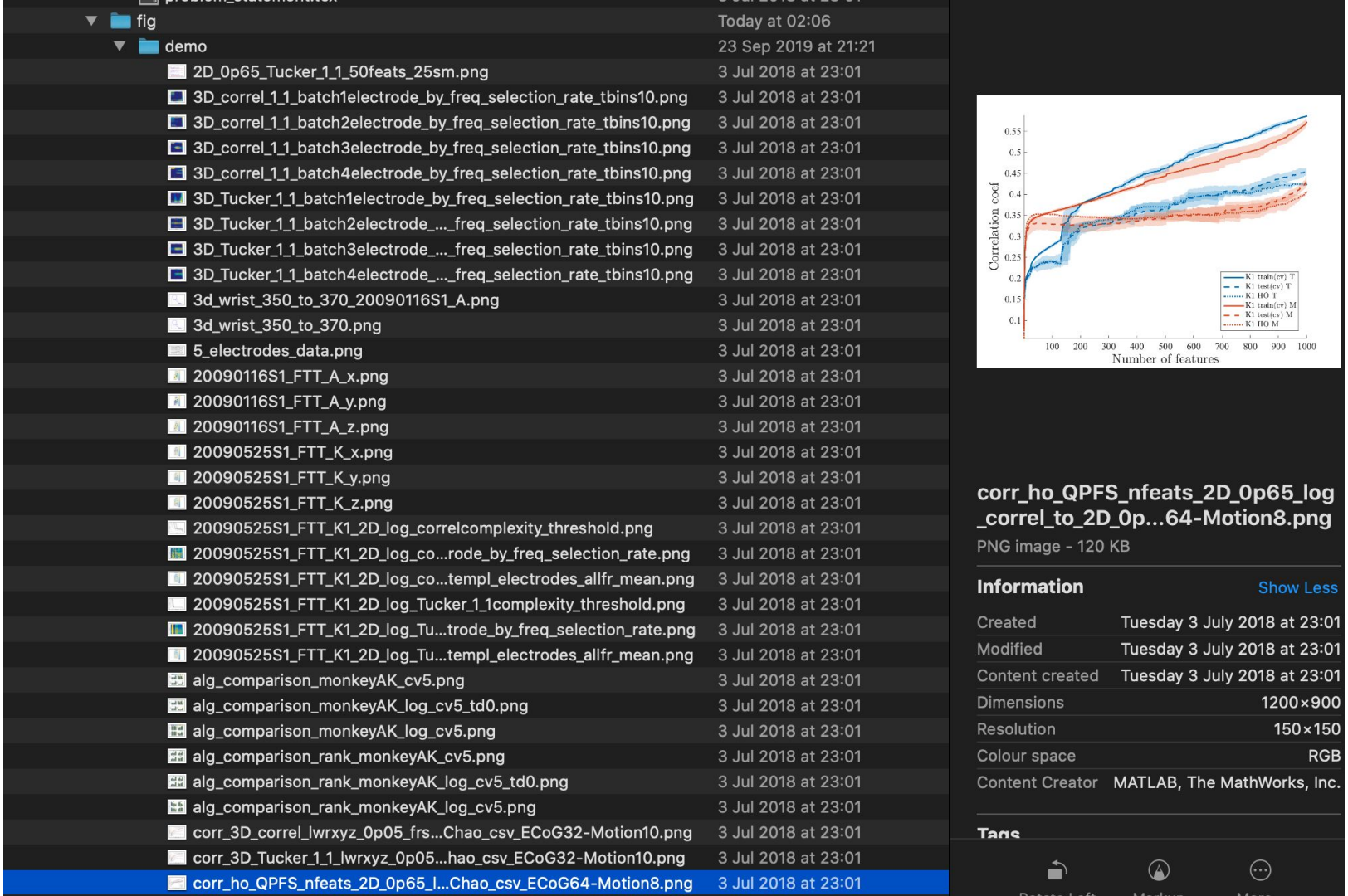
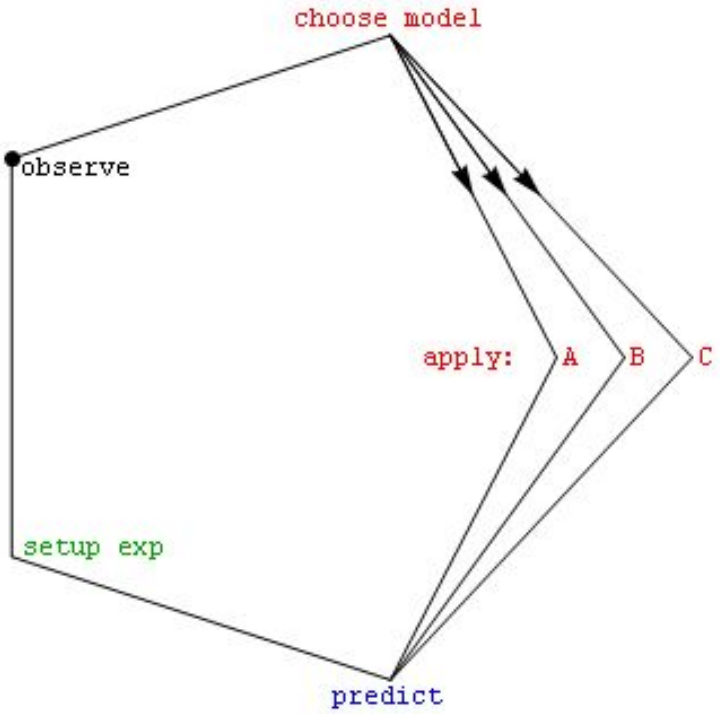
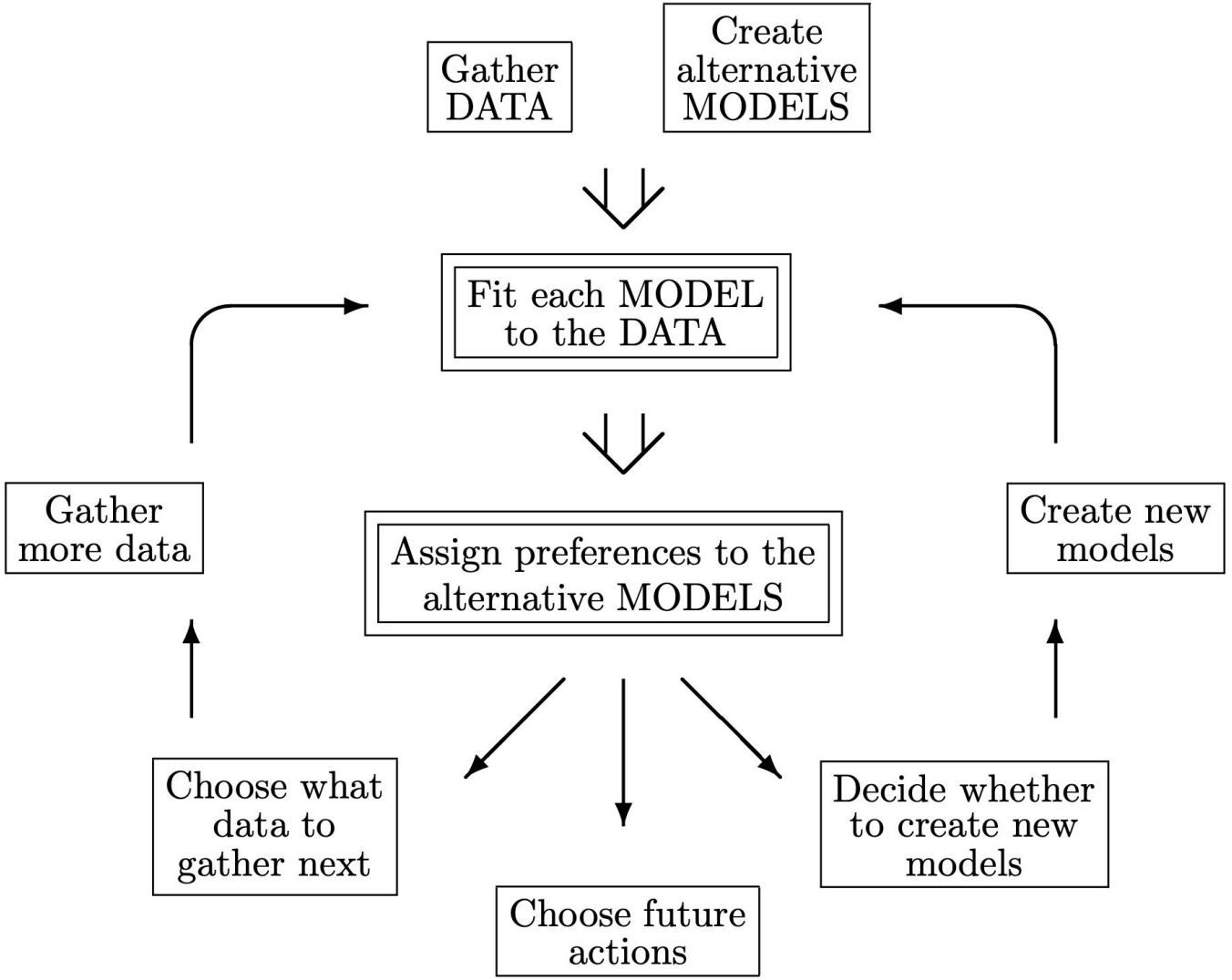


