5.56	5.46	2.51	0.5	0.74	6.28	1.51	0.83	-0.09	218	3550.19
concrete	2.15	944	1.73	193.18	2.28	4.1	165.98	205.55	208	11.46	8.9	9.77	12.26	0.71	0.59	7.36	0.63	0.99	-0.11	504	2300.41
tree	3.11	169	1.47	172.22	2.49	3.35	240.18	127.65	148.83	8.41	10.34	11.5	1.87	0.43	0.55	6.44	1.38	0.7	0.31	174	4707.12
car	1.2	44	0.79	208.8	1.14	1.36	180.95	221.61	223.82	35.42	36.45	35.17	2	0.54	0.89	5.84	1.58	0.63	-0.1	36	4340.51
car	1	88	0.22	234.51	1.11	1.12	208.5	246.48	248.56	12.3	11.29	8.81	2	0.44	0.93	6.08	1.78	0.6	-0.08	42	4213.84
building	1.59	1737	0.67	219.61	1.3	1.64	185.86	233.84	239.13	7.08	7.03	7.28	1.49	0.47	0.9	6.51	2.15	0.51	-0.11	274	2637.73
tree	2.37	153	1.3	120.24	2.85	2.59	184.15	81.5	95.06	10.21	11.99	11.3	2.05	0.56	0.64	6.62	1.57	0.5	0.39	128	4474.26
building	1.57	3552	0.46	213.22	1.32	1.6	173.03	229.84	236.8	5.5	5.54	5.63	1.4	0.11	0.94	6.24	2.3	0.31	-0.14	382	3756.06
asphalt	4.19	418	2.48	83.35	4.21	4.3	68.07	88.8	93.17	6.51	6.3	6.12	1.1	0.46	0.44	6.52	1.31	0.26	-0.13	352	3711.08
building	1.3	1024	0.58	183.93	1.39	1.64	169.57	209.61	172.6	6.15	6.37	6.8	3.95	0.38	0.89	6.46	1.63	0.86	-0.11	210	3703.39
grass	1.14	289	0.38	173.16	1.21	1.21	213.71	145.56	160.23	10.3	11.55	11.1	1.79	0.54	0.91	6.77	2.02	0.52	0.19	82	2642.33
shadow	2.03	249	1	39.62	2.2	2.12	35.86	38.92	44.07	7.72	6.36	6.87	1.35	0.63	0.76	6.42	1.92	0.3	-0.04	134	2542.16
building	1.31	1639	0.68	205.31	1.26	1.37	196.24	239.05	180.65	7.15	7.82	8.37	1.69	0.49	0.9	6.53	2.07	0.56	-0.1	222	2675
tree	2.68	285	1.48	138.01	2.53	2.7	201.84	98.06	114.13	15.24	15.1	14.31	1.01	0.7	0.63	7.28	1.75	0.15	0.35	182	2323.79
soil	2.79	293	1.37	237.11	2.4	3.1	227.22	242.45	241.68	7.9	9	9.86	2.17	0.45	0.67	6.66	1.5	0.73	-0.03	212	3689.82
building	1.21	2797	0.78	244.7	1.34	1.23	229.52	252.21	252.37	7.5	4.98	4.42	1.24	0.48	0.85	6.64	2.2	0.4	-0.05	260	2766.23
shadow	1.13	217	0.32	41.25	1.22	1.22	47.51	34.69	41.55	8.79	8.29	7.77	2.18	0.48	0.92	6.41	1.84	0.68	0.16	72	3132.21
pool	1.28	173	0.85	157.34	1.53	1.41	85.43	161.11	225.49	14.8	12.18	16.76	2.18	0.77	0.76	6.64	1.8	0.65	-0.31	74	1700.07
shadow	3.23	203	2.15	66.75	3.55	3.4	60.3	68.47	71.47	14.52	16.39	14.8	1.46	0.52	0.31	6.83	1.25	0.47	-0.06	194	4149.89
concrete	2.09	178	1.63	215.76	2.4	2.66	188.69	229.03	229.55	5.25	4.95	5.01	3.55	0.54	0.59	6.07	1.06	0.94	-0.1	142	3928.82
tree	2	138	1.31	154.97	2.46	2.04	200.16	124.49	140.25	13.52	13	12.97	1.18	0.58	0.73	6.74	1.7	0.45	0.23	96	3503.29
grass	1.79	203	1.83	152.09	2.41	2.46	193.34	123.78	139.15	14.03	13.18	10.17	4.9	0.56	0.67	6.91	0.96	0.96	0.22	140	2627.62
concrete	2	792	1	238.34	1.88	2.24	211.08	251.96	251.97	6.43	3.73	3.78	2.49	0.51	0.84	6.4	1.69	0.81	-0.09	252	2772.6
