Постановка задач и выбор моделей в машинном обучении

Вадим Викторович Стрижов

Московский физико-технический институт

Осенний семестр 2019

Цель: научиться ставить задачи в области машинного обучения в индустриальных и научных проектах

Семинары включают:

- теорию выбора моделей и анализа ошибки,
- разбор примеров постановок задач,
- обсуждение методов и инструментов планирования при постановке и решении задач.

Семинары не включают:

- разбор отдельных алгоритмов машинного обучения,
- разбор кода на питоне и других языках,

потому что курсы и материалы по этим темам уже имеются в достаточном количестве.

Подробнее см. http://www.machinelearning.ru/wiki/index.php?title=M1

Тематический план

- Модели и их порождение
- Функия ошибки и выбор моделей
- Снижение размерности, порождение признаков
- Метрическое обучение
- Обучение представлений и вложения
- Планирование эксперимента и анализ ошибки

Домашние задания

- Анкета с вопросами на понимание пройденного материала и на подготовку к следующему занятию
- Основное задание — постановка задачи, решение задачи, либо вычислительный эксперимент
- Рецензирование работ — их просмотр и ранжирование

Робот AutoML (сейчас в режиме отладки)

1. Собирает проекты из папок GitHub, огранизация phystech-intelligent-systems.
2. Рассылает ссылки на проекты рецензентам (вам, слушателям этого курса).
3. Каждый рецензент получает от трех до пяти ссылок на проекты и ранжирует их по качеству, заполняя анекту (так же, как это делает Курсера).
4. Рассылает оценки за домашнее задание в конце недели.

Оценивание и дедлайны

- Анкета с вопросами: 0–10 баллов (зависит от числа верных ответов)
- Основное задание: 50–100 баллов (зависит от качества постановки, решения задачи или эксперимента)
- Рецензирование: 10 баллов (просмотр и ранжирование работ)

$$\text{Оценка} = \frac{1}{100}\text{суммы баллов}.$$

Робот AutoML

1) собирает выполненные задания в четверг 9:00 и
2) рассылает оценки в пятницу в 9:00.

Других дедлайнов не предполагается.

Коммуникации

Домашние задания по короткому адресу **bit.ly/PS-ML**

Вопросы по почте **MLalgorithms@gmail.com**

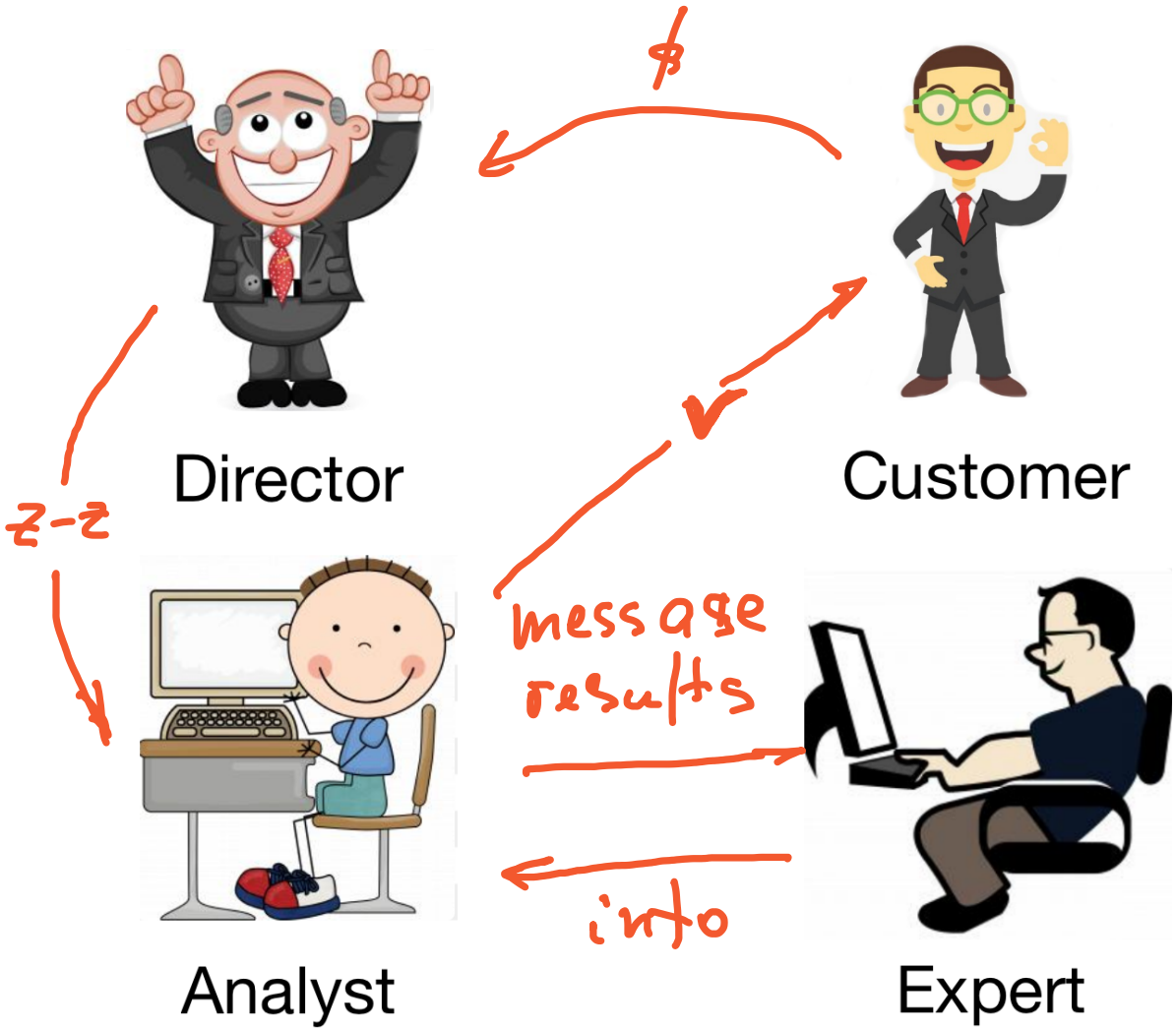
Потеря информации при передаче сообщения[1]

