Systematic Strategies
and
Machine Learning

Kevin Noel

Principal Data Scientist

contact: noelkevin1@gmail.com

PLAN

Introduction

1) Overview of Systematic Strategies
    1-a) Framework
    1-b) Strategies Overview

2) Some overview of Machine Learning
    2-a) Learning principles
    2-b) Some applications : Dependence measure, ...
    2-c) Potential future research

Disclaimers / Context

- Wide coverage or In depth coverage of the field ?
  - Field(s) are too broad to do **wide** and **In depth** (ie equations) coverage in 1 hour.

- Focus on “Wide coverage”
  - Present concepts/principles (vs full details equations)
  - Focus on “core” principles/assumption and intuitive views.
  - Try to provide a different perspective than usual.

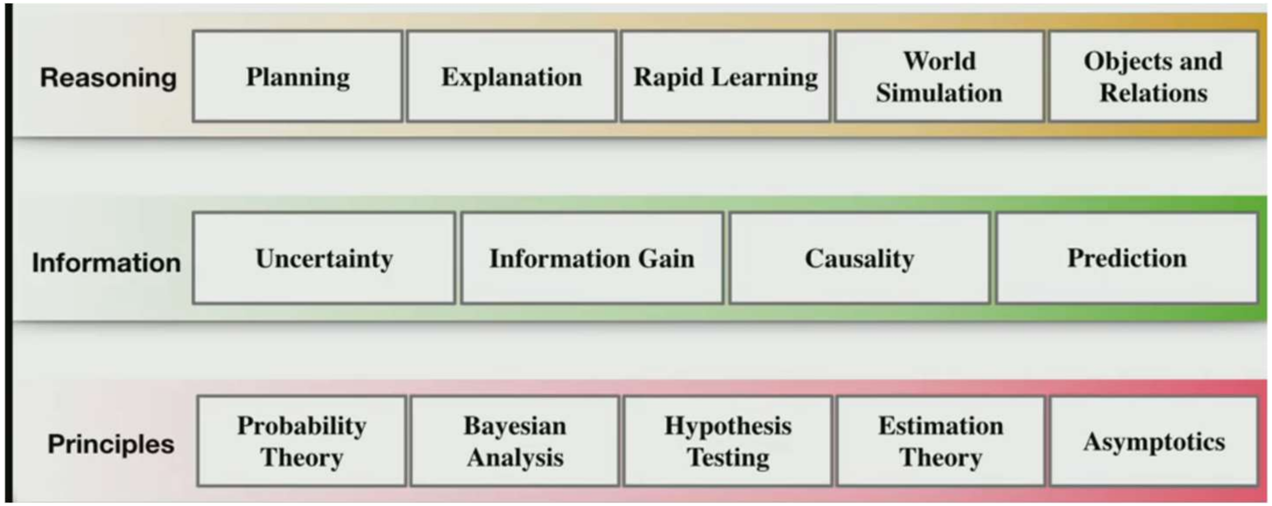
- Any question, please contact the author.

Introduction

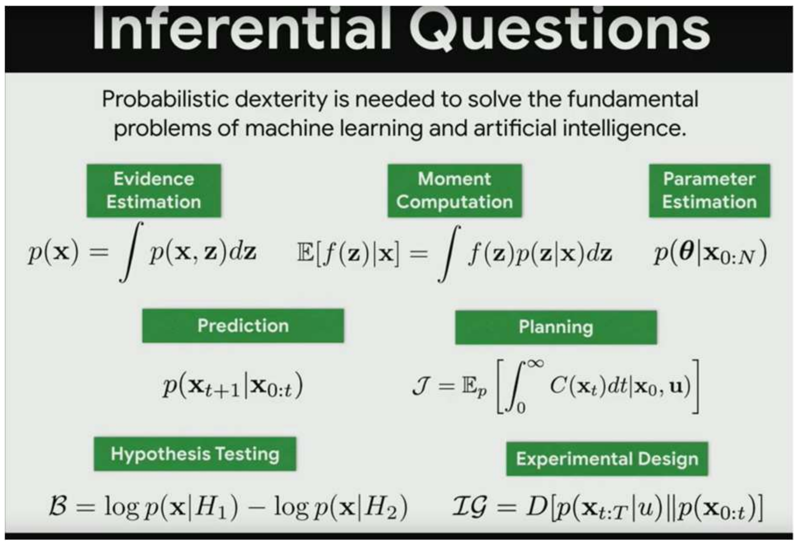
Introduction : Common relationship between fields : ML, Statistical Quant. Fin.

