

My first scientific paper

Week 4

Plan the experiment

Vadim Strijov

Moscow Institute of Physics and Technology

2021

Вычислительный эксперимент

Цель эксперимента

Восстановить плотность пространственной конфигурации пары аминокислота-лиганд с помощью предложенного монотонного многослойного перцептрона.

Данные

В базе данных 20 аминокислотных остатков. Аминокислотные остатки взаимодействуют с лигандами, их 40. Взаимодействие описывается 5 признаками: тип аминокислотного остатка, тип лиганда, расстояние и 2 угла θ и φ .

Гипотеза

Максимумы полученной плотности распределения соответствуют устойчивым пространственным конфигурациям.

Цель

Восстановить плотности распределения пространственных ориентаций различных пар вида аминокислота-лиганд.

Данные

Данные представляют собой 47916041 пятерку значений, элементы каждой пятерки: a — индекс аминокислоты, b — индекс лиганда и тройка r, θ, φ . Индексы аминокислоты и лиганда образуют 840 пар и используются для разделения данных на 840 выборок (r, θ, φ) , каждая из которых соответствует своей взаимодействующей паре.

Описание эксперимента

Для каждой из 840 выборок строится восстановленная плотность $\hat{p}^{a,b}(r, \theta, \varphi) = p(r, \theta, \varphi | \mathbf{w}^*, \mathbf{U}^*)$.

Наборы данных

Набор	Нативные структуры	Модели структур	Разбиение
CASP 9	117	35963	Train, Validation
CASP 10	103	15450	
CASP 11	84	12291	
CASP 12	37	5501	Test

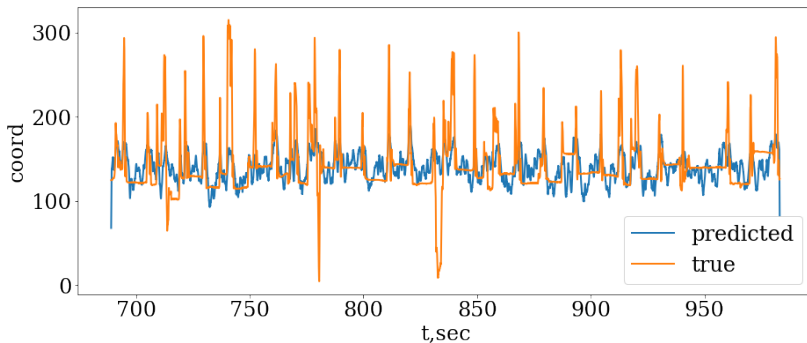
При обучении нейросети анализируются усредненные по T нативным структурам коэффициенты корреляции Пирсона и Спирмена

$$R = R(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{T} \sum_{i=1}^T R_i^{\text{target}} = \frac{1}{T} \sum_{i=1}^T \text{PEARSON}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

$$\rho = \rho(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{T} \sum_{i=1}^T \rho_i^{\text{target}} = \frac{1}{T} \sum_{i=1}^T \text{SPEARMAN}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

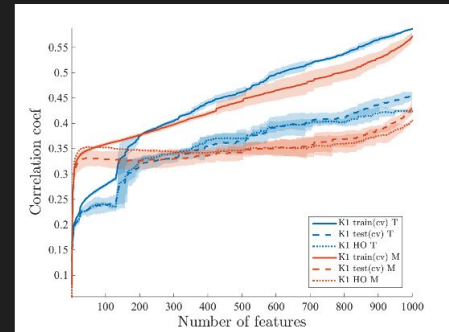
Описание экспериментов

- Требуется восстановить координату движения конечности по сигналу электрокортикограммы.
- 5 временных рядов по 20 минут, первые 15 минут обучение, остальные 5 минут — тест.
- Критерий качества: коэффициент корреляции между предсказанной траекторией и истинной.