grass	3.54	429	1.79	176.66	2.73	3.93	205.01	160.12	164.84	7.52	7.29	7.02	2.28	0.57	0.58	6.58	1.33	0.8	0.12	326	2794.4
building	1.38	1078	0.64	243.74	1.27	1.43	226.39	252.62	252.21	7.17	3.78	4.43	1.42	0.48	0.9	6.58	2.17	0.37	-0.05	188	3001.58
building	1.92	957	0.96	175.33	1.76	2.05	161.9	201.82	162.26	7.79	7.24	6.78	2	0.51	0.8	6.67	1.92	0.62	-0.11	254	2880.27
building	2.74	316	1.92	200.53	2.73	3.32	181.09	208.44	212.05	7.5	8.4	8.22	3.38	0.5	0.57	6.49	1.04	0.93	-0.07	236	3511.27
asphalt	2.94	642	1.43	84.16	2.56	3.02	67.13	90.65	94.7	7.06	7.08	7.15	1.34	0.55	0.66	6.56	1.82	0.33	-0.15	306	2756.09
shadow	1.39	220	0.86	49.89	1.43	1.79	61.68	40.14	47.87	9.51	7.03	7.09	3.89	0.53	0.84	6.54	1.35	0.92	0.21	106	3269.51
building	1.48	3084	0.93	230.71	1.33	2.52	215.62	252.64	223.88	7.04	3.65	7.07	9.33	0.53	0.89	6.65	1	0.98	-0.08	560	2851
grass	2.44	554	1.72	146.12	3.03	2.49	213.73	102.01	122.61	8.67	6.99	7.24	1.05	0.65	0.65	6.7	1.72	0.37	0.35	234	2153.61
tree	2.89	288	1.41	124.98	2.4	2.98	187.27	88.61	99.07	15.77	13.45	13.66	1.3	0.69	0.62	7.25	1.76	0.12	0.36	202	2300.9
building	1.4	1278	0.73	226.77	1.42	2.06	204.75	251.55	224.01	7.03	4.58	7.13	6.29	0.46	0.91	6.61	1.22	0.96	-0.1	294	3145.16
asphalt	2.66	998	1.34	71.58	2.35	2.82	57.98	75.54	81.23	8.51	8.88	8.99	1.82	0.49	0.73	6.88	1.77	0.66	-0.13	356	3476.61
tree	1.66	328	0.9	237.5	1.73	1.74	211.47	250.57	250.48	6.5	4.95	4.99	1.6	0.47	0.83	6.38	1.94	0.54	-0.08	126	3312.05
tree	3.2	373	1.61	148.7	2.78	3.73	213.41	106.93	125.76	13.31	10.07	10.96	2.58	0.68	0.57	7.33	1.3	0.81	0.33	288	2575.69
concrete	1.41	1046	0.76	239.94	1.37	1.9	215.11	252.47	252.25	6.2	3.72	3.71	4.94	0.36	0.9	6.4	1.39	0.93	-0.08	246	3420.48
soil	2.26	397	1.63	236.16	3.43	2.38	233.62	238.42	236.43	8.41	10.71	11.52	1.62	0.47	0.69	6.75	1.66	0.45	-0.01	190	3241.68
tree	1.96	144	0.98	112.87	2.31	2.13	168.78	78.85	90.97	11.45	9.01	8.27	1.88	0.69	0.72	6.54	1.65	0.58	0.36	102	2639.3
building	1.51	869	1.17	189.67	1.56	1.71	164.41	208.08	196.5	7.45	8.08	7.53	2.57	0.54	0.78	6.66	1.57	0.84	-0.12	202	2768.26
building	1.87	3645	0.75	236.75	1.55	1.97	204.59	252.82	252.83	5.72	3.22	3.29	1.78	0.24	0.86	6.32	2.05	0.62	-0.11	476	3634.8
building	1.55	2306	0.74	211.82	1.29	1.71	185.81	232.73	216.93	6.91	6.28	6.21	2.43	0.42	0.9	6.55	2	0.62	-0.11	328	3081.92
grass	1.83	183	0.83	144.61	1.97	1.96	156.36	139.63	137.83	7.98	7.32	7.61	1.6	0.46	0.78	6.52	1.75	0.62	0.06	106	3499
grass	3.72	557	2.13	152.42	3.58	3.94	166.87	145.44	144.96	7.65	7.94	7.7	1.79	0.58	0.55	6.74	1.39	0.52	0.07	372	3096.14
shadow	1.72	663	0.43	45.19	1.7	1.81	47.1	41.36	47.11	9.06	7.43	7.24	1.55	0.7	0.87	6.72	2.13	0.43	0.06	186	2044.52
soil	2.23	224	1.23	239.87	2.24	2.31	225.22	247.39	247	6.67	6.04	6.51	1.41	0.42	0.63	6.3	1.64	0.57	-0.05	138	3484.04