fig
Today at 02:06
demo
23 Sep 2019 at 21:21
2D_0p65_Tucker_1_1_50feats_25sm.png
3 Jul 2018 at 23:01
3D_correl_1_1_batch1electrode_by_freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3D_correl_1_1_batch2electrode_by_freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3D_correl_1_1_batch3electrode_by_freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3D_correl_1_1_batch4electrode_by_freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3D_Tucker_1_1_batch1electrode_by_freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3D_Tucker_1_1_batch2electrode_..._freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3D_Tucker_1_1_batch3electrode_..._freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3D_Tucker_1_1_batch4electrode_..._freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3d_wrist_350_to_370_20090116S1_A.png
3 Jul 2018 at 23:01
3d_wrist_350_to_370.png
3 Jul 2018 at 23:01
5_electrodes_data.png
3 Jul 2018 at 23:01
20090116S1_FTT_A_x.png
3 Jul 2018 at 23:01
20090116S1_FTT_A_y.png
3 Jul 2018 at 23:01
20090116S1_FTT_A_z.png
3 Jul 2018 at 23:01
20090525S1_FTT_K_x.png
3 Jul 2018 at 23:01
20090525S1_FTT_K_y.png
3 Jul 2018 at 23:01
20090525S1_FTT_K_z.png
3 Jul 2018 at 23:01
20090525S1_FTT_K1_2D_log_correlcomplexity_threshold.png
3 Jul 2018 at 23:01
20090525S1_FTT_K1_2D_log_co...rode_by_freq_selection_rate.png
3 Jul 2018 at 23:01
20090525S1_FTT_K1_2D_log_co...templ_electrodes_allfr_mean.png
3 Jul 2018 at 23:01
20090525S1_FTT_K1_2D_log_Tucker_1_1complexity_threshold.png
3 Jul 2018 at 23:01
20090525S1_FTT_K1_2D_log_Tu...trode_by_freq_selection_rate.png
3 Jul 2018 at 23:01
20090525S1_FTT_K1_2D_log_Tu...templ_electrodes_allfr_mean.png
3 Jul 2018 at 23:01
alg_comparison_monkeyAK_cv5.png
3 Jul 2018 at 23:01
alg_comparison_monkeyAK_log_cv5_td0.png
3 Jul 2018 at 23:01
alg_comparison_monkeyAK_log_cv5.png
3 Jul 2018 at 23:01
alg_comparison_rank_monkeyAK_cv5.png
3 Jul 2018 at 23:01
alg_comparison_rank_monkeyAK_log_cv5_td0.png
3 Jul 2018 at 23:01
alg_comparison_rank_monkeyAK_log_cv5.png
3 Jul 2018 at 23:01
corr_3D_correl_lwrxyz_0p05_frs...Chao_csv_ECoG32-Motion10.png
3 Jul 2018 at 23:01
corr_3D_Tucker_1_1_lwrxyz_0p05...hao_csv_ECoG32-Motion10.png
3 Jul 2018 at 23:01
corr_ho_QPFS_nfeats_2D_0p65_l...Chao_csv_ECoG64-Motion8.png
3 Jul 2018 at 23:01
Correlation coef
Number of features
K1 train(cv) T
K1 test(cv) T
K1 HO T
K1 train(cv) M
K1 test(cv) M
K1 HO M
corr_ho_QPFS_nfeats_2D_0p65_log_correl_to_2D_0p...64-Motion8.png
PNG image - 120 KB
Information
Show Less
Created
Tuesday 3 July 2018 at 23:01
Modified
Tuesday 3 July 2018 at 23:01
Content created
Tuesday 3 July 2018 at 23:01
Dimensions
1200×900
Resolution
150×150
Colour space
RGB
Content Creator
MATLAB, The MathWorks, Inc.
Tags