Зависимость предсказанной и истинной траекторий от времени

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_fr...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



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



Планирование коммерческих проектов (m1p.org)

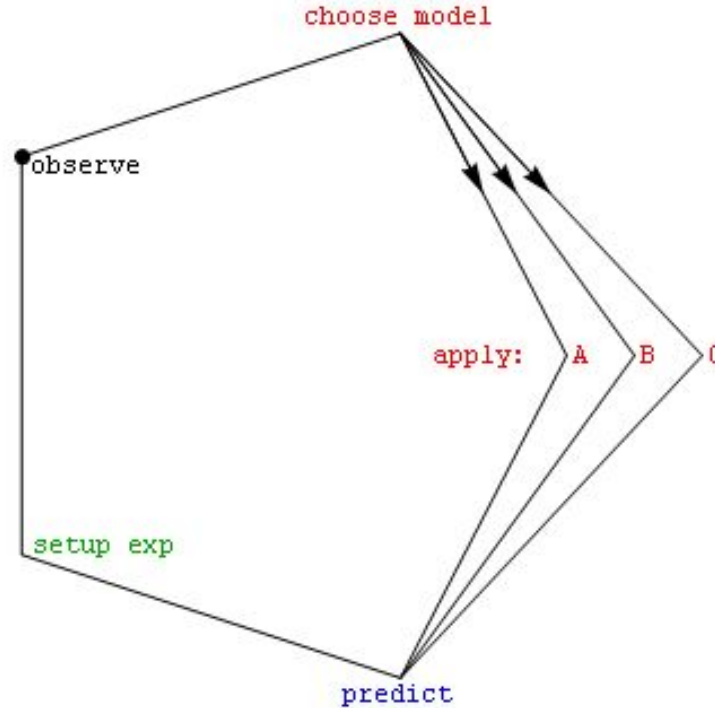
Научный проект: исследует свойства объекта, решает узкую задачу

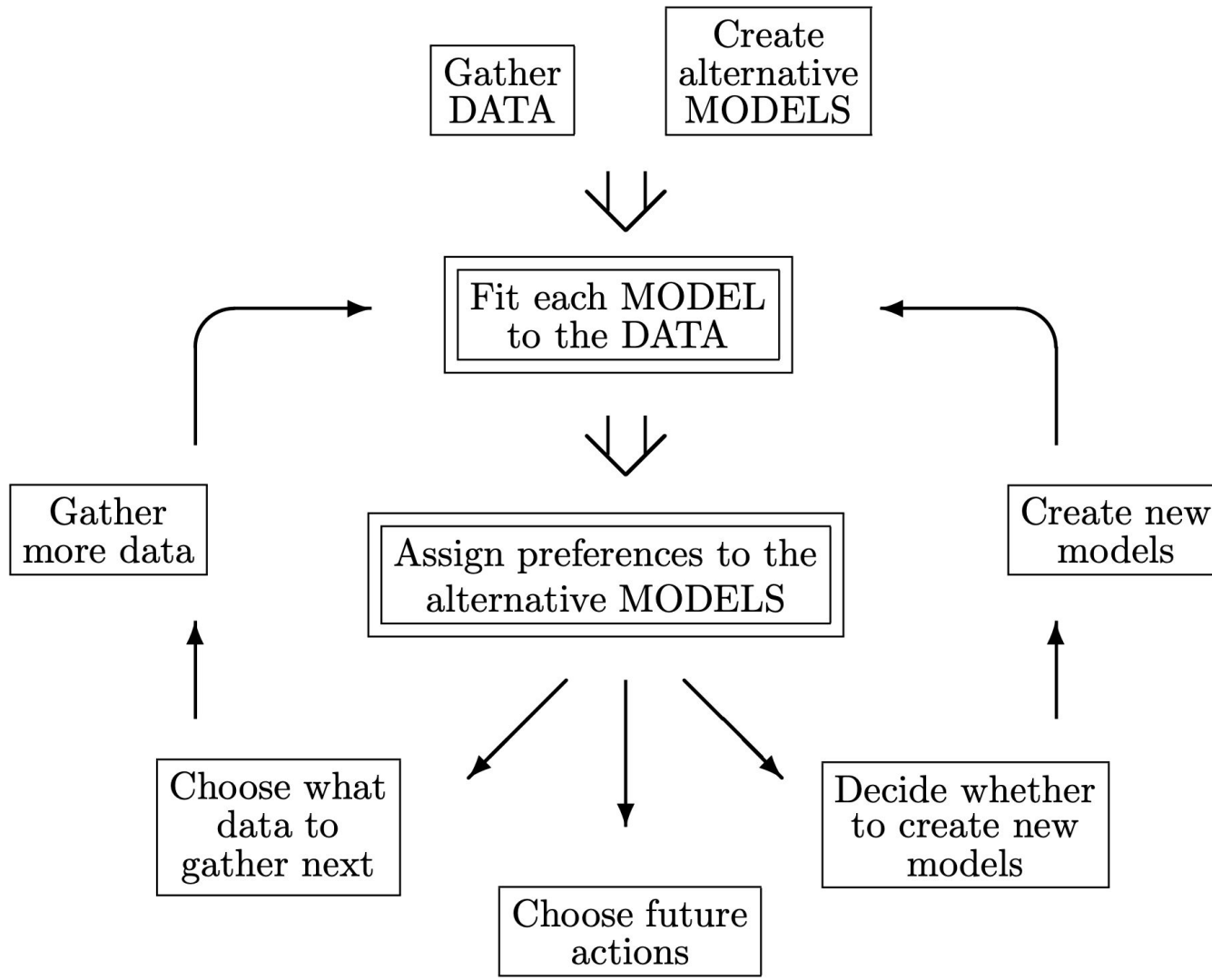
Коммерческий проект: создает систему, продукт