training.csv

Normal View | Ready | Sum=0

Why are we here?

The infamous **Big Data**!



Water, water everywhere, nor any drop to drink

© marketoonist.com

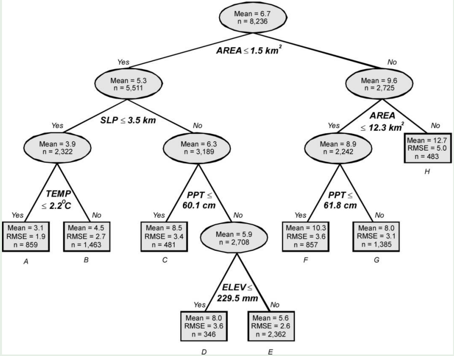
Water, water everywhere, nor any drop to drink

Predicting fish species richness

Olden *et al.* Q Rev Biol 2008, 83(2):171-193

Data: Lake surface area, shoreline perimeter, air temperature, precipitation and elevation

Method: Decision trees (supervised)



What's the problem?

Detection of malarial parasites

Purwar *et al.* Malar J 2011, 10:364

Data: Image intensity

Method: Modified k-means clustering (unsupervised)

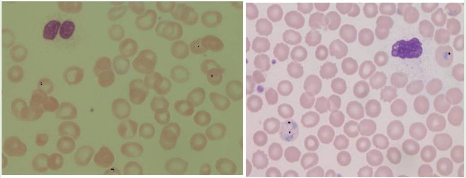
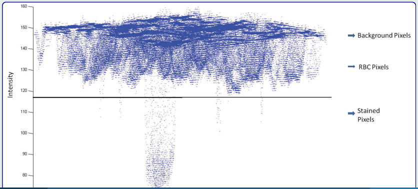


Figure 17 **Parasites marked image.**



Creating carbon-density maps

Baccini *et al.* Nature Clim. Change 2012, 2:182-185

Data: Light detection and ranging (LiDAR) (elevation data)

Method: Random forests (ensemble of decision trees) (supervised)