Небольшая группа программистов работает над новым проектом. Сколько времени пройдет, прежде чем

1) в группе выработается свой уникальный лабораторный жаргон,
2) новый сотрудник сможет разобраться, чем занимается группа,
3) руководитель группы перестанет понимать ход проекта,
4) каждый член группы перестает понимать, чем занимаются его коллеги?

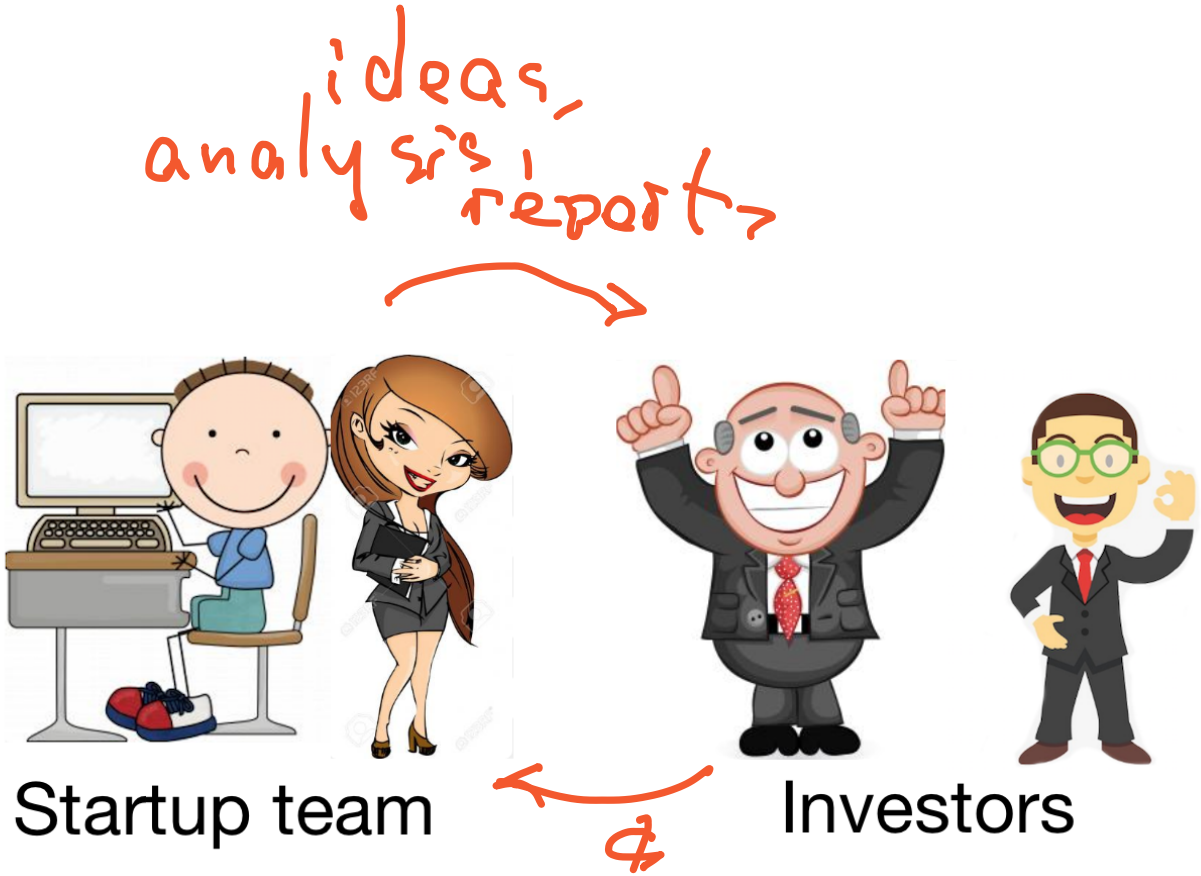
[1] в отсутствие планирования

Исследователь-аналитик в коммерческой компании



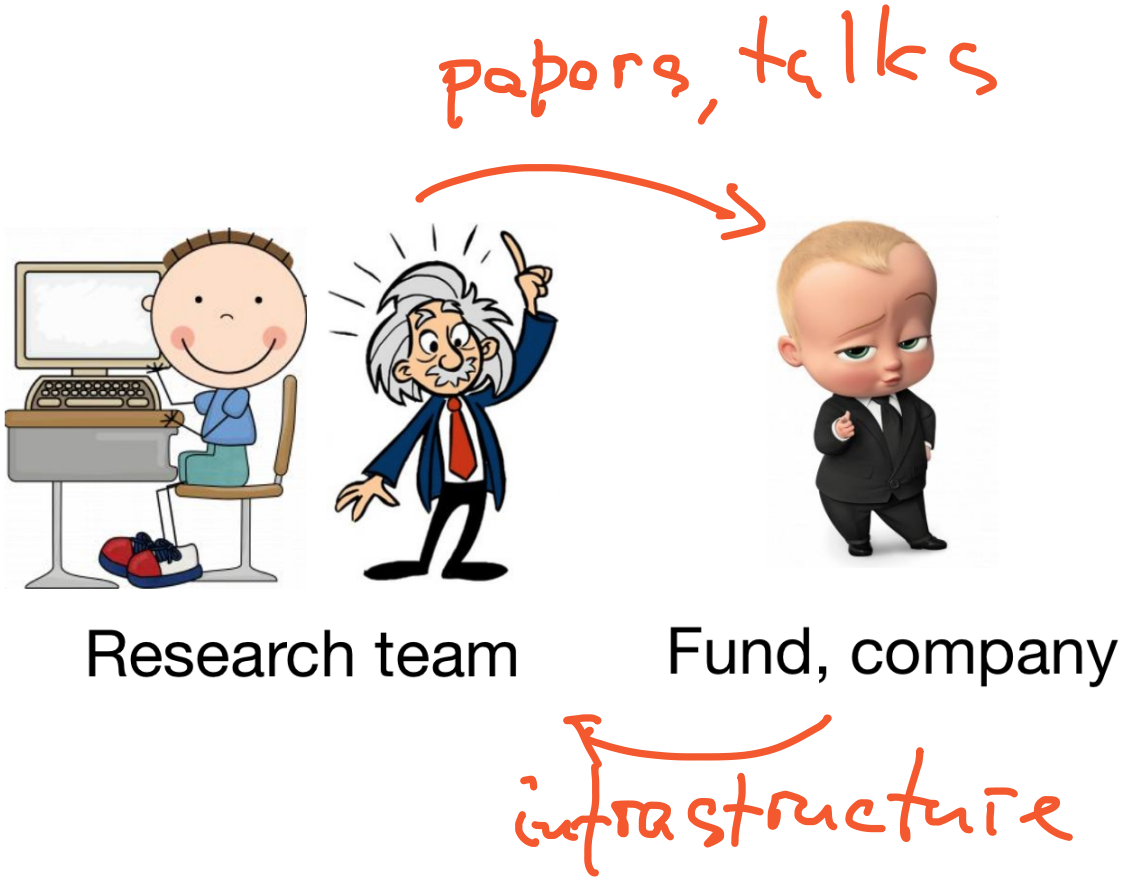
The source:http://rpelm.com

Исследователь-аналитик в стартапе



The source:http://rpelm.com

Исследователь-аналитик в научной группе



The source:http://rpelm.com

Построение скоринговых вероятностных моделей как прикладная задача классификации