Правила проектирования:

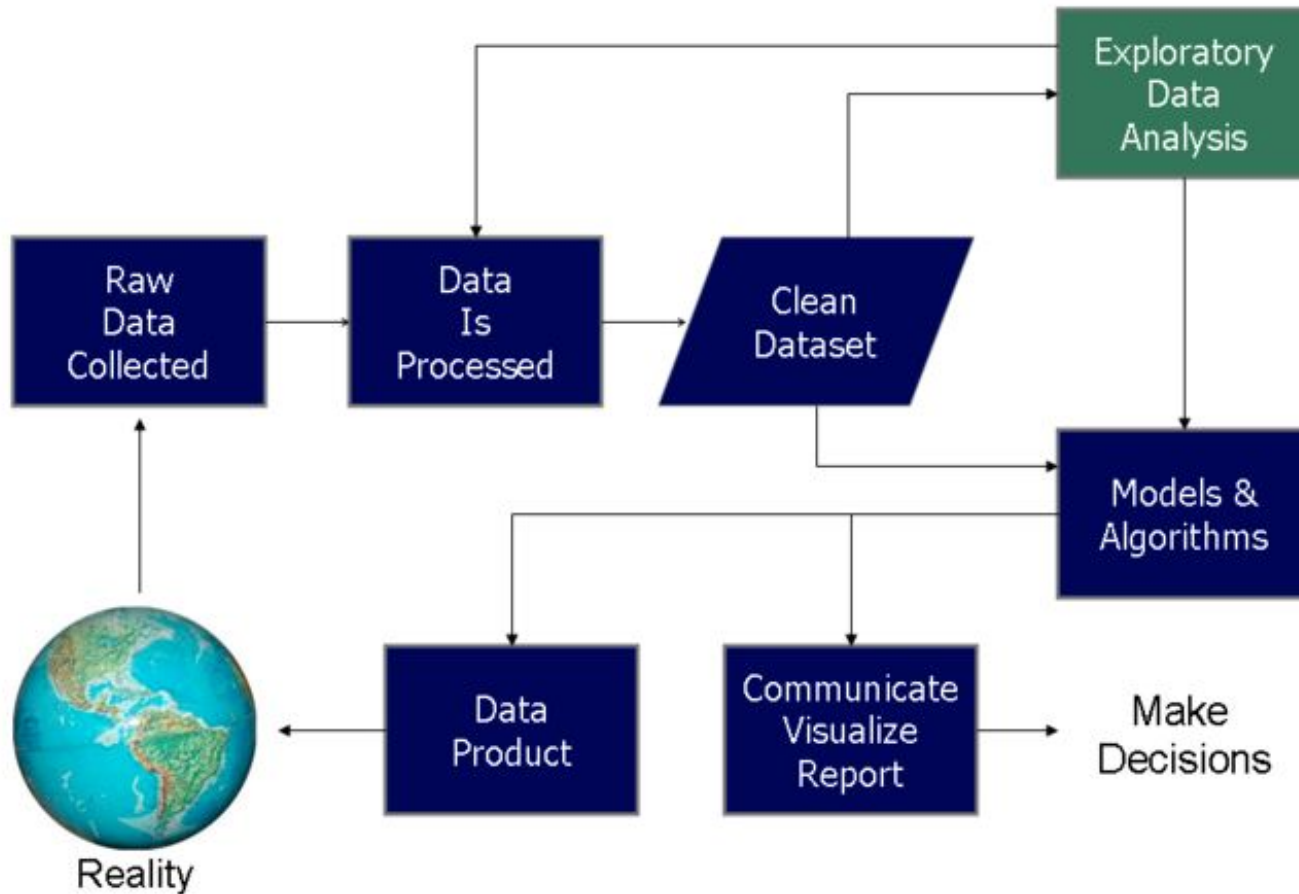
- 1) система полна,
- 2) система иерархична,
- 3) система имеет обратную связь

