Machine Learning for OR & FE

Introduction and Course Overview

Martin Haugh

Department of Industrial Engineering and Operations Research
Columbia University
Email: martin.b.haugh@gmail.com

What is Data Mining / Machine Learning / "Big Data"?

Data mining / machine learning / "big data" are all related to the process of discovering new patterns from large data sets.

There is substantial overlap between all three "fields"

– they all use ideas from statistics and algorithms.

According to Hand, Mannila and Smyth:

> *"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."*

But isn't all this just statistics?!

Differences Between Machine Learning and Statistics?

Traditional statistics:

- First hypothesize, then collect data, then analyze.
- Often model-oriented with an emphasis on parametric models.
- Focus on understanding and hypothesis testing.

Machine learning (ML):

- Few if any a priori hypotheses.
- Typically data has already been collected.
- Analysis is typically data-driven not hypothesis-driven.
- Often algorithm-oriented rather than model-oriented.
- Focus on prediction.
- ML is more associated with artificial intelligence than statistics.

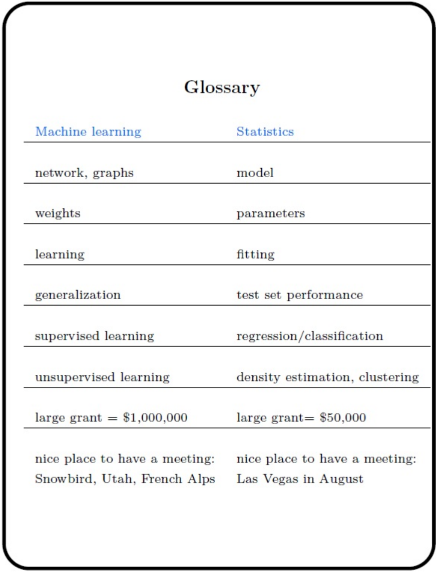
But statistical ideas and algorithms are extremely useful in ML

- – e.g. inference, bootstrapping, over-fitting solutions, regression and classification algorithms.

"Big data" also encompasses data-management technology

- – e.g. cloud computing, NoSQL databases, map-reduce and Hadoop etc.

Robert Tibshirani Compares Machine Learning with Statistics :-)



Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = \$1,000,000	large grant= \$50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

But “uncool” jargon isn’t unique to statistics; witness “neural networks”.

Why the Massive Popular Interest in “Big Data”?

There are many reasons including:

- The explosion of data from many domains such as biology, the internet, astrophysics, telecommunications, engineering and sensor networks.
- The availability of ever cheaper storage
    - developments in database technology such as NoSQL databses
- Faster and cheaper computation to analyze the data.
- Competitive pressure in business: data has value!
- Software developments:
    - commercial products: SAS, SPSS, Google Analytics, IBM, Oracle, Amazon, S-Plus
    - open source products: R, Python libraries such as scikit learn and pandas, Weka.

Don’t necessarily need to be a ML expert to do ML

– but many pitfalls so some competence is required!

Is “Big Data” Over-Hyped?

But “big data” is perhaps over-hyped in the popular media

– and startup scene seems focussed on less important matters:

MIT Technology Review
VOL. 115 NO. 6 | $5.99 US
HAS QUANTUM COMPUTING FINALLY ARRIVED? Upfront p24
HOW TOMORROW'S STARTUPS WILL BE FUNDED Business Report p75
TECH TRANSFORMS MUSIC, ART, AND PROSE Reviews p87
Buzz Aldrin, Apollo 11 moonwalker, would like a word with you.
You Promised Me Mars Colonies. Instead, I Got Facebook.
We've stopped solving big problems. Meet the technologists who refuse to give up. p26

Budd Aldrin, *MIT Technology Review*, October 2012.

But There Are Many Data Mining Successes ...

