Artificial Intelligence, Big Data and the Future

Ruey S. Tsay
Booth School of Business, University of Chicago

December 6, 2019

Outline

1. Artificial Intelligence (AI), Machine learning & Deep learning
2. Big Data (Oil of the information age?)
3. The Marriage of AI & 5G (and beyond)
4. The Impact: Improved living with increasing disparities
5. The Value of Data
6. Why is Statistics Relevant?
7. The Challenges
8. Concluding Remarks

What is AI? Knowledge and Digital Economy

- A tool to make computers think and behave intelligently.
- A science to study and mimic how human brain works.
- A field in which we train machines to understand patterns and behaviors of certain entities.

What is machine learning (ML)? What is deep learning (DL)?
Alternative (or modern) names for AI

Statistical definition of AI (or ML or DL):
A semi-parametric model with sophisticated algorithms for classification and prediction.
A known function with many unknown parameters (weights and biases)

Why is AI important? Attract so much attention?

- Great potential: affects every aspect of our lives
- Good timing: have the necessary ingredients and abilities to train the machines

  Information revolution: [Industrial revolution IV?]
  - Data are widely available and easily accessible.
  - tremendous computing power, storage and optimization
  - Efficient algorithms [Advances in science and technology.]
- News media hypes: AlphaGO, AlphaStar, AI Books (e.g., Lee Kai-Fu), self-driving cars, FinTech
- Exciting events: DeepMind & esport (video games, 100 trillion trillion possible moves)

Three important ingredients of AI

- Data (input) (or rules)
- Computing power and methods (optimization)
- Algorithms (Human input)

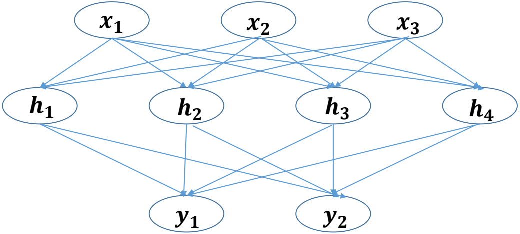
$\boxed{\text{Data}} \Rightarrow \boxed{\text{Deep Learning}} \Rightarrow \boxed{\text{Application}}$

Deep Learning (or Machine Learning)

- A collection of algorithms: See the book AIQ by Polson and Scott (2018).
- Mainly a multiple-layer neural network (deep network)
- A complicated semi-parametric statistical model (with many parameters)
- Enjoy several advantages (nonlinearity)
- Can process all types of data
- Encounter certain difficulties and drawbacks too

What is a neural network?

A simple feed-forward 3-4-2 neural net: 3 input variables, one hidden layer with 4 nodes, and 2 output variables.



How does it process information?

- Network structure: how many layers? number of nodes, etc.
- Activation functions: many available, e.g. logistic function

$$h_i(.) = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta}_i)}{1 + \exp(\boldsymbol{x}'\boldsymbol{\beta}_i)}.$$

- Many types of networks available, e.g., recurrent network (feedback)
- Training: divide data into *training subsample* and *validation subsample*.
- Certain objective function is defined for optimization in the training (classification versus prediction)
- Typical method of training: back-propagation + stochastic gradient decent
- Some difficulties appear for deep networks, but they can be overcome.

Impacts of AI

- Affects our lives in every way possible
- Changes our behavior and thinking, but not overnight
- Suitable for well-defined tasks, e.g. image classification, play games, routine tasks, etc. [Smarter than the current robots.]
- Affects labor markets, business operations, and human relationships
- Reduce distance and time, increase productivity, and more
- 5G accelerates the advances and applications of AI
- Creates more disparities in the society
- Will not overtake human, but will replace many, especially those unprepared.

Limitations of Deep Learning

- One size simply cannot fit all
- Lack theoretical foundation: computation alone cannot solve all problems [global approximation.]
- Black box: lack of guidance in building or comparing DL
- Lack uncertainty quantification: Real world is stochastic
- Available data cannot cover all future events
- Algorithms are rational, human decisions are not.
- Lack artificial irrationality (or stupidity)
- Dangers of self-fulfilling prophecy? Diminishing data values
- Security concern: Vulnerable to outside attack and misuse

Big Data

Forbes, May 21, 2018 (B. Marr)

1. 2.5 quintillion bytes data collected each day, i.e.

$$2,500,000,000,000,000,000$$

   How big is this?