- Выдача кредита (Application scoring)
- Динамика состояния (Behavioral scoring)
- Просроченная задолженность (Collection scoring)

Типы кредитов для физических лиц:

- Потребительский (POS)
- Кредит наличными
- Автокредит
- Ипотечный

Типичное число клиентских записей в базе данных:

- $\sim 10^4$ для «тяжелых» долгосрочных кредитов,
- $\sim 10^6$ для «легких» кредитов,
- $\sim 10^7$ для банковских карт.

Процедура получения кредита с точки зрения банка

Заявка-анкета, кредитная история

↓

Скоринговый балл: вероятность невозврата (PD)

Вариант: кредитный контроль

↓

Назначение кредитного лимита

↓

Сообщение о принятии заявки или об отказе

↓

Заключение договора

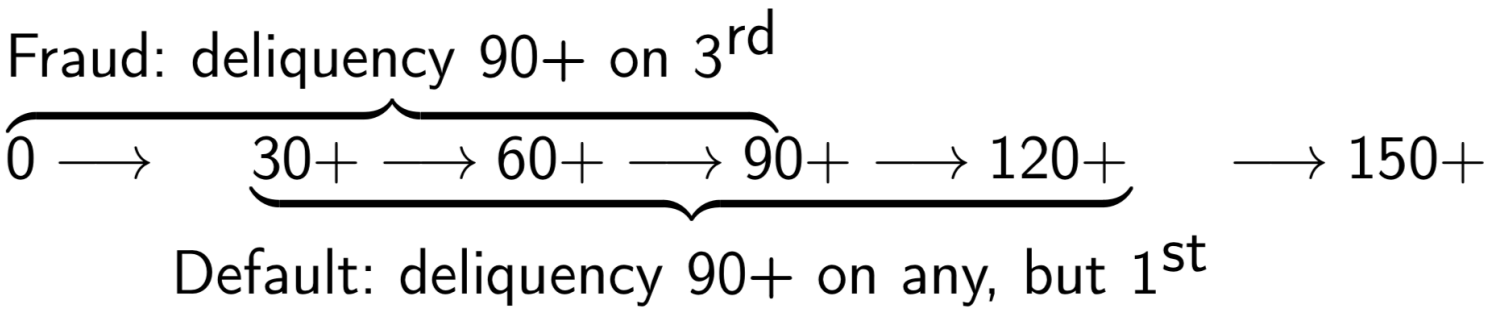
↓

...