- Google: the entire company!
- Market Basket (WalMart).
- Recommender Systems: Amazon.com, Netflix.
- Fraud Detection in Telecommunications: AT&T.
- Targeted Marketing: Target.
- Financial Markets: Hedge funds using text mining, Bloomberg.
- DNA Microarray analysis.
- Smart grid: dynamic demand response.

Taxonomy of Machine Learning Problems

Supervised Learning

- Observe training samples, and learn a function mapping inputs to outputs.
- Examples include regression, classification, ranking, etc.

Unsupervised Learning

- Observe data and construct a low complexity description of the data.
- Clustering, dimensional reduction techniques such as principal components analysis, non-negative matrix factorization etc.

Reinforcement Learning (RL)

- RL attempts to learn a map from system states to actions with the goal of maximizing total cumulative reward.
- A given action in a given state influences what the next state will be
  - so need to trade off exploitation (of current knowledge) with exploration (to learn new information).
- Closely related to dynamic programming except that we do not know the parameters of the model.

Will focus on supervised and unsupervised problems in this course.

Data Types

Data comes in many forms ...

- Flat files or vector data.
- Text data; e.g. an article or collection of articles.
- Transactional data: web logs, phone calls.
- Relational data from relational databases.
- Time series data; e.g. financial data, auction data from EBay.
- Image data and video data.
- Spatial temporal data e.g. geophysical data.
- Network data: physical networks, social networks etc.

Regardless of original data-type, typically need to convert it into quantitative data

– hopefully without losing important information.

Exploratory Data Analysis (EDA)

Goal: To obtain a general sense and understanding of the data

- EDA is a vital and necessary first step for any machine learning / statistics task.

EDA tasks:

- What do the marginal distributions look like? Normal, skewed, etc.
- What is the quality of the data? Is some of it missing?
  - If so, why is it missing? Is it missing at random (MAR), missing completely at random (MCAR) or missing not at random (MNAR)?
  - A very important issue!
- Are there outliers? What should we do with them?
- What are the dependencies / correlation between different variables?
- What subsets of the variables are of particular interest?
- What sort of (functional) relationships are we looking for?

General guidelines for EDA:

- Make it data-driven and typically model-free ...
- Think interactive and visual ...
- Many dimensions to play with: x, y, z, space, color, time ...

Some EDA Strategies

Summary statistics:

- mean, median, variance, skewness, kurtosis

Single variable visualization:

- Histograms
- Density estimates
- Box plots

Two variable visualization:

- Scatter plot
- Binning
- Transparent plots
- Contour plots
- Bar charts: categorical variables

Bottom line: it is always well worth looking at your data!

  - EDA will not be a focus of this course but it should always be the start of any analysis.

Data Visualization

Data visualization is an increasingly important subfield of “big data”

– if “a picture is worth a thousand words” what is a video worth?

See Hans Rosling’s famous talk on the joy of statistics (and global health dynamics) at `http://www.youtube.com/watch?v=jbkSRLYSojo`.

Data visualization also becoming increasingly important in journalism

– see `http://www.theguardian.com/data` if you need convincing.

Domain Knowledge

An expertise in machine learning / statistics generally not enough to apply these methods successfully.

Domain-specific knowledge generally also required:

- How can we apply techniques to data from biology, the internet, astrophysics, telecommunications, engineering and sensor networks, etc. without understanding these domains?
- How do we know what the right questions are?
- How can we make sense of the results?
- Or figure out if the results even make sense?

The most successful ML applications almost always based on good domain knowledge

- also helps in the EDA process.

Should avoid the temptation of blindly applying your favorite ML tool

- a hammer looking for a problem is not a good recipe for success!