   - 2.5 quintillion pennies cover earth 5 times
   - Our brains have 100 billion neurons

2. 90% of available data were collected over the last 2 years
3. 3.7 billion people use Internet daily
4. Google processes more than 40,000 search per second.
5. By 2020, 1.7MB data created every second per person on earth. [DOMO's report, Irfan Ahmad, 6/15/2018]

Mainly: Internet of Things

Forms of Big Data

Many different types of data: Structured and un-structured.

1. Numerical values
2. Videos
3. Images
4. Texts
5. Sentiments
6. Instagram (photos), etc.

Concerns over big data

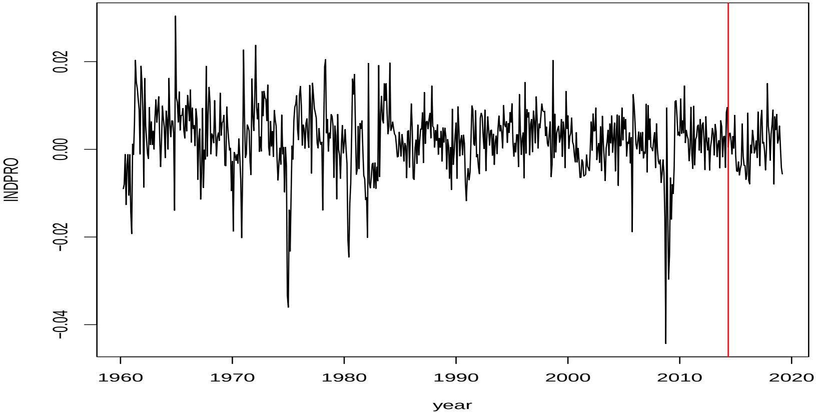
$$\boxed{\text{More Data}} \overset{?}{\Longrightarrow} \boxed{\text{More Information}}$$

- Size is not everything
  Value Data are more useful than Big Data
- Internet of Things provide observational data.
  They have limitations: For instance,
  - confounding: Not well-known in general
  - hard to make causal inference (critical in decision making)
- High heterogeneity
- Often encounter selection bias
- Misinformation: Fake news, manipulated rating (Amazon?)
- Diminishing value in data: More data are generated by AI

An illustration of deep learning: forecasting IP growth rate

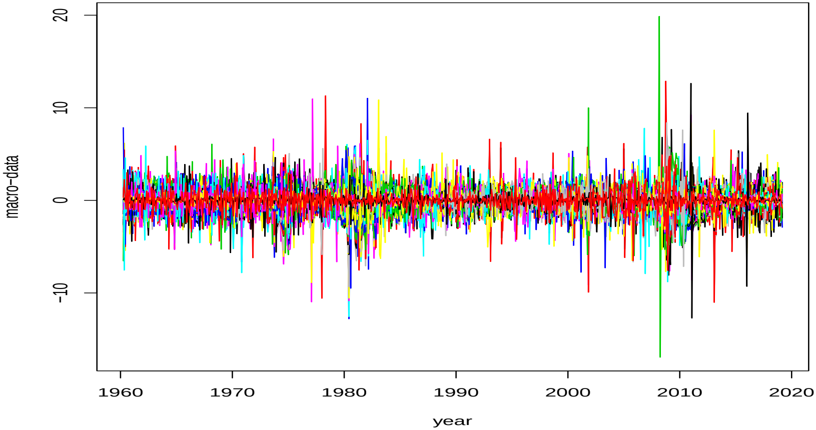
- Variable of interest: US monthly growth rate of industrial production index
- Data period: March 1960 to March 2019 with 708 observations
- Predictors: 122 monthly US macroeconomic variables (Lag-1 to Lag-4: 488 predictors)
- Training subsample: First 650 observations
- Forecasting subsample: 651 to 708
- Number of layers: 2, 3, to 6, each with 100 nodes
- Activation functions: Hyperbolic tangent + identity
- Learning rates: e^{-2} to e^{-3}
- Benchmark: OLS and Ridge regression