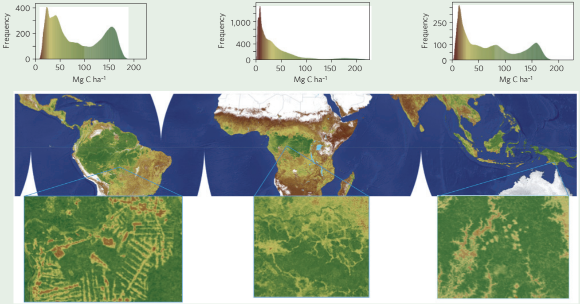


Figure 1 | Carbon contained in the aboveground live woody vegetation of tropical America, Africa and Asia (Australia excluded). The upper panels show the frequency distribution of carbon in units of Mg C ha^{-1} for each region. Inset figures across the bottom provide higher-resolution examples of the spatial detail present in the satellite-derived biomass data set. Carbon amount is represented in the maps as a colour scheme from dark brown (low carbon) to dark green (high carbon). See upper panels for numeric values.

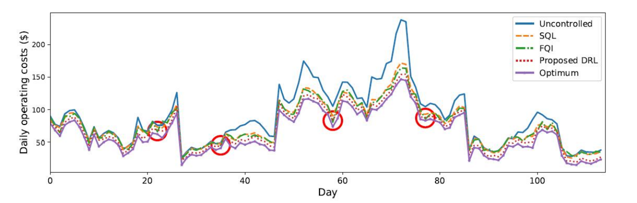
What's the problem?

Real-Time Energy Management of a Microgrid Using Deep Reinforcement Learning

Ji *et al.* Energies 2019, 12(12):2291-2312

Data: State of the Micro Grid

Method: Deep reinforcement learning (supervised)



What my mum thinks machine learning is

Who uses machine learning?

Who uses machine learning?

Machine Learning in Ecosystem Informatics and Sustainability

Thomas G. Dietterich
School of Electrical Engineering and Computer Science
Oregon State University
tgd@cs.orst.edu