Course Highlights I: Classification and Reduced-Rank LDA

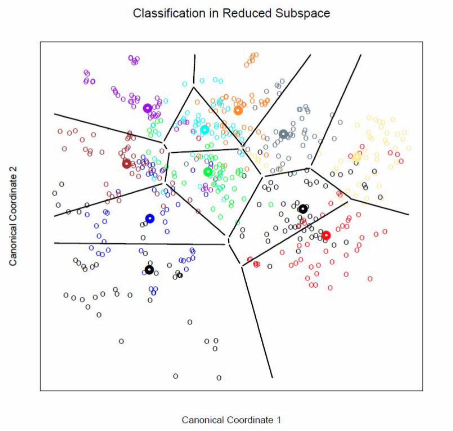


Figure 4.11 from HTF: Decision boundaries for the vowel training data, in the two-dimensional subspace spanned by the first two canonical variates. Note that in any higher-dimensional subspace, the decision boundaries are higher-dimensional affine planes, and could not be represented as lines.

Course Highlights II: Clustering and Vector Quantization

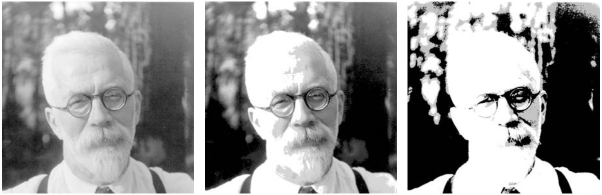


FIGURE 14.9. *Sir Ronald A. Fisher (*$1890 - 1962$*) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a* 1024×1024 *grayscale image at* 8 *bits per pixel. The center image is the result of* 2×2 *block VQ, using* 200 *code vectors, with a compression rate of* 1.9 *bits/pixel. The right image uses only four code vectors, with a compression rate of* 0.50 *bits/pixel*

Figure 14.9 taken from HTF

Course Highlights III: Dimension Reduction Techniques

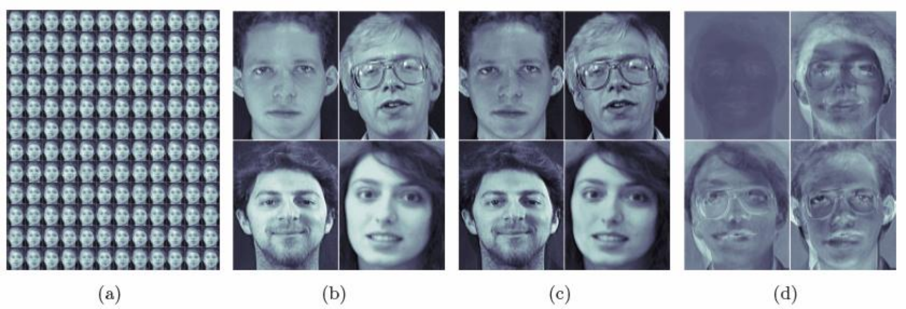


Figure 15.15 from Barber: (a) Training data, consisting of a positive (convex) combination of the base images. (b): The chosen base images from which the training data is derived. (c): Basis learned using conditional PLSA on the training data. This is virtually indistinguishable from the true basis. (d): Eigenbasis (sometimes called ‘eigenfaces’).