US monthly growth rate of industrial production (IP) index



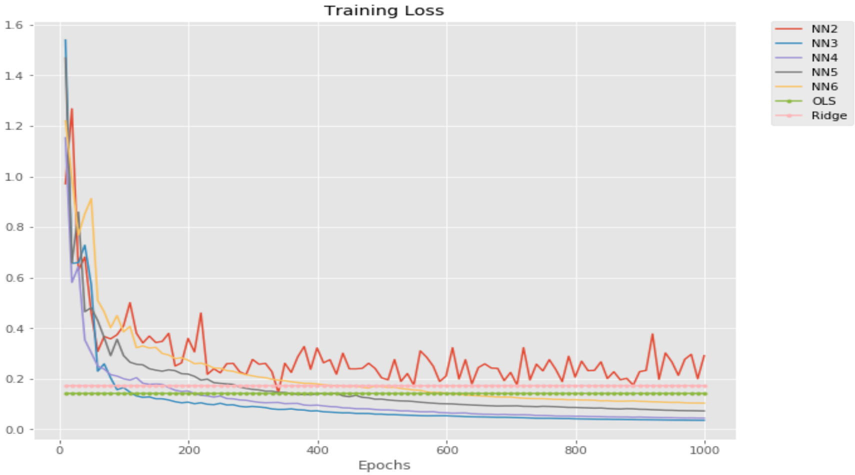
Last 58 observations form the forecasting subsample

Macro variables used: 122 with 708 observations

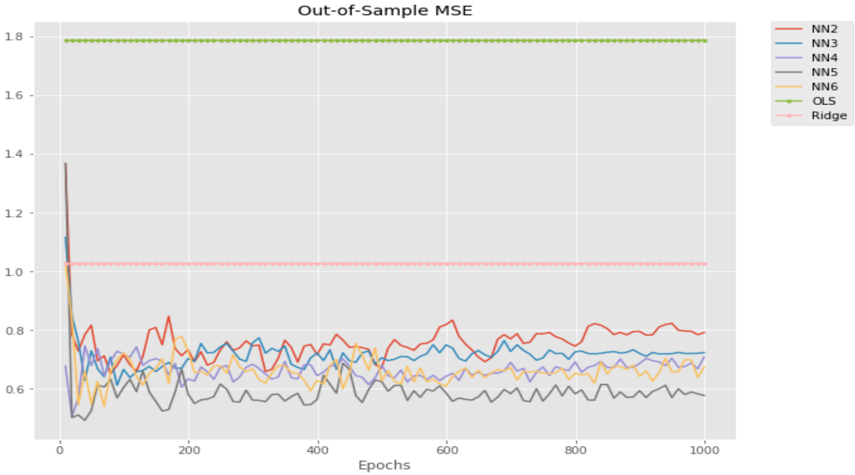


Series are standardized: March 1960 to March 2019

Training


Training Loss
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
Epochs
NN2
NN3
NN4
NN5
NN6
OLS
Ridge

Forecasting



Why is Statistics Relevant?

There is no true model in real applications!

Statistics can contribute in the following areas:

1. Understand the limitations of methods used in AI and devise ways to improve them
2. Assess the value of data
3. Select important predictors and model comparison
4. Quantify uncertainty

Example: omitted variables in linear regression

Problem statement: Variable of interest Y

- True model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, where X_1 and X_2 are dependent (related)
- Omitted variable: X_2
- Model used: $Y = \gamma_0 + \gamma_1 X_1 + e$.

No matter how big the sample size is, $\hat{\gamma}_1$ does not converge to β_1.

(1). Framework: Assume two correlated factors v_{1t} and v_{2t}. Assume also risk-free interest rate is available and data are demeaned.
The model is (in excess returns)

$$\boldsymbol{r}_t = \boldsymbol{\beta}\boldsymbol{\gamma} + \boldsymbol{\beta}\boldsymbol{v}_t + \boldsymbol{u}_t,$$

where $\boldsymbol{u}_t$ is the idiosyncratic noise, $\boldsymbol{\beta}$ is a matrix of risk exposures and $\boldsymbol{\gamma}$ is the vector of risk premia for the two factors.

Omitted variables continued.

(2). Suppose g_t is a proxy of factor v_{1t}. [v_{1t} may not be tradable.]

(3). Goal: to estimate γ_1 (the risk premium of g_t).

Methods available in finance to estimate $\boldsymbol{\gamma}$:

- Two-pass regression: Fama and MacBeth (1973)
  1. Time-series regression: regress g_t on $\boldsymbol{v}_t$ to obtain $\widehat{\boldsymbol{\beta}}$
  2. Cross-section regression: regress average excess returns on $\widehat{\boldsymbol{\beta}}$ to obtain $\widehat{\boldsymbol{\gamma}}$.
- Mimicking-portfolio approach: Haberman et al. (1987)
  1. Regress g_t on a set of tradable asset returns to construct a tradable portfolio [maximum correlation with g_t.]
  2. Cross-section regression: project average excess returns on the constructed tradable porfolio.

Omitted-variable bias exists in both methods.