<u>Goal:</u> Process information to discover "new relationships” and applied them, …

--> Some analogy with Science dev. (*from DeepMind*):



Q.Finance Theory



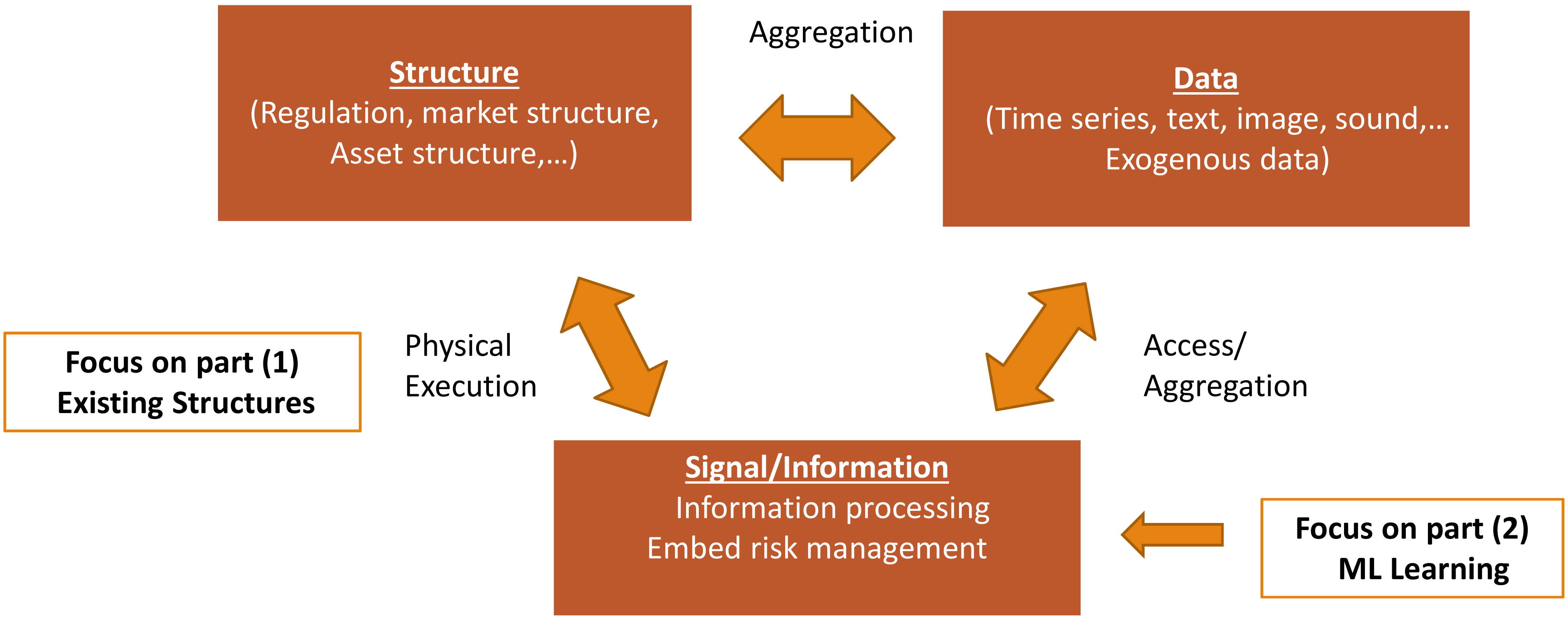
1) Overview of Systematic Strategies

Blank : request the author for details.