↓

Появление кредитной истории

Виды просрочек возврата кредита



- Fraud — мошенничество
- Default — возврат кредита просрочен

Потери от просрочек возврата потребительского кредита

Примерная просрочка (от недели и выше) по потребительским кредитам на некоторый момент времени

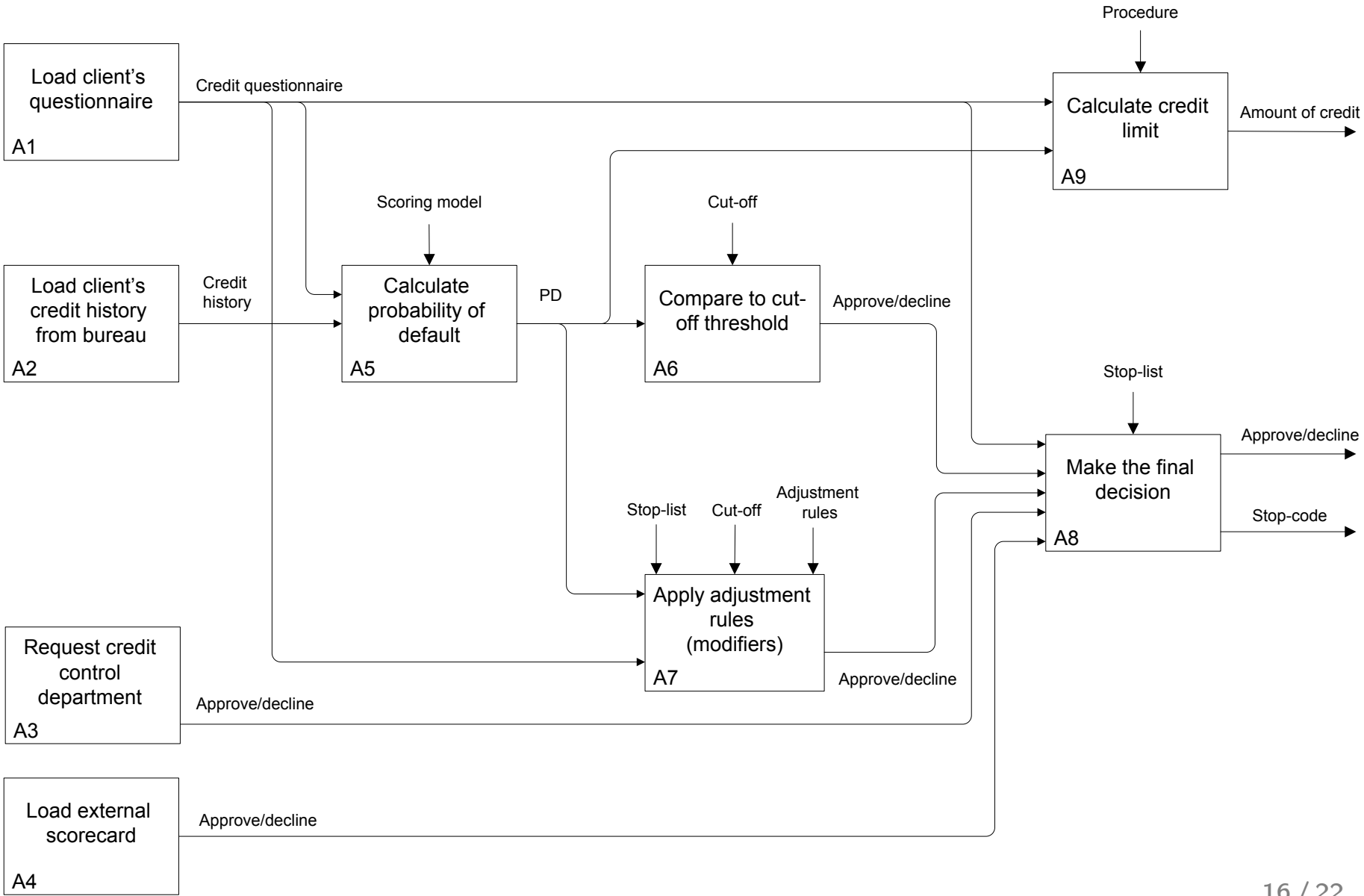
Категория	Количество	Сумма
Все категории товаров	100 000	2 100 М
Бытовая техника	30 000	350 М
Мебель	20 000	300 М
Одежда	15 000	200 М
Телевизоры	10 000	100 М
Мобильные телефоны	15 000	80 М
Фотоаппараты	2 000	20 М

Причины отказа в выдаче кредита

Некоторые типичные причины:

- недостаточный скоринговый балл,
- не прошел кредитный контроль,
- в черном списке банка,
- просрочка по данным бюро кредитных историй,
- не гражданин России,
- маленький личный доход,
- клиент моложе (старше) определенного возраста и сумма слишком велика,
- мобильный телефон найден у другого клиента.