Big data come to rescue

Based on the paper by Giglio and Xiu (2019)

Basic ideas: Use principal component analysis & diffusion indexes

- Many asset returns are available (they use 647 portfolios)
- Assume that the omitted variables are in the column space of the available returns [Omitted variables are linearly measurable functions of observed returns.]
- Use a three-pass method to overcome the omitted variable bias

Ideas and Methods used

1. Big return data:
$$\boldsymbol{R} = \boldsymbol{\beta}\boldsymbol{\gamma}\boldsymbol{1}_T' + \boldsymbol{\beta}\boldsymbol{V} + \boldsymbol{U},$$
where $\boldsymbol{R}$ is $n \times T$, $\boldsymbol{1}_T$ is T-dimensional vector of 1s, $\boldsymbol{V}$ is $p \times T$ matrix of factors, n is the number of observed asset returns, T is the sample size, and p is the number of factors.
2. Let $\bar{\boldsymbol{A}}$ as demeaned variables. The data become
$$\bar{\boldsymbol{R}} = \boldsymbol{\beta}\bar{\boldsymbol{V}} + \bar{\boldsymbol{U}}. \tag{1}$$
3. The factors of interest are related to the true factors as
$$\boldsymbol{G} = \boldsymbol{\delta} + \boldsymbol{\eta}\boldsymbol{V} + \boldsymbol{Z},$$
which can be written, via demeaned variables, as
$$\bar{\boldsymbol{G}} = \boldsymbol{\eta}\bar{\boldsymbol{V}} + \bar{\boldsymbol{Z}}. \tag{2}$$
4. The risk premia of $\boldsymbol{G}$ is
$$\boldsymbol{\gamma}_g = \boldsymbol{\eta}\boldsymbol{\gamma}. \tag{3}$$

Empirical procedure

Use big return data to obtain proxy of all factors & Apply two-pass regression: True factors are latent variables.

(a). Perform PCA of $\bar{\boldsymbol{R}}$ and use information criterion (Bai and Ng (2002)) to select p, the number of factor.

$$\hat{\boldsymbol{V}} = T^{1/2}(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_{\hat{p}})'$$

where $\boldsymbol{\epsilon}_i$ are the eigenvectors corresponding to the largest $\hat{p}$ eigenvalues,

$$\hat{\beta} = T^{-1}\bar{\boldsymbol{R}}\hat{\boldsymbol{V}}'.$$

(b). Cross-section regression: Regression average excess returns $\bar{r}$ on $\widehat{\boldsymbol{\beta}}$ to obtain risk premia of latent factors

$$\widehat{\gamma} = (\widehat{\boldsymbol{\beta}}'\widehat{\boldsymbol{\beta}})^{-1}\widehat{\boldsymbol{\beta}}'\bar{r}.$$

Empirical procedure continued

(c). Time-series regression: regress g_t on the extracted factors to estimate $\boldsymbol{\eta}$.

$$\hat{\boldsymbol{\eta}} = \bar{\boldsymbol{G}}\hat{\boldsymbol{V}}'(\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}')^{-1}.$$

(d). Finally,

$$\widehat{\gamma}_g = \hat{\boldsymbol{\eta}}\widehat{\boldsymbol{\gamma}}.$$

Empirical findings

Giglio and Xiu (2019): 647 portfolios including equities sorted by many characteristics, bonds, and currencies.

- Risk premia vary substantially depend on whether one considers omitted variables or not.
- Risk premium of the market portfolio is positive, significant, and close to the time-series average of market excess returns.

Example II: Variable selection and model comparison

Over-fitting in ML

- Theoretical work: blessing of high dimensionality (asynptotic zero-penalty)
- Applications: over-fitting pays a price (finite sample)

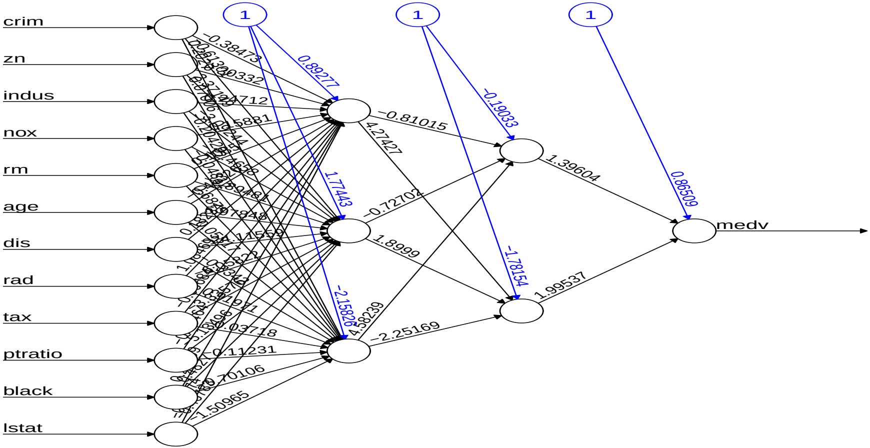
A well known result: Autoregressive model
True model: AR(p), but fit an AR($p + h$) model with $h > 0$.
Forecasting MSE increases by a factor of $\frac{h}{T}$, where T is the sample size.
If h is fixed, $h/T \to 0$.
if h increases with T, e.g. $h/T \to c > 0$, MSE increases