MACHINE LEARNING IN THE LIFE SCIENCES

©BRAND X, PHOTODISC

Machine Learning in the Life Sciences

How it is Used on a Wide Variety of Medical Problems and Data

KRZYSZTOF J. CIOS, LUKASZ A. KURGAN, AND MAREK REFORMAT

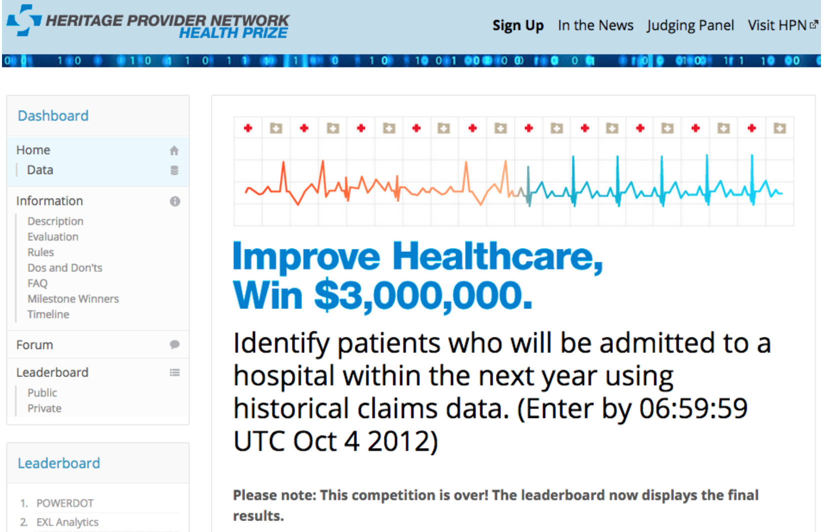
VOLUME 83, NO. 2 *THE QUARTERLY REVIEW OF BIOLOGY* JUNE 2008

MACHINE LEARNING METHODS WITHOUT TEARS: A PRIMER FOR ECOLOGISTS

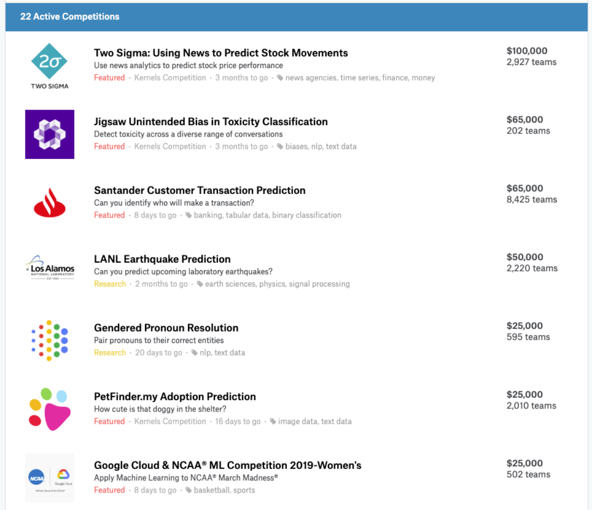
Data Analysis and Mining in the Life Sciences

Nam Huyn
SurroMed, Inc.
2375 Garcia Ave, Mountain View, CA 94043, USA
phuyn@surromed.com

There are even lucrative competitions!



Lots of them actually...



Source: http://www.kaggle.com/

So what is machine learning?

So what is machine learning?

A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E

— Tom Mitchell

So what is machine learning?

A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E

— Tom Mitchell

The scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead

— Wikipedia

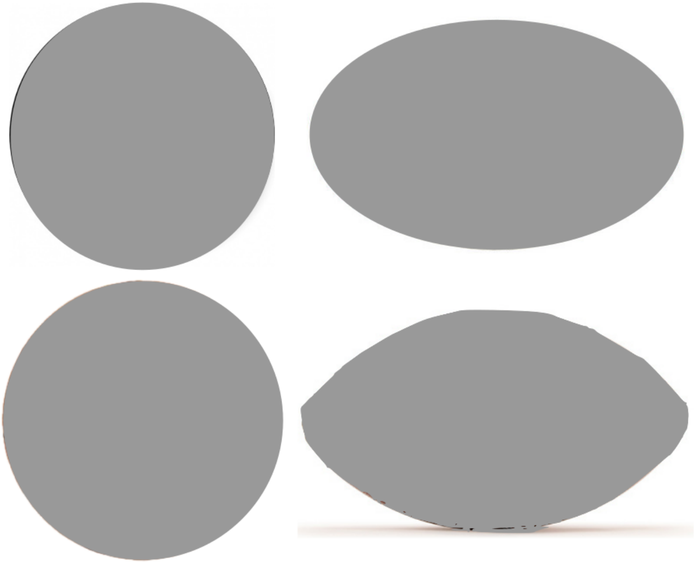
So what is machine learning?

> A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E
>
> — Tom Mitchell

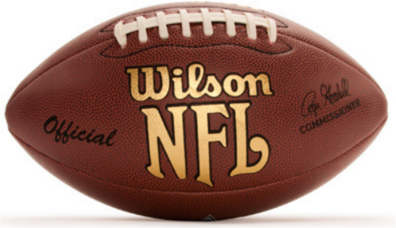
> The scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead
>
> — Wikipedia

Machines learn using flashcards

Group by shape (unsupervised learning)



Add labels (supervised learning)



Types of machine learning methods: Unsupervised learning

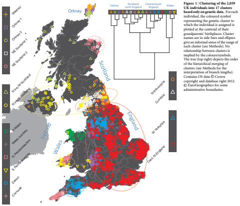
Inputs have *no* corresponding output labels

- **Clustering** - discovering groups having similar attributes
- **Density Estimation** - determine the distribution of data
- **Dimensionality Reduction** - identify and remove redundant dimensions