2) Some overview of Machine Learning

a) Learning principles

1.b) Strategies Overview : Summary



2.a) Learning principles : Overview

- Machine learning, is usually considered (in usage) :
  - As extension of statistical signal processing technics.
  - As a way to quantify un-structured data: text, images.

BUT, There are **fundamental differences** in the approach

(vs structured model + Calibration based).

2.a) Learning principles: vs Model calibration

- Model calibration (Math/Physics concept)
  - Structural model/equation explaining a phenomenon: BScholes, Newtonian Dynamics….
  - Minimize error between measurement and model macro-prediction : **Ground Thruth**
  - **Uniqueness, Existence** of optimal parameters (i.e. math problem is well defined).
  - Implicit notion of **Ground Truth** recovery from model (+calibration).

- Learning process (ML concept, i.e. more "Bayesian view")
  - Defined from PAC Theory: **Probability Approximated Correct Learning**
  - Find a function approximator based on samples with error bounded in probability, in polynomial compute time.
  - Notions of : **Approximation, Probabilistic and High Dimensionality.**
  - "Learn based on samples" --> Learn a representation of the "concept".

2.a) Learning principles : Some remarks

- Calibration meaning in "Machine Learning":
  - Normalization of "output probability": Calibrate estimated probability vs actual probability.
  - → Related to normalization of output across models.

- Learning assumption : **IID Independent Identical Distributed** :
  - For 2 sample sets : (X1) ,(X2), we have : distribution(X1) = distribution(X2)
  - Does **NOT** hold in finance time series : major issues.

  → Sol: Make the feature IID, Adjust the learning process (No IID learning).

- High dimensionality : Find approximation in very high dimensional spaces (in ML)
  - If low dimensional → Estimation with traditional statistics process is **more efficient.**
  - Efficient "representation" : and re-use this representation later for new data.
  - IID Hypothesis **P(new Data) = P(Old data)** (i.e. at least in the representation space).
  - Learning in Polynomial time.

2.a) Learning principles : Overview of ML models

Type of ML models	Example	Remarks
Formulae based	Regression, polynomial	Nb of parameters are limited/fixed (Structural or local model)
Parametric Structure (size is undefined)	Tree, SVM, RF, ...	Model structure is fixed (Sparse structure)
Compositional and Architecture based	Neural Network	Model can have dense and flexible structure. Universal Approximators
…. (ex: Open compositional Architecture, transfer learning)	...	….

2.a) Learning principles: Many learning (ways)

Learning type	Description	Challenges
Supervised Learning **Semi-supervised**	Input X, label y pairs : closest to modeling. Label engineering.	Find **Good Labels** (!)
Un-supervised/ **Self-supervised**	Clustering, AE, Dimension reduction methods. Extract label from raw data (ie NLP)	Find stable patterns, Efficient computing
Reinforcement **/Control Learning**	Evaluate sample action with Environment : State, Action, Feedback.	Find efficient training/computing
Differential Learning	Learn from samples, while restricting granularity of learning. (i.e. learning from sensitive, private data...),	Tradeoff Accuracy vs Privacy
Meta Learning	'Learning to learn" : Learn a tasks (ie classification) using very few samples. ie : Image recognition, NLP	Mostly on-going research.
Transfer Learning	Use pre-trained Model to train other new models : actively used in Image, NLP	Find "Big generic" model
Adversarial Learning	Training from pair models (Generator, Discriminator). Generator generates 'adversarial samples" Discriminator tries to differentiate fake samples vs actual samples.	Find efficient training process, Computational challenges
Causality Learning, **Stable learning**	DAG Learning of Direct Acyclic Graph and Invariance.	"Causality", DAG formalism is very difficult.
"Graph Learning"	Learns graph representation, graph related task, from samples.	Computational, On going research
Machine Teaching	Control of machine learning. M. teaching designs the optimal training data to drive the learning algorithm to a target model.	On going research (Microsoft)

2.a) Learning principles : Overview : Application in finance/sys. strategies

- Application in finance requires extra work/caution :
  - Since the signal ratio noise/information is high. More chance to learn noise than actual information.
  - No stationary of phenomenon: Rules does not hold over very long time (outlier).
  - Not IID and lack of samples for low frequency strategy.