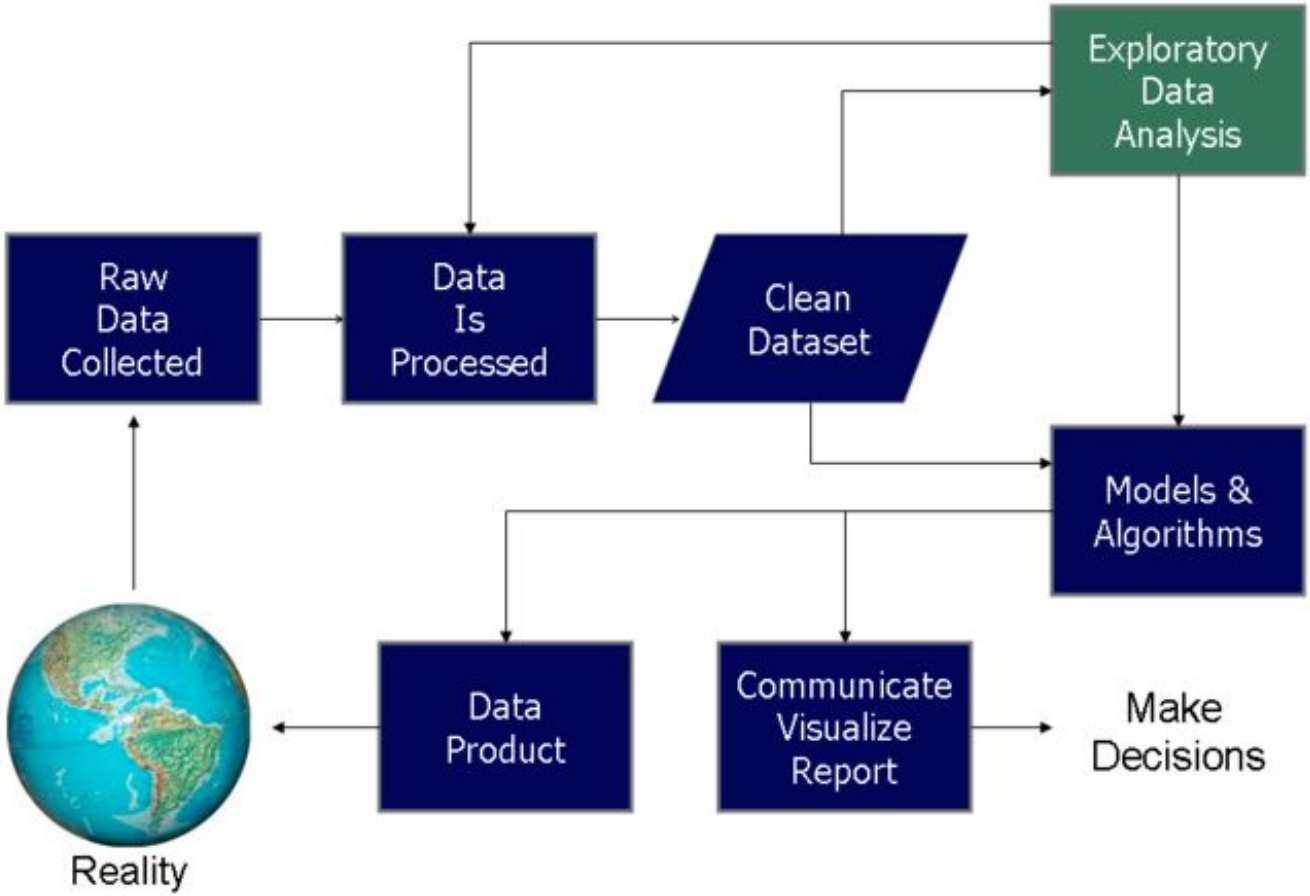
The scientific observation cycle (model selection)



wikipedia.org



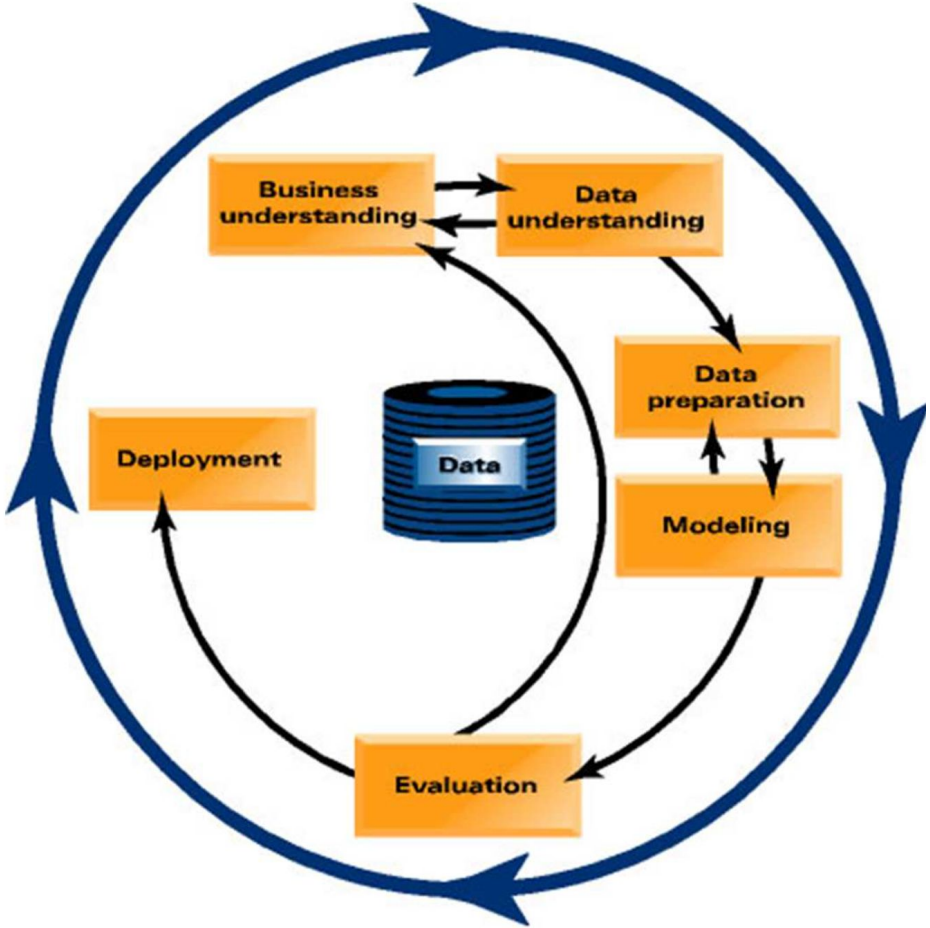
Data Science Process



Cross-industry standard process for data mining (CRISPR-DM)