Types of machine learning methods: Unsupervised learning

Inputs have *no* corresponding output labels

- **Clustering** - discovering groups having similar attributes



- **Density Estimation** - determine the distribution of data

- **Dimensionality Reduction** - identify and remove redundant dimensions

Types of machine learning methods: Unsupervised learning

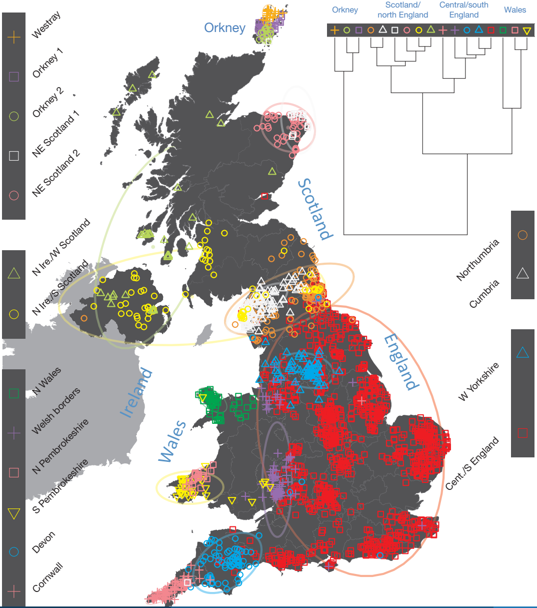
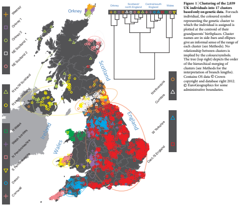


Figure 1 | Clustering of the 2,039 UK individuals into 17 clusters based only on genetic data. For each individual, the coloured symbol representing the genetic cluster to which the individual is assigned is plotted at the centroid of their grandparents' birthplaces. Cluster names are in side-bars and ellipses give an informal sense of the range of each cluster (see Methods). No relationship between clusters is implied by the colours/symbols. The tree (top right) depicts the order of the hierarchical merging of clusters (see Methods for the interpretation of branch lengths). Contains OS data © Crown copyright and database right 2012. © EuroGeographics for some administrative boundaries.

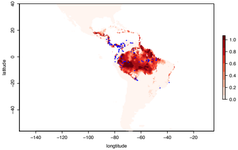
Types of machine learning methods: Unsupervised learning

Inputs have *no* corresponding output labels

- **Clustering** - discovering groups having similar attributes

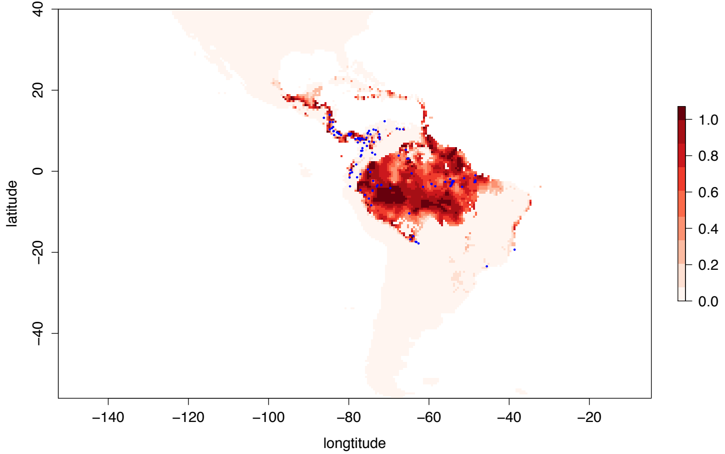


- **Density Estimation** - determine the distribution of data



- **Dimensionality Reduction** - identify and remove redundant dimensions

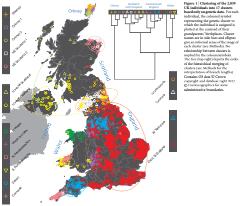
Types of machine learning methods: Unsupervised learning



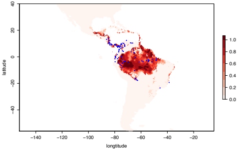
Types of machine learning methods: Unsupervised learning