The scientific observation cycle (model selection)





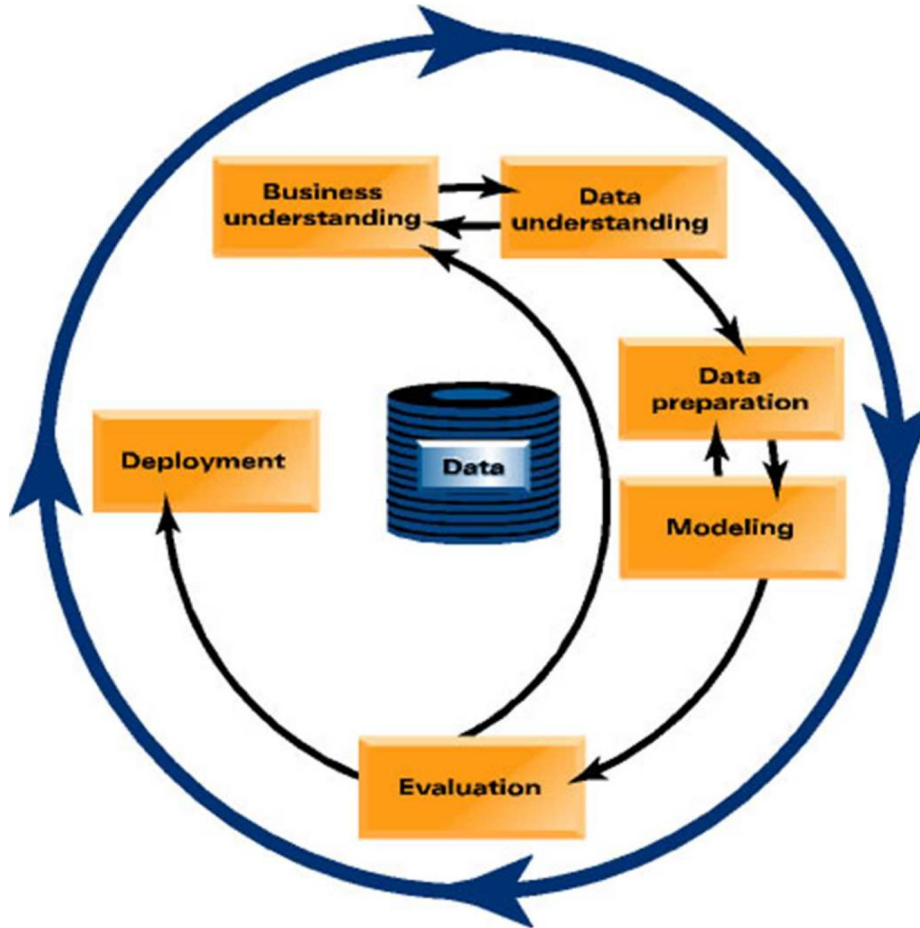
Data Science Process



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

Six major phases:

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



Project life cycle



Р 50.1.028—2001

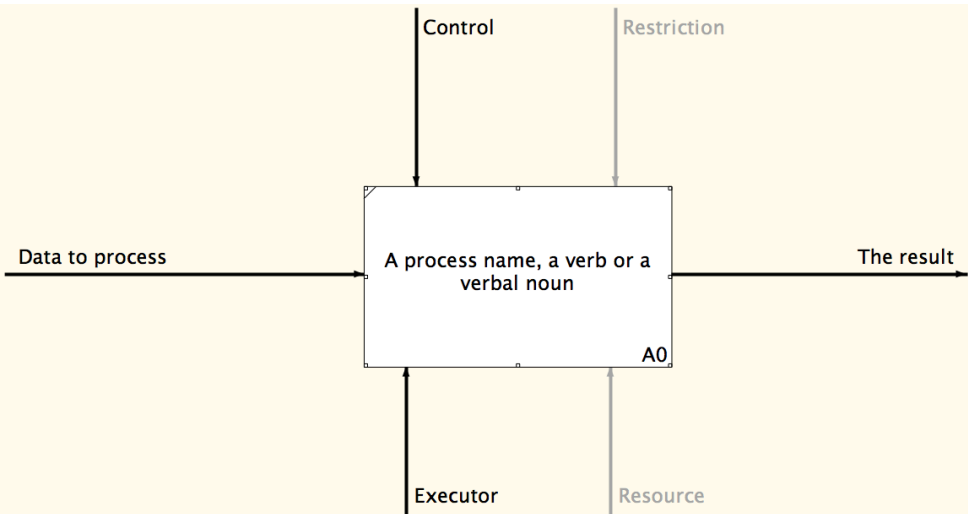
РЕКОМЕНДАЦИИ ПО СТАНДАРТИЗАЦИИ

**Информационные технологии поддержки жизненного
цикла продукции**

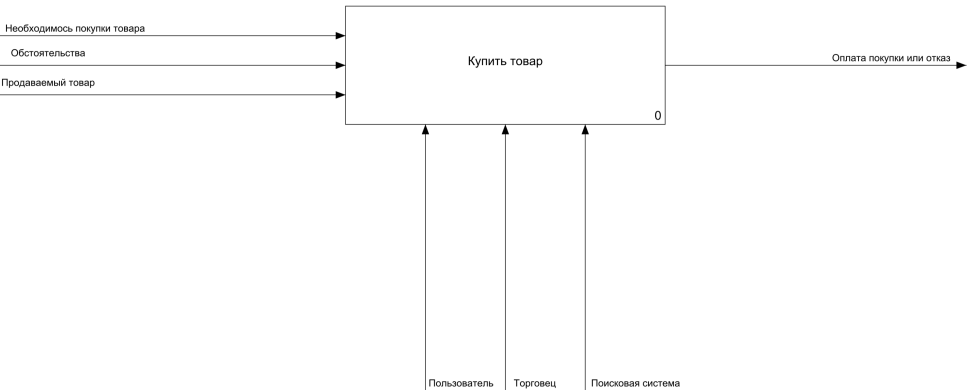
**МЕТОДОЛОГИЯ ФУНКЦИОНАЛЬНОГО
МОДЕЛИРОВАНИЯ**

Издание официальное

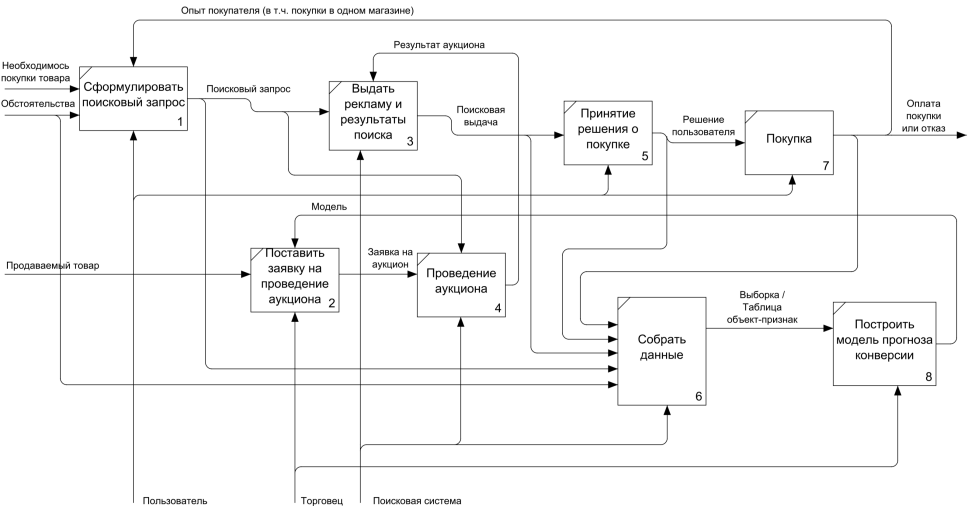
Функциональная схема IDEF0, лист A-0



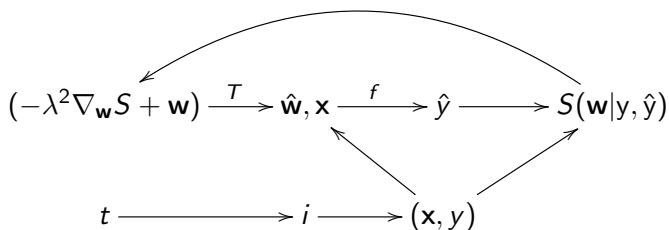
Непрерывный аукцион в поисковых системах, лист А-0