  - Dynamic system with varying fundamental conditions (i.e. regulation, …)
  - Zero sum game: Edge in a relative way rather absolute way (vs competitors).

    Higher entry level : longer the persistence of alpha.

2) Overview of Machine Learning Technics

b) Some applications

2.b) Some applications

- 3 main applications (simplified):
  - Structured or Semi-structured data processing: time series, tabular data,…
  - Un-structured data processing: Text, Image, Sound
  - Prediction and Control : RL and Control,….

- Goal : Structured / Un-structured data into signals (ie score, …).
  Quantify the state of the “market & economy” system and “predict some outcome”.

2.b) Some applications: General dependency measure in noisy signals

Assessing dependencies in noisy data (numerical, categorical,..)

Some examples of measures

Correlation based	r^2	Squared Pearson's correlation	–	✗	$\mathcal{O}(n)$
	ACE	Alternative conditional expectation	Breiman and Friedman (1985)	✗	$\mathcal{O}(n)$
	dCorr	Distance correlation	Székely and Rizzo (2009)	✓	$\mathcal{O}(n \log n)$
	RDC	Randomized dependency coefficient	Lopez-Paz et al. (2013)	✓	$\mathcal{O}(n \log n)$

Information theory based	MIC	Maximal information coefficient	Reshef et al. (2011)	✗	$\mathcal{O}(n)$
	GMIC	Generalized mean information coefficient	Luedtke and Tran (2013)	✗	$\mathcal{O}(2^n)$
	MID	Mutual information dimension	Sugiyama and Borgwardt (2013)	✗	$\mathcal{O}(n \log n)$
	MIC_e	Maximal information coefficient	Reshef et al. (2015b)	✗	$\mathcal{O}(n)$
	TIC_e	Total information coefficient	Reshef et al. (2015b)	✗	$\mathcal{O}(n)$
	RIC	Randomized information coefficient	–	✓	$\mathcal{O}(n^{1.5})$

(Reference : https://link.springer.com/article/10.1007/s10994-017-5664-2)

2.b) Some applications: General dependency measure in noisy signals

Randomized Information Coefficient (RIC) for 2 random variables X, Y :

$$\mathrm{NI}((\mathbf{X},\mathbf{Y}|\mathbf{G}) \triangleq \frac{\mathbf{I}((\mathbf{X},\mathbf{Y})|\mathbf{G})}{\mathbf{max}\{,\mathbf{H}(\mathbf{X}|\mathbf{G}),\mathbf{H}(\mathbf{Y}|\mathbf{G})\}}$$

$$\mathrm{RIC}(\mathbf{X},\mathbf{Y}) \triangleq \frac{\mathbf{1}}{\boldsymbol{K}} \sum_{\mathbf{1}}^{\boldsymbol{K}} \mathrm{NI}((\boldsymbol{X},\boldsymbol{Y})|G_{\boldsymbol{k}})$$

RIC is computed by randomly generating *K* discretization grids $G_k G_k$ and averaging the normalized mutual information NI.

where *I* and *H* are respectively the mutual information and the entropy function for discrete variables, normalized by *max{H(X),H(Y)}.*

The intuition behind RIC is that on average a random grid can encapsulate the relationship between X and Y. Both **random discretization** and **ensembles of classifiers** have been shown to be effective in machine learning.

- Low-variance statistics based on information and ensemble theory.

- **Measure's strong performance on**
  - (i) Discrimination between strong and weak noisy relationships.
  - (ii) Ranking of relationships, to its low variance estimation of mutual information.

2.b) Some applications: NLP/NLU (Natural Language Understanding …)

- Text data is the "**new oil**" in Finance (!) : very large volume (beyond news processing).