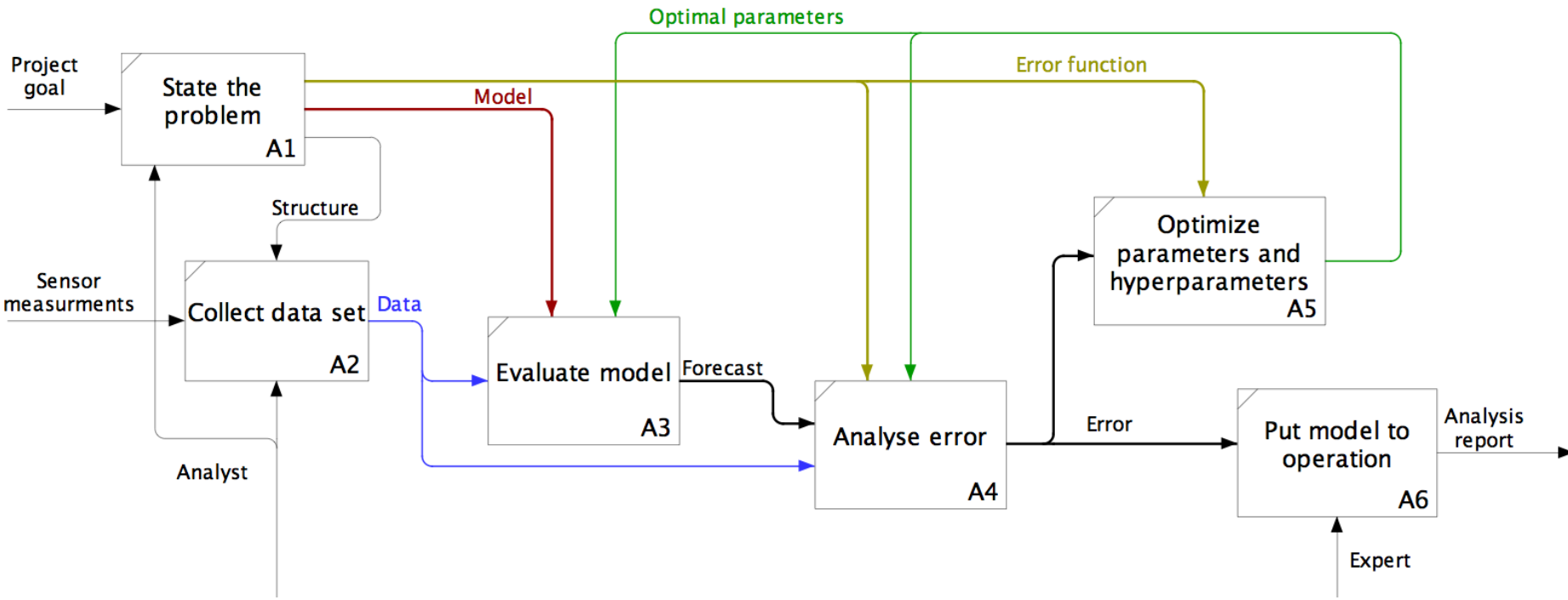
Функциональная схема IDEF0 скоринговой карты с вероятностной моделью классификации



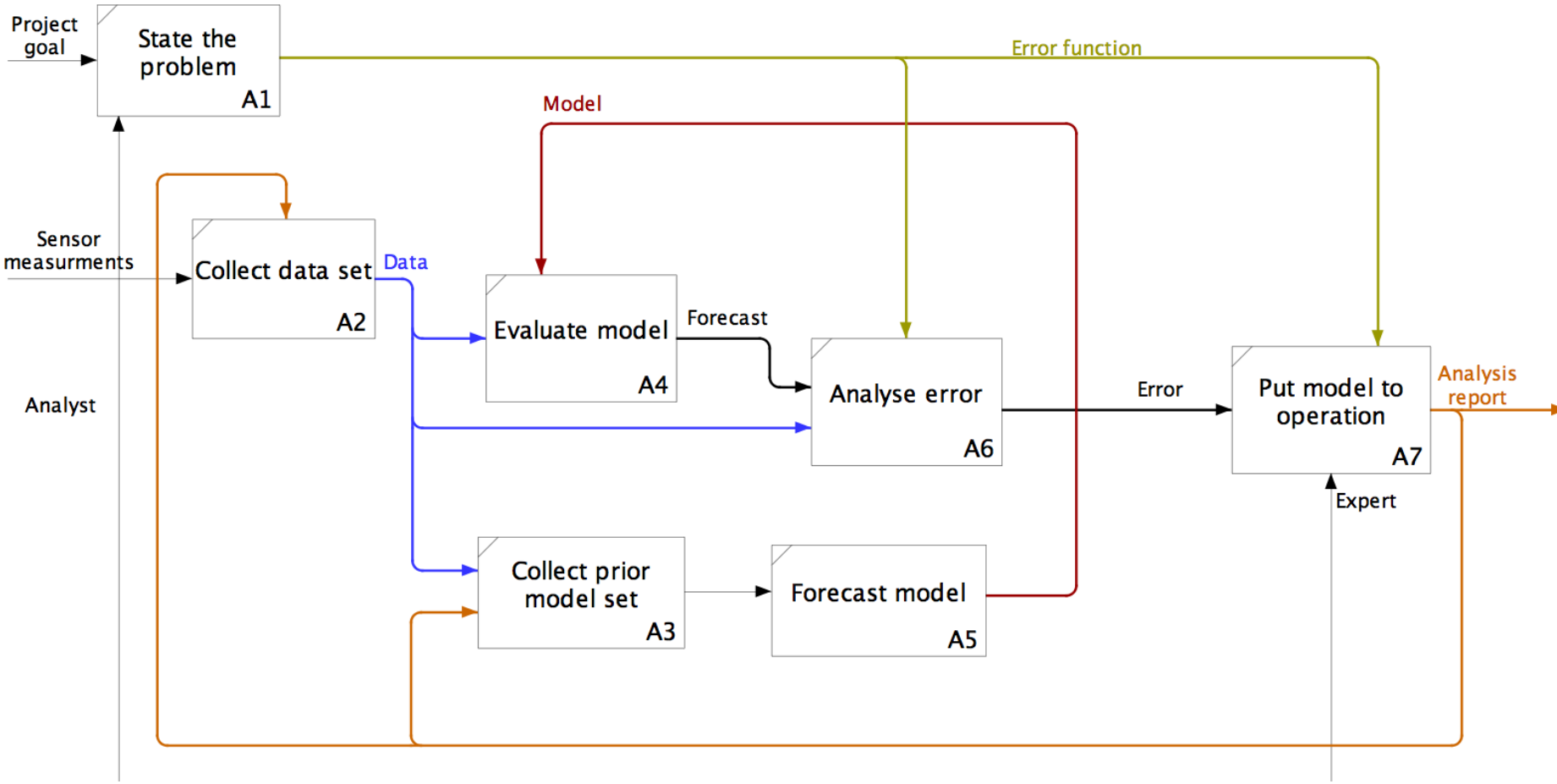
До начала планирования исследования аналитик и (эксперт) обсуждают ключевые вопросы