Inputs have *no* corresponding output labels

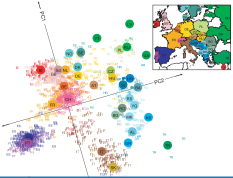
- **Clustering** - discovering groups having similar attributes



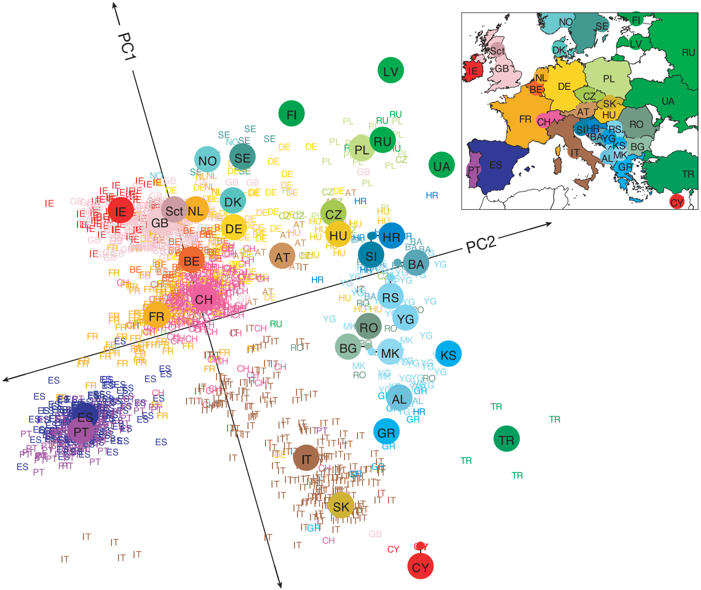
- **Density Estimation** - determine the distribution of data



- **Dimensionality Reduction** - identify and remove redundant dimensions



Types of machine learning methods: Unsupervised learning



Types of machine learning methods: Supervised learning

Inputs have corresponding output labels

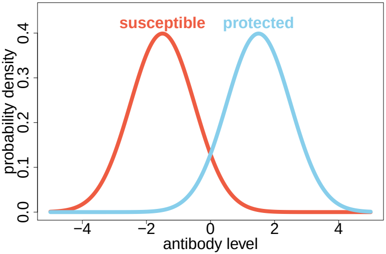
- **Classification** - output is categorical

- **Regression** - output is continuous

Types of machine learning methods: Supervised learning

Inputs have corresponding output labels

- **Classification** - output is categorical

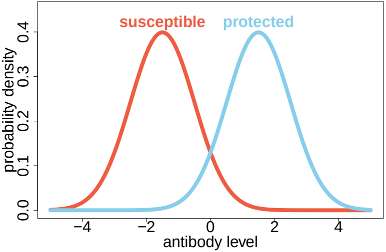


- **Regression** - output is continuous

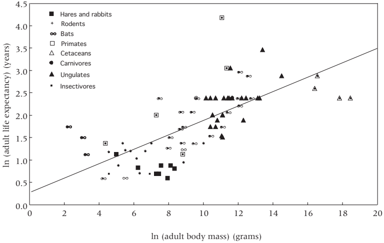
Types of machine learning methods: Supervised learning

Inputs have corresponding output labels

- **Classification** - output is categorical



- **Regression** - output is continuous

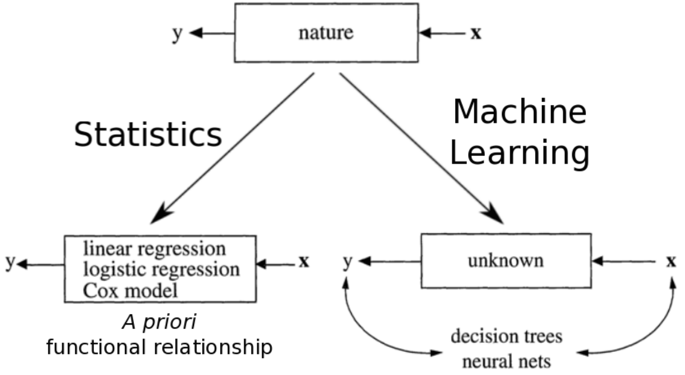


Statistics vs Machine Learning (not mutually exclusive)