Boston housing price example

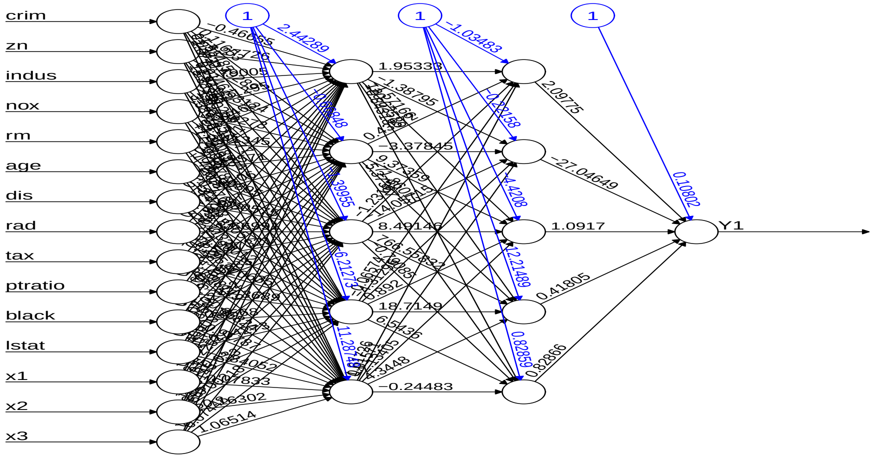
Well known data set in Statistics: Harrison and Rubinfeld (1978)

- Dependent variable: median housing price (log)
- Predictors: 13 variables include crime rate, % lower status of the population, tax, age, etc.
- Sample size: 506
- Artificial predictors: three random variables drawn from $N(0, 1)$

A simple neural network: Boston data



A simple neural network: Boston data with 3 noise variables



How to identify the three irrelevant noise predictors?

Out-of-sample prediction: Run 11 iterations of the following:

- Training sample: randomly select 360 data points
- Forecasting sample: the remaining 146 data points
- Variables used:
  - All 15 variables (12 predictors + 3 noises)
  - Use 12 variables: omit 13-15, 12-14, 11-13, etc.
- Compute the MSE of predictions for each model, each iteration

Median MSE of prediction

15 predictors (12+3 noise): 0.0564

12 predictors:

Model	13-15	12-14	11-13	10-12
MSE	0.0336	0.0628	0.0794	0.0841
Model	9-11	8-10	7-9	6-8
MSE	0.0533	0.0635	0.0652	0.0523
Model	5-7	4-6	3-5	2-4
MSE	0.0562	0.0668	0.0661	0.0567
Model	1-3			
MSE	0.0655			

Model 13-15 means predictors 13 to 15 are removed.
The same result holds for average MSE

The Challenges

1. Social and economical inequalities: increasing wealth gap
   - Richest 1% own 50%+ of world's wealth (CNBC)
   - In US, the median household income of the richest states is about double that of the poorest states (Census)
2. Income inequality: with/without AI background
3. Wealthy vs poor economies: poor becomes poorer
4. Strong competition at light speed: knowledge & information
5. Protection of human right and privacy

Keys to success

- Education! Education! Education!
- Cannot satisfy with being an AI user, but a creator. Taiwan is behind several countries in AI related patent applications.
- Team work with effective communication skills
- Mathematics, Computer Science, Statistics, Optimization (skills)
- Liberal arts (a better person)
- Domain knowledge (specialty)

Concluding Remarks

1. No short cut. Success is for those who prepare well.
2. AI is a two-edged sword

Things to do:

- Education
- Team work with excellent communication skills
- Embrace AI, create AI, and make good use of AI
- Study the failure of AI, especially develop theory for artificial irrationality (or stupidity).
- Protect human right, democracy, and privacy. [even from government agencies.]
  [Prevent misuse of AI.]