Course Highlights IV: MCMC and Cryptography

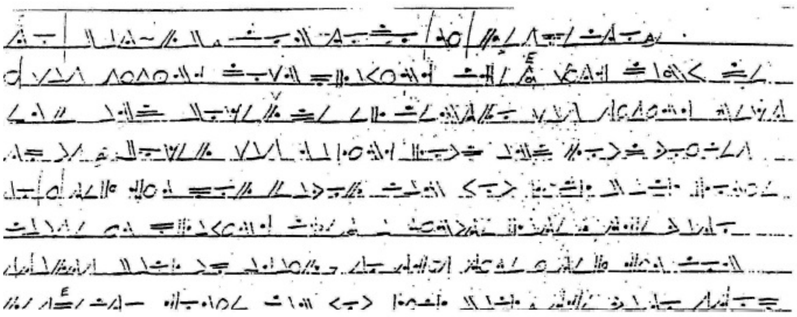
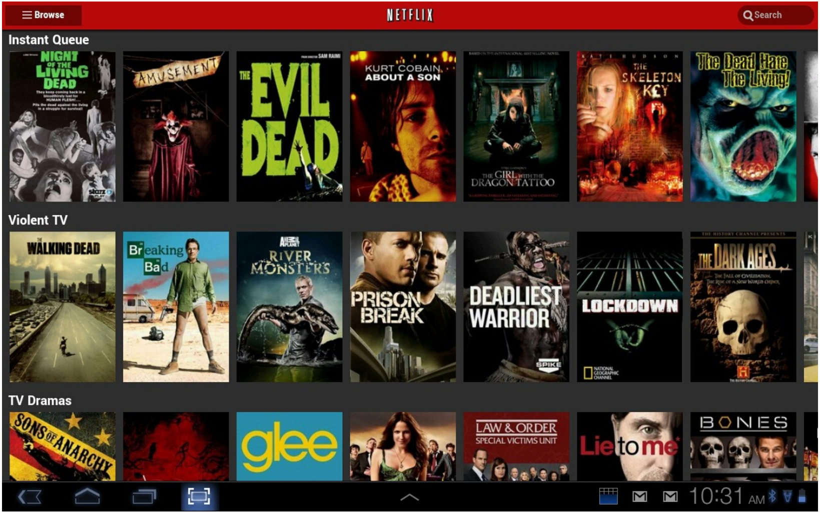


Figure taken from “The Markov Chain Monte Carlo Revolution”, by Persi Diaconis in the *Bulletin of the American Mathematical Society* (2008).

- An unusual application of MCMC: decoding the coded message.

Course Highlights V: Collaborative Filtering and Recommender Systems



Source: *Netflix*

Course Highlights VI: HMMs and Burglar Tracking

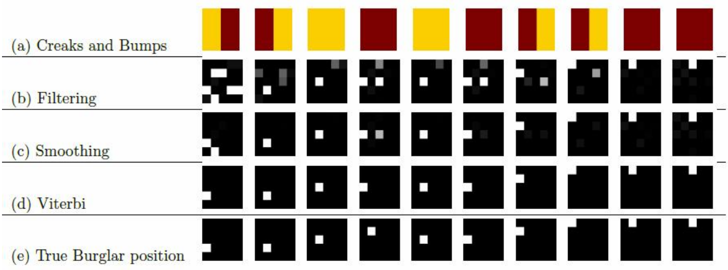
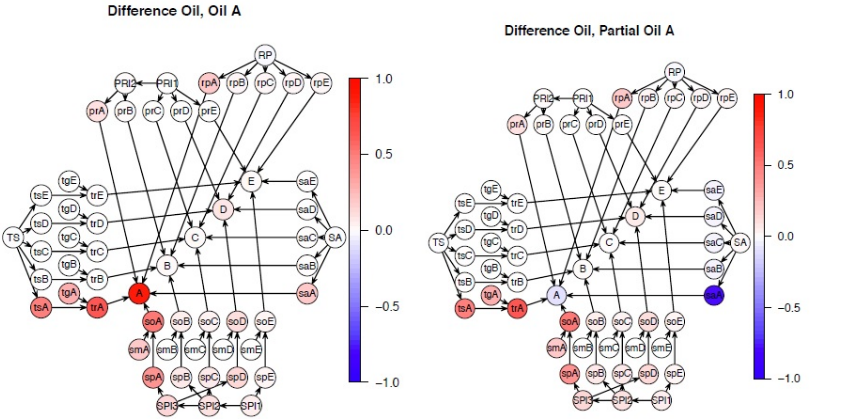