1. Цель проекта. (Ожидаемый результат разработки.) **Ожидаемая цель исследования.**
2. Прикладная задача, решаемая в проекте. (Как результат будет использован?) **Чем результат будет проиллюстрирован?**
3. Описание исторических измеряемых данных. (Форматы и тайминг.) **Алгебраическая структура данных.**
4. Критерии качества. (Как измеряется качество полученного результата, что будет в отчете?) **Функция ошибки, что будем оптимизировать.**
5. Выполнимость проекта. (Как показать, что проект выполним, список возможных рисков.) **План анализа ошибки.**
6. Условия, необходимые для успешного выполнения проекта. (Организация работ.) **Требования к выборке.**
7. Методы решения. (Библиотеки процедур.) **Поставленные гипотезы, оптимальные вероятностные модели.**

Аналитик создает модель, чтобы эксперт ее использовал



Модель выбирается из множества допустимых моделей



Парадигма глубокого обучения: уменьшение роли аналитика в создании оптимальной модели и автоматизация научных исследований в машинном обучении.

НИР или ОКР? Новизна или технологичность

Эксперт:

(Как долго будет эксплуатироваться модель? Что заменит ее в дальнейшем?)

Аналитик:

Какое влияние окажет исследование на область знаний? Насколько она будет полезна?

За какую задачу браться?

- Масштабность: решение задачи должно влиять на большое число людей, специалистов, лиц принимающих решения.
- Заброшенность (популярность) задачи. Общая ошибка: решать популярные задачи.
- Решаемость задачи. Следует выбирать просто и элегантно решаемые задачи.
- Наша квалификация и готовность к решению: похожие задачи мы уже решали.

Case 1. Energy consumption and price forecasting, 1-day ahead hourly

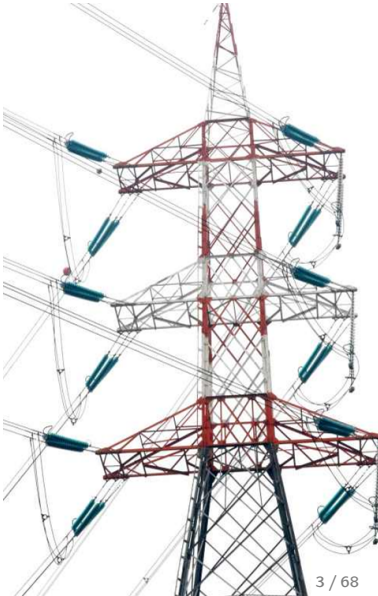
The components of multivariate time series with periodicity

Time series:

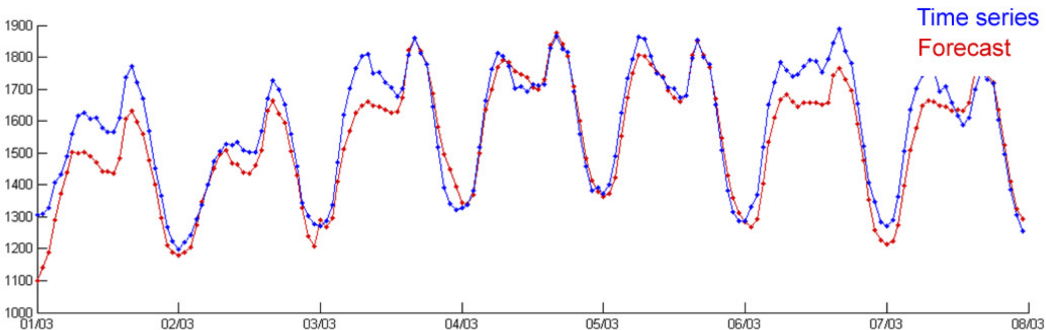
- energy price,
- consumption,
- daytime,
- temperature,
- humidity,
- wind force,
- holiday schedule.

Periodicity:

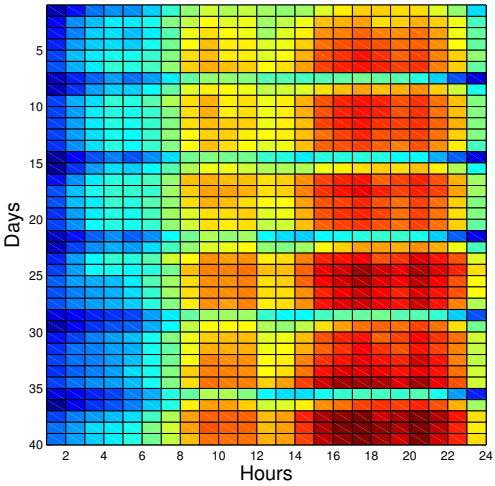
- one year seasons (temperature, daytime),
- one week,
- one day (working day, week-end),
- a holiday,
- aperiodic events.