Six major phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment



Project life cycle

Neural architecture search

automating the design of artificial neural networks (ANN). Methods for NAS are categorized as

- The *search space* defines the type(s) of ANN that can be designed and optimized.
- The *search strategy* defines the approach used to explore the search space.
- The *performance estimation strategy* evaluates the performance of a possible ANN from its design (without constructing and training it)

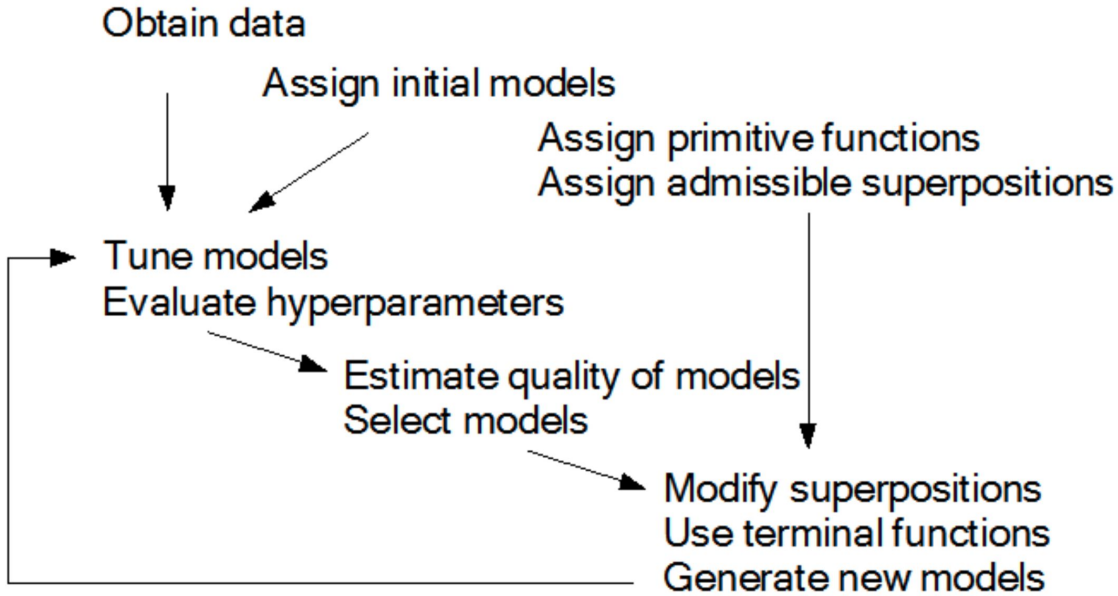
Reinforcement learning

Evolution

Hill-climbing

Multi-objective search

Supernetwork search



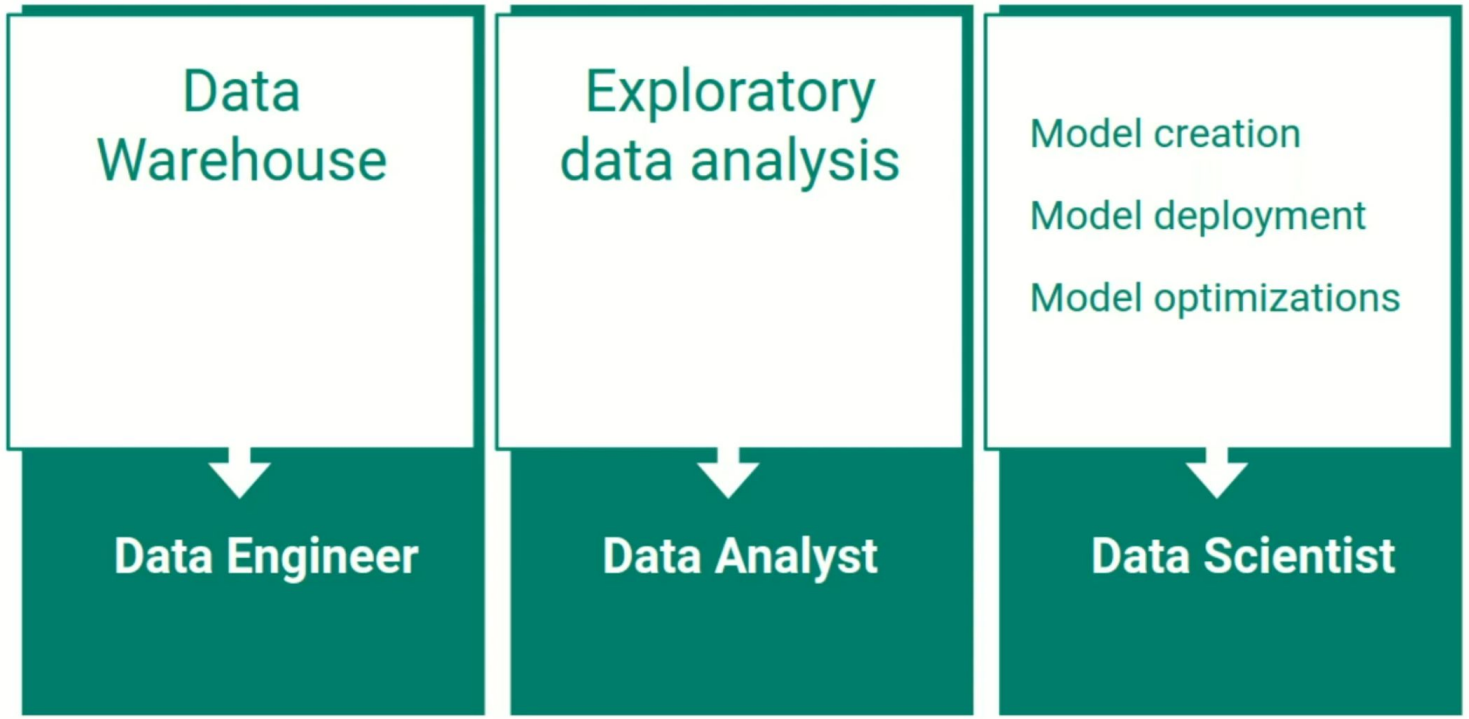
Automated machine learning (AutoML)