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman



Statistics vs Machine Learning (not mutually exclusive)

Statistics

- **Philosophy** - provide humans a set of data analysis tools
- **Focus** - what is the relationship between the data and the outcome?
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2, chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2, Cp statistic, AIC)

Machine Learning

- **Philosophy** - replace humans in the processing of data
- **Focus** - how can we predict the outcome using the data?
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning (not mutually exclusive)

Statistics

- **Philosophy** - provide humans a set of data analysis tools
- **Focus** - what is the relationship between the data and the outcome?
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2, chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2, Cp statistic, AIC)

Machine Learning

- **Philosophy** - replace humans in the processing of data
- **Focus** - how can we predict the outcome using the data?
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning (not mutually exclusive)

Statistics

- **Philosophy** - provide humans a set of data analysis tools
- **Focus** - what is the relationship between the data and the outcome?
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2, chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2, Cp statistic, AIC)

Machine Learning

- **Philosophy** - replace humans in the processing of data
- **Focus** - how can we predict the outcome using the data?
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning (not mutually exclusive)

Statistics

- **Philosophy** - provide humans a set of data analysis tools
- **Focus** - what is the relationship between the data and the outcome?
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2, chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2, Cp statistic, AIC)

Machine Learning

- **Philosophy** - replace humans in the processing of data
- **Focus** - how can we predict the outcome using the data?
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning (not mutually exclusive)

Statistics

- **Philosophy** - provide humans a set of data analysis tools
- **Focus** - what is the relationship between the data and the outcome?
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2, chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2, Cp statistic, AIC)

Machine Learning

- **Philosophy** - replace humans in the processing of data
- **Focus** - how can we predict the outcome using the data?
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning (not mutually exclusive)

Statistics

- **Philosophy** - provide humans a set of data analysis tools
- **Focus** - what is the relationship between the data and the outcome?
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2, chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2, Cp statistic, AIC)

Machine Learning

- **Philosophy** - replace humans in the processing of data
- **Focus** - how can we predict the outcome using the data?
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning (not mutually exclusive)

Statistics

- **Philosophy** - provide humans a set of data analysis tools
- **Focus** - what is the relationship between the data and the outcome?
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2, chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2, Cp statistic, AIC)

Machine Learning

- **Philosophy** - replace humans in the processing of data
- **Focus** - how can we predict the outcome using the data?
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics and machine learning complement each other

The best solution could be an algorithmic model (machine learning), or maybe a data model, or maybe a combination. But the trick to being a **scientist** is to be open to using a wide variety of tools.

— Leo Breiman

The objective is not just to get a better fit to the data but to have a predictive model that *generalises* well, that is, gives good predictions to *unseen* data

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonious model i.e a model which is complex enough to describe "well" our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Features: Covariates, Predictors, Inputs, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generalisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonious model i.e a model which is complex enough to describe “well” our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Features: Covariates, Predictors, Inputs, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generalisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonious model i.e a model which is complex enough to describe "well" our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Features: Covariates, Predictors, Inputs, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generalisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonious model i.e a model which is complex enough to describe "well" our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Features: Covariates, Predictors, Inputs, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generalisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonious model i.e a model which is complex enough to describe "well" our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Features: Covariates, Predictors, Inputs, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generalisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonious model i.e a model which is complex enough to describe “well” our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Features: Covariates, Predictors, Inputs, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generalisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonious model i.e a model which is complex enough to describe "well" our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Features: Covariates, Predictors, Inputs, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generalisation error

A bird's-eye view of machine learning