Figure 23.7 from Barber: Localising the burglar through time for 10 time steps. (a): Each panel represents the visible information $v_t = \left(v_t^{\text{creak}}, v_t^{\text{bump}}\right)$, where $v^{\text{creak}} = 1$ means that there was a 'creak in the floorboard' ($v^{\text{creak}} = 2$ otherwise) and $v^{\text{bump}} = 1$ meaning 'bumped into something' (and is in state 2 otherwise). There are 10 panels, one for each time $t = 1, \ldots, 10$. The left half of the panel represents v_t^{creak} and the right half v_t^{bump}. The yellow shade represents the occurrence of a creak or bump, the red shade the absence. (b): The filtered distribution $p(h_t \mid v_{1:t})$ representing where we think the burglar is. (c): The smoothed distribution $p(h_t \mid v_{1:10})$ that represents the distribution of the burglar's position given that we know both the past and future observations. (d): The most likely (Viterbi) burglar path $\text{argmax}_{h_{1:10}} \, p(h_{1:10} \mid v_{1:10})$. (e): The actual path of the burglar.

Course Highlights VII: Graphical Models and Oil Exploration



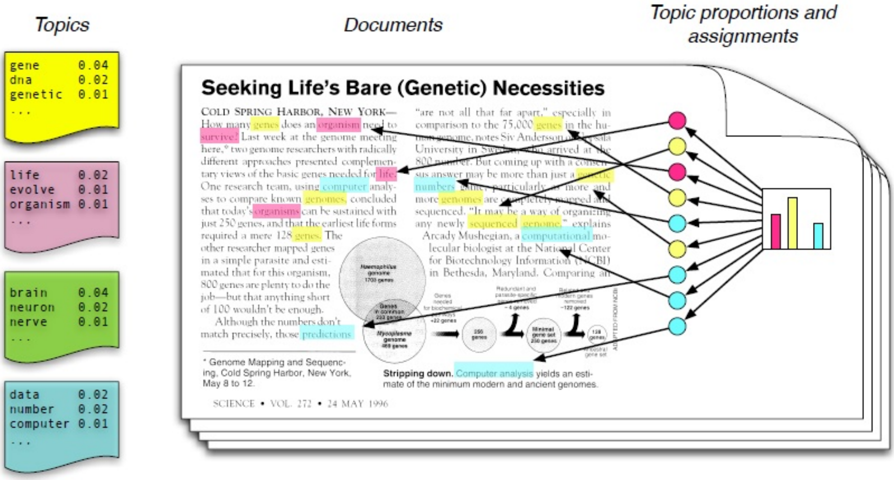
Difference Oil, Oil A

Difference Oil, Partial Oil A

Figure 4: Difference between conditional probabilities given evidence and prior probabilities. Evidence is observed in prospect A, and follows the explanations in section 3.4 . Figure shows the effect of Evidence 1 (left) and 2 (right).

Figure taken from “Strategies for Petroleum Exploration Based on Bayesian Networks: a Case Study”, by Martinelli et al. (2012).

Topic Modeling: an Unlikely Highlight :-(But ...



Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 468 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

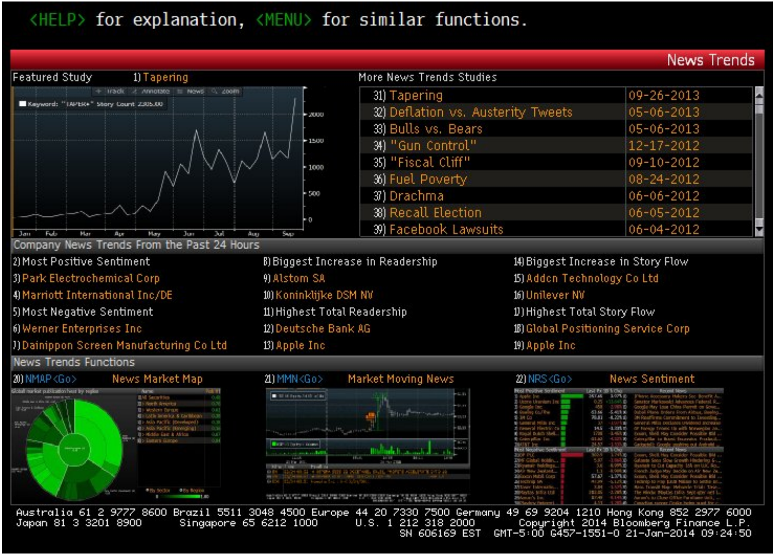
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

[*] Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Figure taken from "Introduction to Probabilistic Topic Models", by D.M. Blei. (2011).