- Automated data preparation and ingestion (from raw data and miscellaneous formats)
  - Automated column type detection; e.g., boolean, discrete numerical, continuous numerical, or text
  - Automated column intent detection; e.g., target/label, stratification field, numerical feature, categorical text feature, or free text feature
  - Automated task detection; e.g., binary classification, regression, clustering, or ranking
- Automated feature engineering
  - Feature selection
  - Feature extraction
  - Meta learning and transfer learning
  - Detection and handling of skewed data and/or missing values
- Automated model selection
- Hyperparameter optimization of the learning algorithm
- Automated pipeline selection under time, memory, complexity constraints
- Automated selection of evaluation metrics / validation procedures
- Automated problem checking
  - Leakage detection
  - Misconfiguration detection
- Automated analysis of results obtained
- User interfaces and visualizations for automated machine learning

“Петька, - сказал Чапаев, - ну как ты можешь не быть собой, когда ты и есть абсолютно все, что только может быть?”

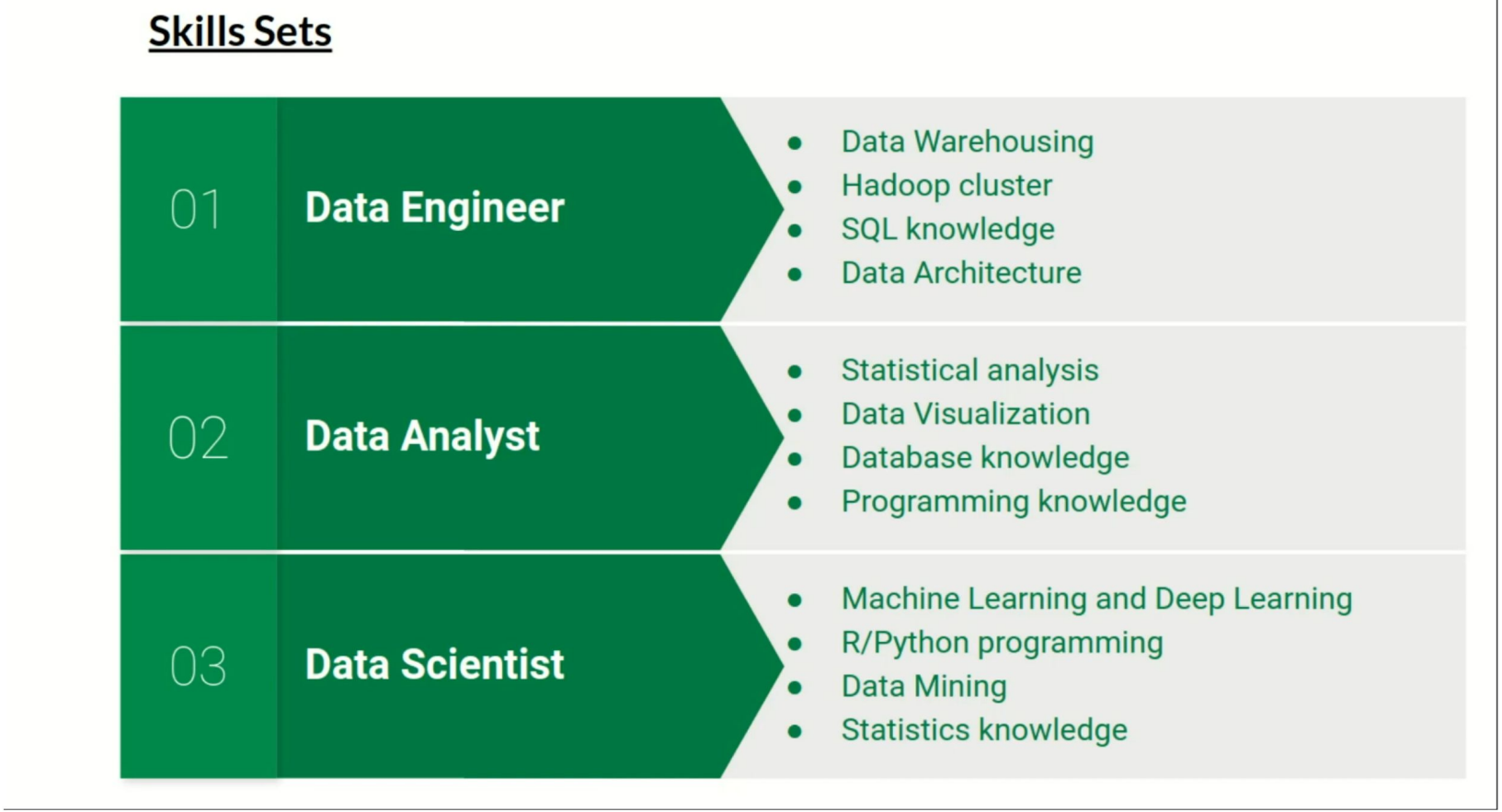
Виктор Пелевин, Дмитрий Козлов

wikipedia.org



Data Engineer vs Data Analyst vs Data Scientist | Salary Comparison | Roles and Responsibilities

125 views • Jan 14, 2020



Data Engineer vs Data Analyst vs Data Scientist | Salary Comparison | Roles and Responsibilities

125 views • Jan 14, 2020

1 0 SHARE SAVE ...