- Recent progress: Generating scores from raw texts (News, PRelease, Earning Calls …)
  - **Multiple Sentiment Scores**
  - **Auto-Tagging of text (ie category)**

- Major efforts required in :
  - Data pre-processing: Normalization (i.e. incorporate Domain knowledge)
  - Calibration/Measure of score impact (along with label engineering).

- Neural Network (LSTM, Attention,…) taking over Traditionnal ML (SVM, RF)

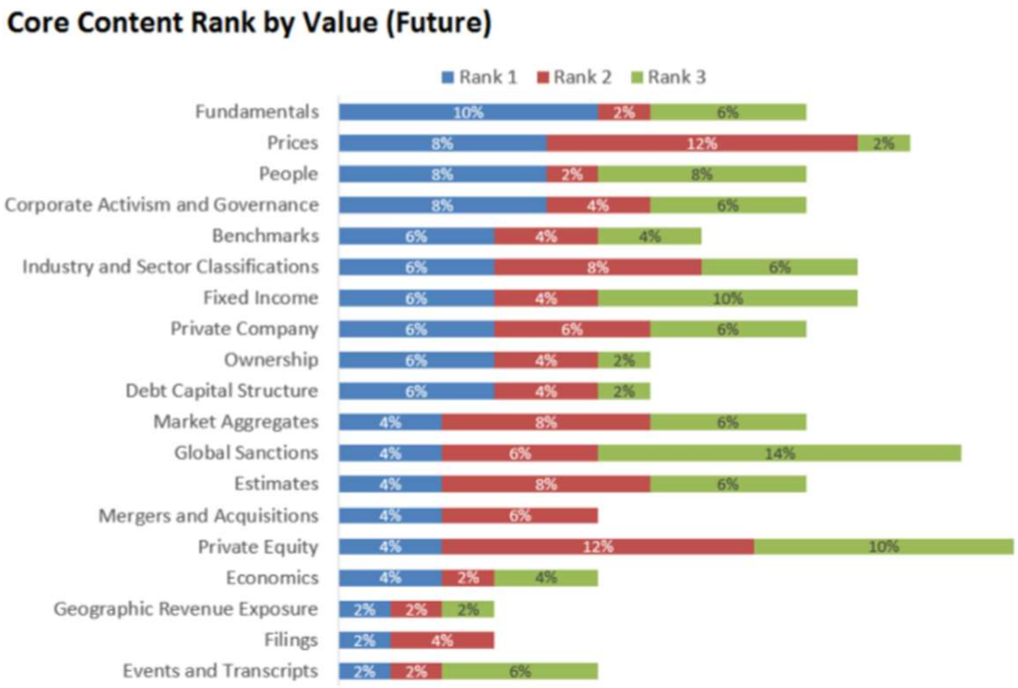
2.b) Some applications : NLP/NLU

Remarks about Text data :

- Amount of Raw Text data is immense :
  A lot are not accessed directly.

- Raw text data is very diversified :
  Twitter to Regulatory fillings

- Many challenges :
  Quantifying “Nuances” and Context,…
  Text generation…



From Factset

2.b) Some applications : NLP/NLU , active usage of Transfer Learning

- Transfer Learning to reduce the amount of (training samples, labels) pairs.

Pre-trained model

Large trained NLP model M1 (ie BERT, …) on large dataset. → New specialized Dataset → Specialized model M1 (ie specific tagging)

Transfer by distillation :

Large NLP model M1 (ie BERT, …) trained on large dataset. → Use Smaller NLP model M2 which learns From Big Model

Learning by distillation
(i.e. regression on softmax)