Lots of Machine Learning in Finance but . . .



<HELP> for explanation, <MENU> for similar functions.

News Trends

Featured Study 1) Tapering

More News Trends Studies

31) Tapering	09-26-2013
32) Deflation vs. Austerity Tweets	05-06-2013
33) Bulls vs. Bears	05-06-2013
34) "Gun Control"	12-17-2012
35) "Fiscal Cliff"	09-10-2012
36) Fuel Poverty	08-24-2012
37) Drachma	06-06-2012
38) Recall Election	06-05-2012
39) Facebook Lawsuits	06-04-2012

Company News Trends From the Past 24 Hours

2) Most Positive Sentiment
3) Park Electrochemical Corp
4) Marriott International Inc/DE
5) Most Negative Sentiment
6) Werner Enterprises Inc
7) Dainippon Screen Manufacturing Co Ltd

8) Biggest Increase in Readership
9) Alstom SA
10) Koninklijke DSM NV
11) Highest Total Readership
12) Deutsche Bank AG
13) Apple Inc

14) Biggest Increase in Story Flow
15) Addcn Technology Co Ltd
16) Unilever NV
17) Highest Total Story Flow
18) Global Positioning Service Corp
19) Apple Inc

News Trends Functions

20) NMAP <Go> News Market Map

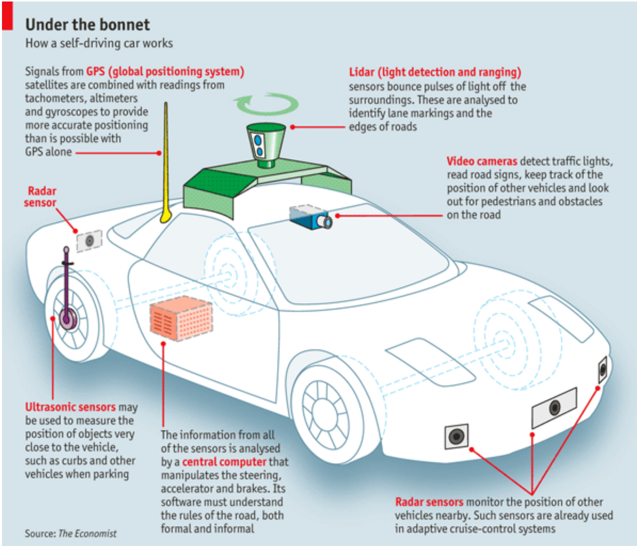
21) MMN <Go> Market Moving News

22) NRS <Go> News Sentiment

Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2014 Bloomberg Finance L.P.
SN 606169 EST GMT-5:00 G457-1551-0 21-Jan-2014 09:24:50

Source: *Bloomberg*

The *Google* Smart Car (Not a Course Highlight!)



Under the bonnet

How a self-driving car works

Signals from **GPS (global positioning system)** satellites are combined with readings from tachometers, altimeters and gyroscopes to provide more accurate positioning than is possible with GPS alone

Lidar (light detection and ranging) sensors bounce pulses of light off the surroundings. These are analysed to identify lane markings and the edges of roads

Video cameras detect traffic lights, read road signs, keep track of the position of other vehicles and look out for pedestrians and obstacles on the road

Radar sensor

Ultrasonic sensors may be used to measure the position of objects very close to the vehicle, such as curbs and other vehicles when parking

The information from all of the sensors is analysed by a **central computer** that manipulates the steering, accelerator and brakes. Its software must understand the rules of the road, both formal and informal

Radar sensors monitor the position of other vehicles nearby. Such sensors are already used in adaptive cruise-control systems

Source: *The Economist*

Source: *The Economist*