P

Programming Yogi

84 subscribers

SUBSCRIBE

Список модулей для построения модели

1. Семплирование выборки
2. Процедура скользящего контроля
3. Заполнение пропусков
4. Фильтрация выбросов
5. Бинаризация признаков
6. Сегментация признаков
7. Эмбеддинг
8. Порождение признаков, прямое
9. Порождение признаков, метрическое
10. Прогноз, вычисление значений прогностической модели
11. Оптимизация параметров модели
12. Выбор признаков
13. Анализ ошибки
14. Вычисление внешних критериев, отчет

План действий

1. Распределить блоки
2. Самостоятельно для своего блока (10 минут)
   a. Нарисовать схему в стандарте IDEF0
   b. Внутри блока написать действие
   c. Стрелки назвать и продублировать обозначениями
   d. Рядом кратко описать алгоритм действия
   e. Загрузить в http://bit.ly/m1p_file2discuss
3. В парах (10 минут)
   a. Согласовать общую схему для одного или нескольких блоков
   b. Нарисовать схему
   c. Загрузить
4. Общая схема (15 минут)
   a. Нарисовать общую схему на большом листе
   b. Загрузить
5. Обсуждение

Подсказка

Типы стрелок:

- независимая переменная
- целевая переменная
- индексы объектов
- индексы признаков
- ошибки
- прогнозы
- модели
- …

Вместе с порожденными данными (объектами и признаками) порождаются и их описания (индексы)

Открытые данные

Statlog (German Credit Data) Data Set by Dr. Hans Hofmann

- Data Set Characteristics: Multivariate
- Attribute Characteristics: Categorical, Integer
- Associated Tasks: Classification
- Number of Instances: 1000
- Number of Attributes: 20

Список переменных

Variable	Type	Categories
Loan currency	Nominal	3
Applied amount	Linear	
Monthly payment	Linear	
Tetm of contract	Linear	
Region of the office	Nominal	7
Day of week of scoring	Linear	
Hour of scoring	Linear	
Age	Linear	
Gender	Nominal	2
Marital status	Nominal	4
Education	Ordinal	5
Number of children	Linear	
Industrial sector	Nominal	27
Salary	Linear	
Place of birth	Nominal	94
...	...	...
Car number shown	Nominal	2

Преобразование шкал

- Область деятельности заемщика, номинальная шкала

Nominal	Tourism	Banking	Education
John	1	0	0
Thomas	0	1	0
Sara	0	0	1

- Образование заемщика, ординальная шкала

Ordinal	Primary	Secondary	Higher
John	1	0	0
Thomas	1	1	0
Sara	1	1	1

Группировка признаков: оптимизационная задача

Мы имеем начальную модель, заданную набором индексов $\mathcal{A}$. Добавим полученные в результате группировки признаки и рассмотрим улучшение функционала качества.

$$\begin{array}{rccccccl} \xi = & 1 & 2 & 3 & \dots & c, & c \text{ число категорий}, \xi \in C; \\ & \downarrow & \downarrow & \downarrow & & \downarrow & \\ x_j = & \gamma_1 & \gamma_2 & \gamma_3 & \dots & \gamma_c, & |\Gamma| \text{ число групп}, \gamma \in \Gamma. \end{array}$$

Требуется найти функцию

$$h : C \to \Gamma.$$

Задача оптимизации ставится так:

$$(h, |\Gamma|) = \arg \max_{h \in H} S(w)_{\mathcal{A} \cup j}$$

и решается методом полного перебора или генетическим алгоритмом.

Список порождающих функций

Description	In	N in	Out	N out	Comm	Param
Nominal to binary	nom	1	bin	1–4	-	Yes
Ordinal to binary	ord	1	bin	1–4	-	Yes
Linear to linear segments	lin	1	lin	1–4	-	Yes
Linear segments to binary	lin	1	bin	1–4	-	Yes
Get one column of n-matrix	bin	1–4	bin	1	-	Yes
Conjunction	bin	2–6	bin	1	Yes	-
Dijsunction	bin	2–6	bin	1	Yes	-
Negate binary	bin	1	bin	1	-	-
Logarithm	lin	1	lin	1	-	-
Hyperbolic tangent sigmiod	lin	1	lin	1	-	-
Logistic sigmoid	lin	1	lin	1	-	-
Sum	lin	2–3	lin	1	Yes	-
Difference	lin	2	lin	1	No	-
Multiplication	lin,bin	2–3	lin	1	Yes	-
Division	lin	2	lin	1	No	-
Inverse	lin	1	lin	1	-	-
Polynomial transformation	lin	1	lin	1	-	Yes
Radial basis function	lin	1	lin	1	-	Yes
Monomials: $x\sqrt{x}$, etc.	lin	1	lin	1	-	-

Задача порождения признаков

Даны

- измеряемые признаки $\Xi = \{\xi\}$,
- заданные экспертами порождающие функции $G = \{g(\mathbf{b}, \xi)\}$,
$$g : \xi \mapsto x;$$
- правила порождения: $\mathcal{G} \supset G$, где суперпозиция $g_k \circ g_l \in \mathcal{G}$ построена с учетом ограничений на число типы входных и выходных переменных ;
- правила упрощения суперпозиций: g_u не принадлежит $\mathcal{G}$, если существует правило
$$r : g_u \mapsto g_v \in \mathcal{G}.$$

Результат

набор «композитных» признаков $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_n\}$.

Внимание! Число порожденных признаков может превосходить число клиентов!