Energy consumption one-week forecast for each hour



The autoregressive matrix, five weeks



The autoregressive matrix and the linear model

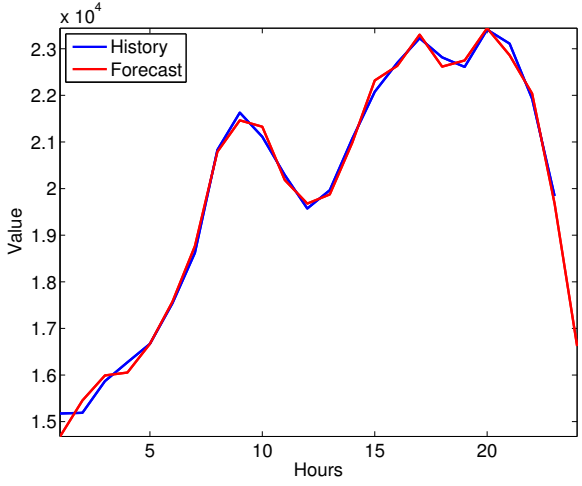
$$\underset{(m+1)\times(n+1)}{\mathbf{X}^*} = \left[\begin{array}{c|ccc} \hat{s}_T & s_{T-1} & \cdots & s_{T-\kappa+1} \\ \hline s_{(m-1)\kappa} & s_{(m-1)\kappa-1} & \cdots & s_{(m-2)\kappa+1} \\ \cdots & \cdots & \cdots & \cdots \\ s_{n\kappa} & s_{n\kappa-1} & \cdots & s_{n(\kappa-1)+1} \\ \cdots & \cdots & \cdots & \cdots \\ s_{\kappa} & s_{\kappa-1} & \cdots & s_1 \end{array}\right] = \left[\begin{array}{c|c} \underset{1\times 1}{\hat{s}_T} & \underset{1\times n}{\mathbf{x}_{m+1}} \\ \hline \underset{m\times 1}{\mathbf{y}} & \underset{m\times n}{\mathbf{X}} \end{array}\right].$$

In terms of linear regression:

$$\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w}) = \mathbf{X}\mathbf{w},$$

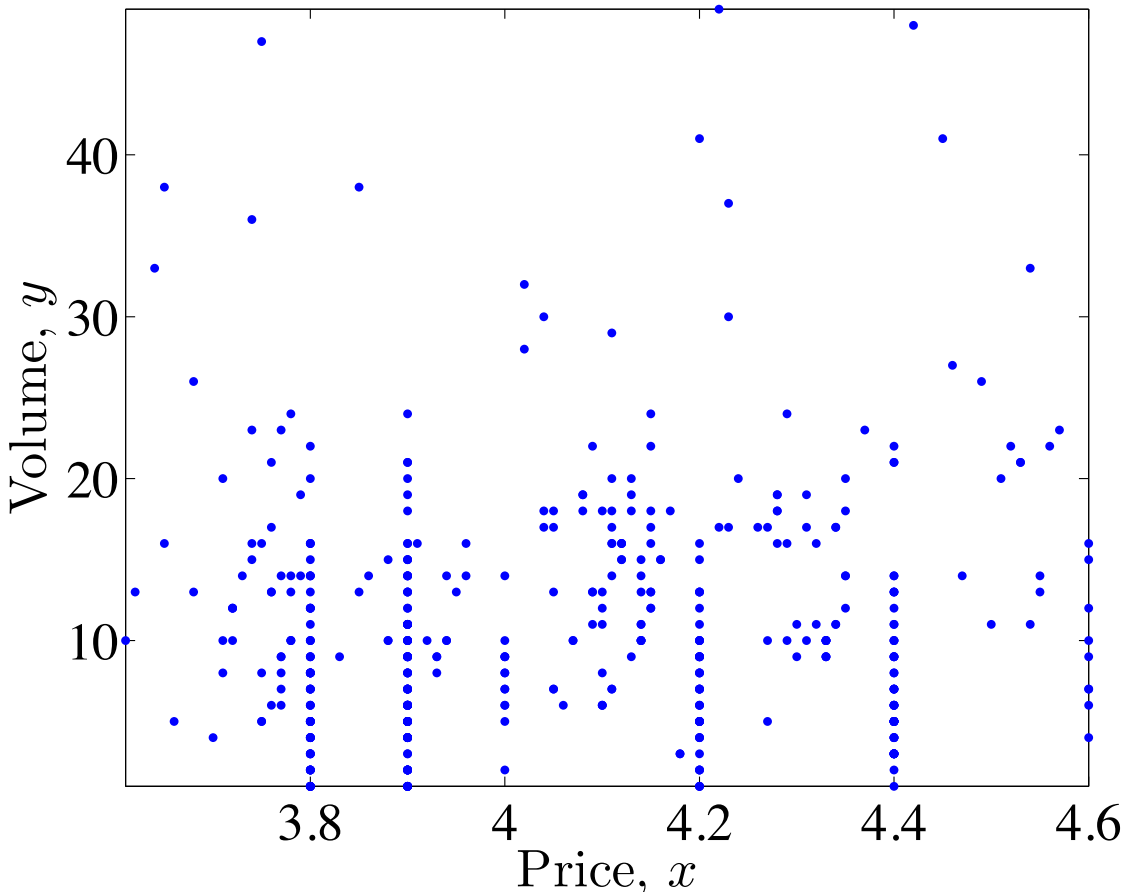
$$\hat{y}_{m+1} = \hat{s}_T = \langle \mathbf{x}_{m+1}, \hat{\mathbf{w}} \rangle.$$