2.b) Some applications: NLP/NLU , active usage of Transfer Learning

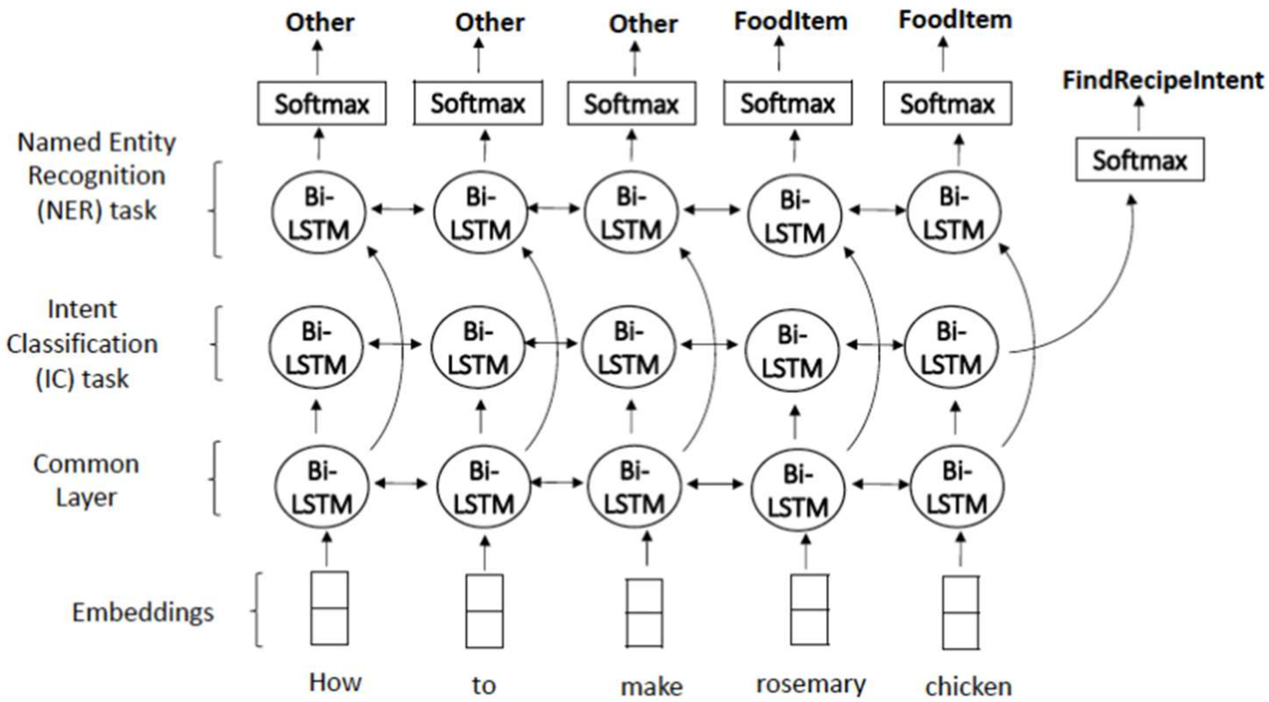
- Example of Multi-task learning : with transfer Learning

- Unit cell are Bi-directional LSTM.

- Common layer “**transfers**” **probability** to other layer tasks” :

  Task 1 : Predict intention

  Task 2 : Name Entity



Stacked multitask bi-LSTM model architecture

From Amazon

2.b) Some applications : NLP/NLU

- Recent progress in NLP/ NLU:

  - Attention mechanism : “Learned” and “Identified” specific part of text
  - Bi-directional prediction: Text as bi-directional token data (left→right, right→ left)
  - Auto-regressive : Text as sequential auto-regressive tokens.
  - Enlarge the capacity size of model :

    BERT : 340 million parameters (4-6 days of training)

    GPT2 : 350 million parameters

    MLtron (Nvidia) : 8 billion parameters. (512GPU, 10 days training, 150go of Text data).

Reference : https://nv-adlr.github.io/MegatronLM

2) Overview of Machine Learning Technics

c) Future research topics

2.c) Future Research Topics : Neural control

❑ Idea

- Replace “parametric based” control/allocation by “Neural Network“ control.
- Learning to Control (ie related to RL) : stability/confidence level is also a requirement (major challenge).

❑ Actual research applications:

- Choice of optimal collateral choice in Collateralized (i.e. swap and MtM). (P.H. Labordere paper)
- Successful in some dynamic system control : Drone control, Physics simulation,…
- Option pricing for american exercice (approximate the price).