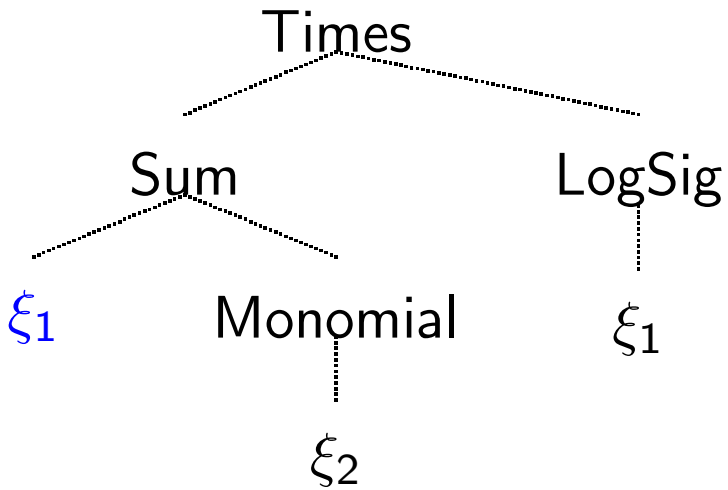
Примеры композитных признаков

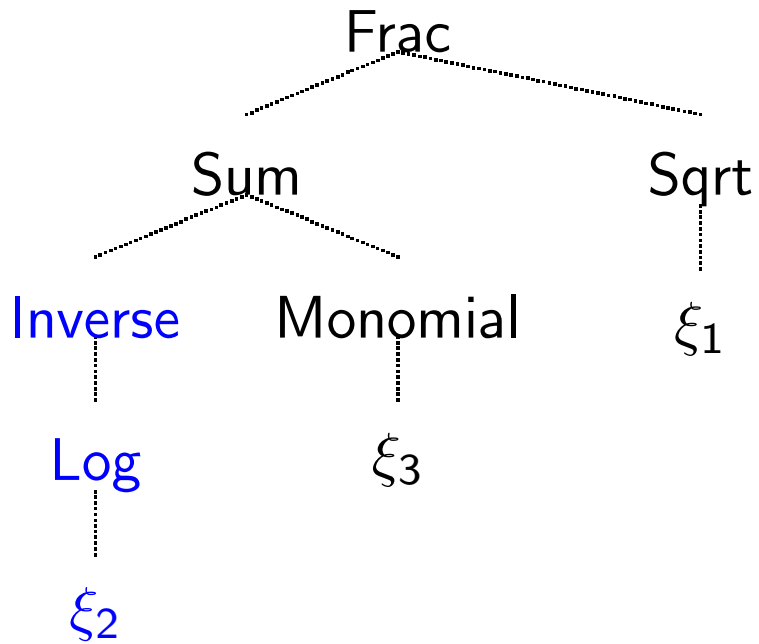
- **Frac**(Period of residence, Undeclared income)
- **Frac**(**Seg**(Period of employment), Term of contract)
- **And**(Income confirmation, Bank account)
- **Times**(**Seg**(Score hour), **Frac**(**Seg**(Period of employment), Salary))

Алгоритм случайного порождения признаков

1. Выбрать случайно узлы двух суперпозиций,
2. обменять соответствующие поддеревья,
3. изменить порождающую функцию на случайном узле.

Любые операции должны учитываю условия допустимости суперпозиций.





Структурные параметры и выбор моделей

Полный перебор порожденных обобщенных линейных моделей

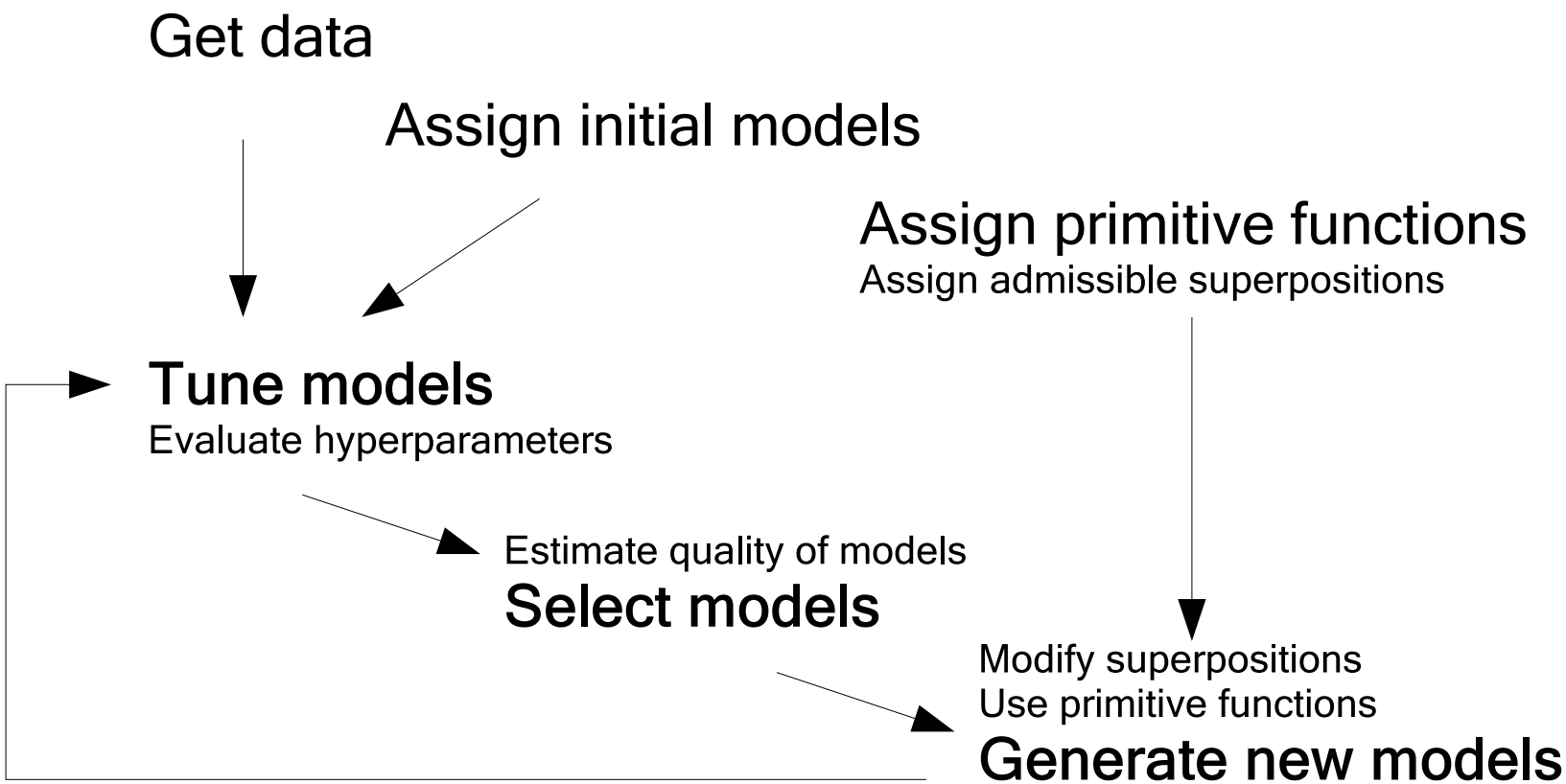
$$\mu(y) = w_0 + \alpha_1 w_1 x_1 + \alpha_2 w_2 x_2 + \ldots + \alpha_R w_R x_R.$$

Здесь $\alpha \in \{0, 1\}$ — структурный параметр.