Непрерывный аукцион в поисковых системах, лист А0



The simplest problem statement in machine learning



f is the forecasting model,

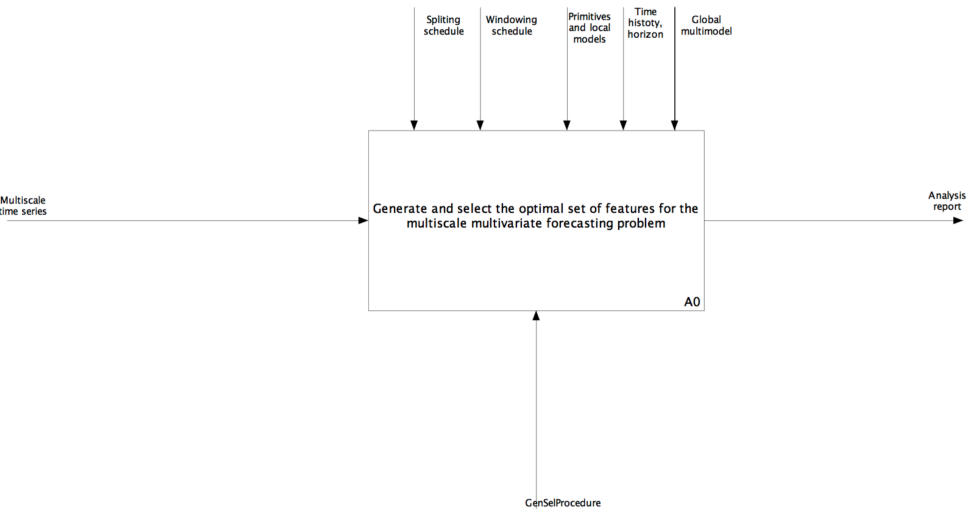
S is the criterion,

T is an optimization algorithm,

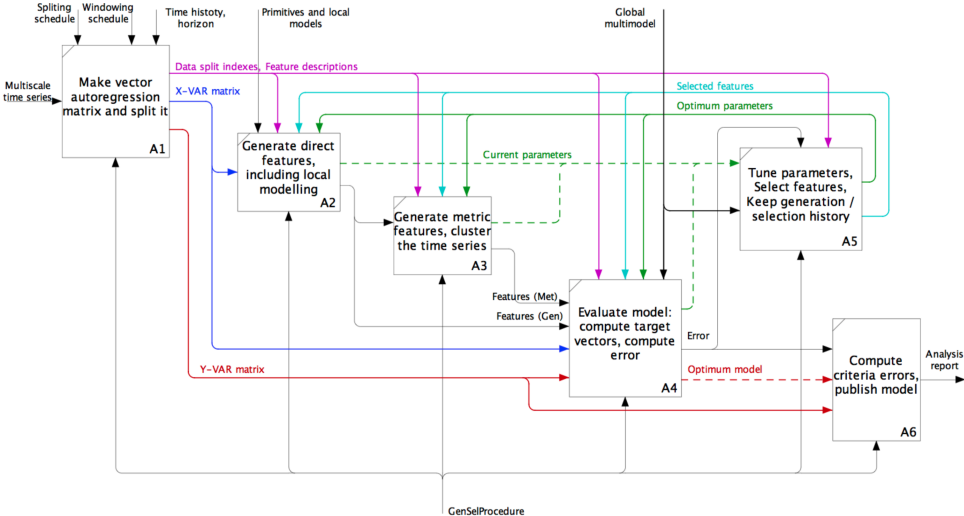
$\hat{\mathbf{w}}$ is some solution,

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

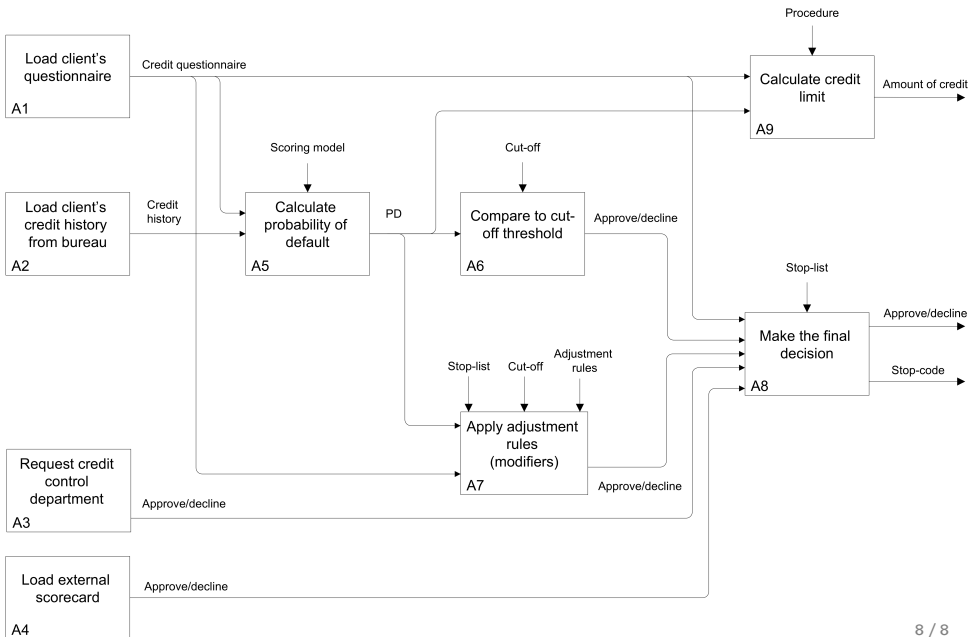
Порождение и выбор моделей, лист A-0



Порождение и выбор моделей, лист A0

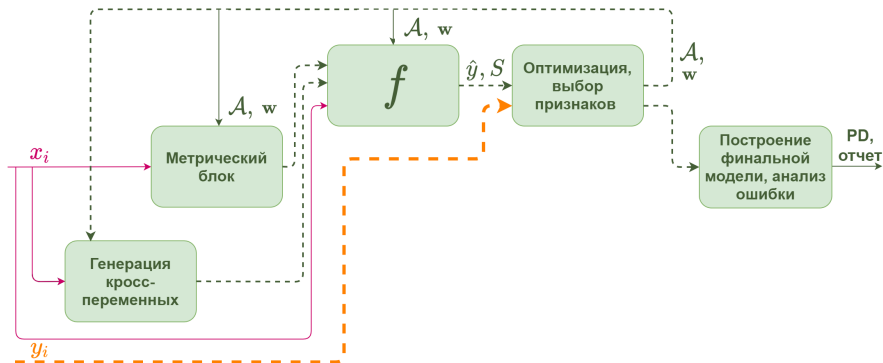


Функциональная схема скоринговой карты



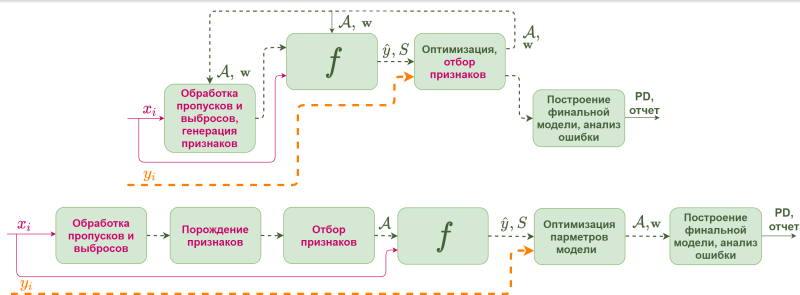
Идея. Новизна

Система построена в единой оптимизационной цепочке с использованием модуля построения модели: процедуры обработки пропусков и выбросов, порождения и отбора признаков выполняются не на этапах предобработки, а в ходе оптимизации