2.c) Future Research topics : Neural Control

Challenges :

- Low number of training samples (vs High frequency data and RL application).
- Stability of control: theoretical bounds (?) when having complex Neural Network architecture.
- Incorporate “Prior Knowledge” : In structural model (explicit), prior knowledge is easier to integrate.

RL vs Optimal control

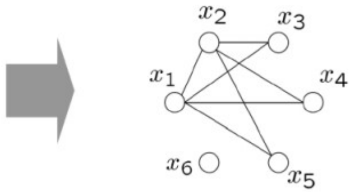
- RL is usually common terminology in CS and refer to “model free” control and reward.
- Optimal Control: terminology used in Math/Operation areas, associated with “model based” control.

2.c) Future Research Topics : Graph Learning

- Represent discrete of rules or relation as Knowledge Graph ; a set of (node, vertex) with labels
  - Raw data sample (i.e. text) into Graph (node, vertex) or into DAG (direct Acyclic Graph) :

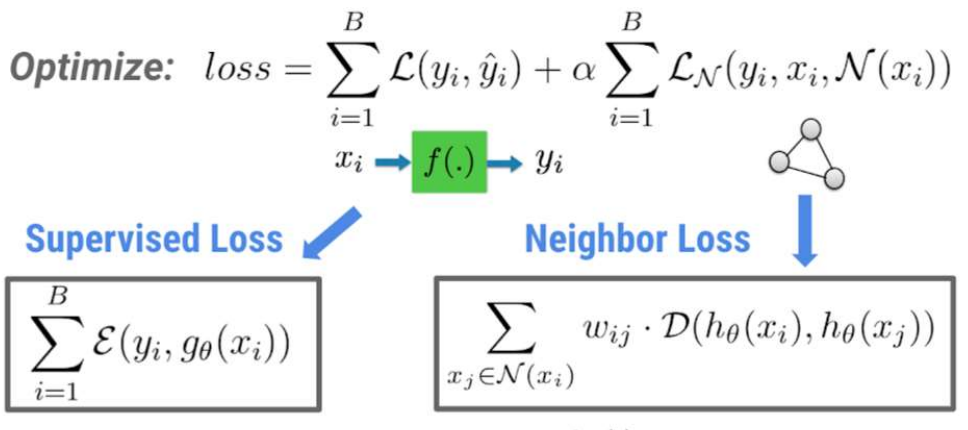
$$Q_{i,j} = 0 \quad \Rightarrow \quad X_i \perp X_j | \mathbf{X}_{-ij} \quad \text{or} \quad p(X_i, X_j | \mathbf{X}_{-ij}) = p(X_i | \mathbf{X}_{-ij}) p(X_j | \mathbf{X}_{-ij})$$

$$\mathrm{Q} = \begin{pmatrix} * & * & * & * & * & 0 \\ * & * & * & * & * & 0 \\ * & * & * & 0 & 0 & 0 \\ * & * & 0 & * & 0 & 0 \\ * & * & 0 & 0 & * & 0 \\ 0 & 0 & 0 & 0 & 0 & * \end{pmatrix}$$



Sparse Precision matrix into graph

  - Reuse the Graph in the training of other models : Embed prior knowledge as loss regularization (ex: neighbor loss).



2.c) Future Research Topics : Graph Learning

❑ Some remarks

- Overall active research
- An efficient way to store of “knowledge” and rules (from various data).
- Challenges in converting raw data to graph.
- Directed Acyclic Graph (DAG) are “difficult” (i.e. causality learning)., active research

Some References

- Transfer Learning, Meta learning : Deep Learning Book (Yoshua Bengio)
- Reinforcement Learning : **Reinforcement Learning**: An **Introduction**. Sutton, Barto c 2014, 2015.
- XLNET : Autoregressive Language model https://arxiv.org/abs/1906.08237
- Neural control : Optimal Posting of Collateral with Recurrent Neural Networks (PH Labordere).

For details, contact : noelkevin1@gmail.com