Найти модель, заданную множеством индексов активных признаков $\mathcal{A} \subseteq \mathcal{J}$:

α_1	α_2	$\ldots$	$\alpha_{\lvert\mathcal{J}\rvert}$
1	0	$\ldots$	0
0	1	$\ldots$	0
$\ldots$	$\ldots$	$\ldots$	$\ldots$
1	1	$\ldots$	1

Процедура построения модели



Examples of nonparametric transformation functions

- Univariate

Formula	Output dimension
$\sqrt{x}$	1
$x\sqrt{x}$	1
$\arctan x$	1
$\ln x$	1
$x \ln x$	1

- Bivariate

Plus	$x_1 + x_2$
Minus	$x_1 - x_2$
Product	$x_1 \cdot x_2$
Division	$\frac{x_1}{x_2}$
	$x_1\sqrt{x_2}$
	$x_1 \ln x_2$

Nonparametric aggregation: sample statistics

Nonparametric transformations include basic data statistics:

- Sum or average value of each row $\mathbf{x}_i$, $i = 1, \ldots, m$:

$$\phi_i = \sum_{j=1}^{n} x_{ij}, \text{ or } \phi_i' = \frac{1}{n} \sum_{j=1}^{n} x_{ij}.$$

- Min and max values: $\phi_i = \min_j x_{ij}$, $\phi_i' = \max_j x_{ij}$.
- Standard deviation:

$$\phi_i = \frac{1}{n-1} \sqrt{\sum_{j=1}^{n} (x_{ij} - \text{mean}(\mathbf{x}_i))^2}.$$

- Data quantiles: $\phi_i = [X_1, \ldots, X_K]$, where

$$\sum_{j=1}^{n} [X_{k-1} < x_{ij} \leq X_k] = \frac{1}{K}, \text{ for } k = 1, \ldots, K.$$

Nonparametric transformations: Haar's transform

Applying Haar's transform produces multiscale representations of the same data.

Assume that $n = 2^K$ and init $\phi_{i,j}^{(0)} = \phi_{i,j}'^{(0)} = x_{ij}$ for $j = 1, \ldots, n$.

To obtain coarse-graining and fine-graining of the input feature vector $\mathbf{x}_i$, for $k = 1, \ldots, K$ repeat:

- data averaging step

$$\phi_{i,j}^{(k)} = \frac{\phi_{i,2j-1}^{(k-1)} + \phi_{i,2j}^{(k-1)}}{2}, \quad j = 1, \ldots, \frac{n}{2^k},$$

- and data differencing step

$$\phi_{i,j}'^{(k)} = \frac{\phi_{i,2j}'^{(k-1)} - \phi_{i,2j-1}'^{(k-1)}}{2}, \quad j = 1, \ldots, \frac{n}{2^k}.$$

The resulting multiscale feature vectors are $\phi_i = [\phi_i^{(1)}, \ldots, \phi_i^{(K)}]$ and $\phi_i' = [\phi_i'^{(1)}, \ldots, \phi_i'^{(K)}]$.

Examples of parametric transformation functions

Function name	Formula	Output dim.	Num. of args	Num. of pars
Add constant	$x + w$	1	1	1
Quadratic	$w_2 x^2 + w_1 x + w_0$	1	1	3
Cubic	$w_3 x^3 + w_2 x^2 + w_1 x + w_0$	1	1	4
Logarithmic sigmoid	$1/(w_0 + \exp(-w_1 x))$	1	1	2
Exponent	$\exp x$	1	1	0
Normal	$\frac{1}{w_1\sqrt{2\pi}} \exp\left(\frac{(x-w_2)^2}{2w_1^2}\right)$	1	1	2
Multiply by constant	$x \cdot w$	1	1	1
Monomial	$w_1 x^{w_2}$	1	1	2
Weibull-2	$w_1 w_2 x^{w_2-1} \exp -w_1 x^{w_2}$	1	1	2
Weibull-3	$w_1 w_2 x^{w_2-1} \exp -w_1 (x - w_3)^{w_2}$	1	1	3
...	...	...	...	...

Monotone functions

- **By grow rate**

Function name	Formula	Constraints
Linear	$w_1 x + w_0$	
Exponential rate	$\exp(w_1 x + w_0)$	$w_1 > 0$
Polynomial rate	$\exp(w_1 \ln x + w_0)$	$w_1 > 1$
Sublinear polynomial rate	$\exp(w_1 \ln x + w_0)$	$0 < w_1 < 1$
Logarithmic rate	$w_1 \ln x + w_0$	$w_1 > 0$
Slow convergence	$w_0 + w_1/x$	$w_1 \neq 0$
Fast convergence	$w_0 + w_1 \cdot \exp(-x)$	$w_1 \neq 0$

- **Other**

Soft ReLu	$\ln(1 + e^x)$	
Sigmoid	$1/(w_0 + \exp(-w_1 x))$	$w_1 > 0$
Softmax	$1/(1 + \exp(-x))$	
Hiberbolic tangent	$\tanh(x)$	
softsign	$\frac{\lvert x \rvert}{1+\lvert x \rvert}$	

Parametric transformations

Optimization of the transformation function parameters $\mathbf{b}$ is iterative:

1. Fix the vector $\hat{\mathbf{b}}$, collected over all the primitive functions $\{g\}$, which generate features ϕ:

$$\hat{\mathbf{w}} = \arg\min S\big(\mathbf{w}|\mathbf{f}(\mathbf{w},\mathbf{x}),\mathbf{y}\big), \quad \text{where} \quad \phi(\hat{\mathbf{b}},\mathbf{s}) \subseteq \mathbf{x}.$$

2. Optimize transformation parameters $\hat{\mathbf{b}}$ given model parameters $\hat{\mathbf{w}}$

$$\hat{\mathbf{b}} = \arg\min S\big(\mathbf{b}|\mathbf{f}(\hat{\mathbf{w}},\mathbf{x}),\mathbf{y}\big).$$

Repeat these steps until vectors $\hat{\mathbf{w}}, \hat{\mathbf{b}}$ converge.

Parameters of the local models

More feature generation options:

- Parameters of SSA approximation of the time series $\mathbf{x}^{(q)}$.
- Parameters of the FFT of each $\mathbf{x}^{(q)}$.
- Parameters of polynomial/spline approximation of each $\mathbf{x}^{(q)}$.

Metric features: distances to the centroids of local clusters

Apply kernel trick to the time series.

1. For given local feature vector $\mathbf{x}_i^{(q)}$, $q = 1, \ldots, Q$ compute k-means centroids $\mathbf{c}_p^{(m)}$, $p = 1, \ldots, P$.
2. With the selected k-means distance function ρ construct the feature vector

$$\phi_i^{(q)} = [\rho(\mathbf{c}_1^{(q)}, \mathbf{x}_i^{(q)}), \ldots, \rho(\mathbf{c}_P^{(q)}, \mathbf{x}_i^{(q)})] \in \mathbb{R}_+^P.$$

The procedure may be applied to each $\mathbf{x}^{(q)}$ or directly to the $\mathbf{x} = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(Q)}]$, resulting in only P additional features instead of $Q \cdot P$