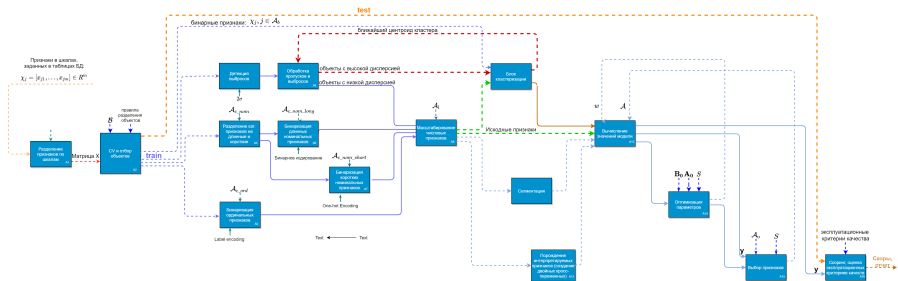
Уникальность модели как скоринговой системы

- Модель как суперпозиция локальных моделей
- Адаптивная, итеративная система (формирование \mathcal{A} на каждой итерации обеспечивает адаптивность)
- Структура модели определяет способность системы накапливать информацию об опорных клиентах и информативных признаках
- Модель не требует разбиения данных на обучающую и контрольную выборку



Сравнение предложенной модели (сверху) и базового решения (снизу)

Универсальная модель как суперпозиция локальных моделей



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. Эмбединг
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)