The one-day forecast: expected error is 3.1% working day, 3.7% week-end

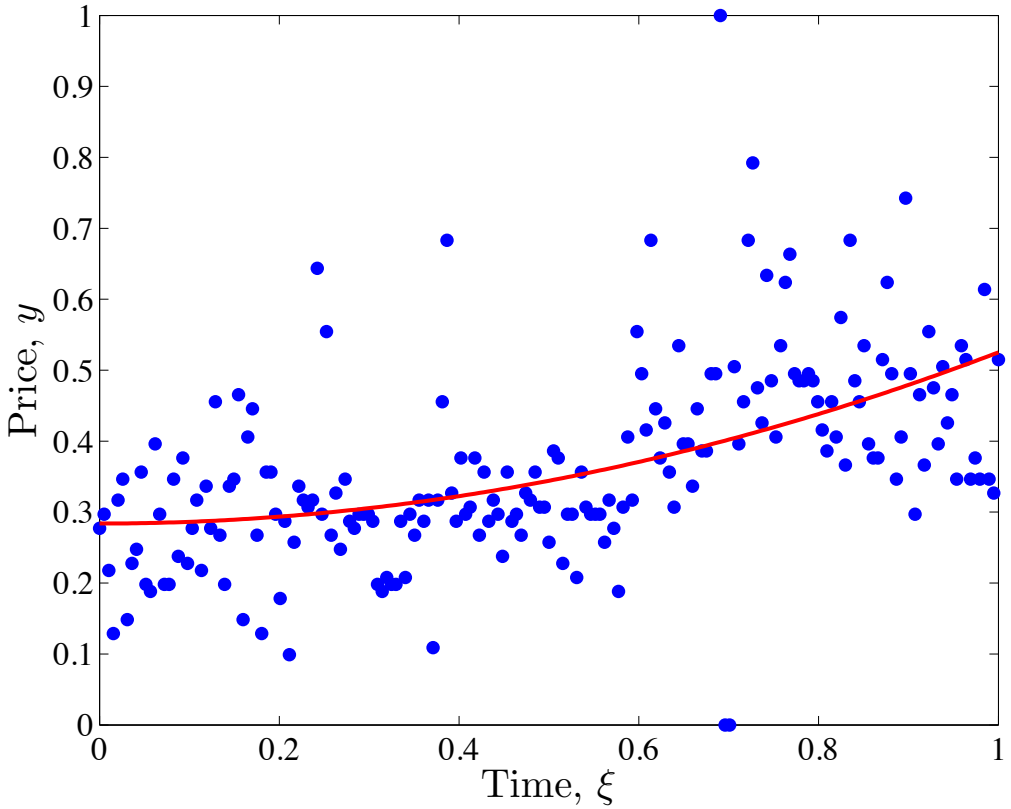


The model $\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w})$ could be a linear model, neural network, deep NN, SVN, ...

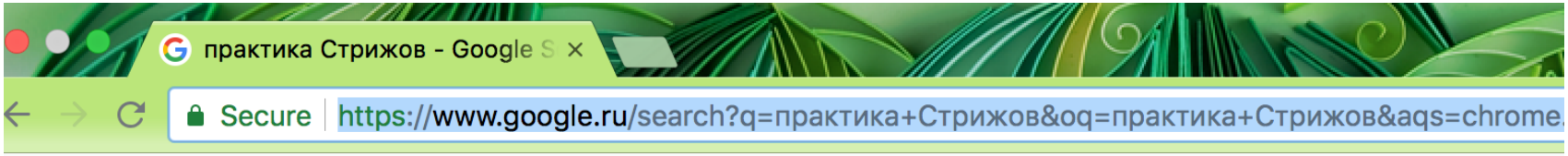
Пример выборки: зависимость объема продажи товара от цены



Одномерная регрессия: зависимость цены товара от времени



Функция регрессии: $y = w_1 + w_2\xi^2 + \varepsilon(\xi)$. Оптимальные параметры: $\mathbf{w}_0 = [0.2839, 0.2412]^\mathsf{T}$. Регрессионная модель: $f = \mathbf{x}^\mathsf{T}\mathbf{w}$, где $x_1 = \xi^0, x_2 = \xi^2$.



Google

практика Стрижов

All Videos Images News Maps More Settings Tools

About 10,400 results (0.39 seconds)

Численные методы обучения по прецедентам (практика, В.В ...

www.machinelearning.ru/.../index.php?...%28практика%2C...Стри... ▾ Translate this page

Jul 7, 2018 - Katrutsa A.M., **Strijov** V.V. Stresstest procedure for feature selection algorithms // Chemometrics and Intelligent Laboratory Systems, 2015, 142 ...

Задачи · Домашние задания · Подготовка к экзамену · Справочные материалы

Численные методы обучения по прецедентам (практика, В.В ...

www.machinelearning.ru/.../index.php?...%28практика%2C...Стри... ▾ Translate this page

Aug 24, 2017 - Карасиков М.Е., **Стрижов** В.В. Классификация временных рядов в пространстве параметров порождающих моделей // Информатика и ...

(практика, В.В. Стрижов)/Группа 474, весна 2018 - MachineLearning.ru

www.machinelearning.ru/.../index.php?...%28практика%2C...Стри... ▾ Translate this page

Jun 20, 2018 - Короткая ссылка bit.ly/2JkLqlo. Два курса весеннего семестра. Выбор моделей в задачах регрессии и классификации, лекции по теории ...

Численные методы обучения по прецедентам (практика, В.В ...

www.machinelearning.ru/.../index.php?...%28практика%2C...Стри... ▾ Translate this page

May 23, 2018 - Адуенко А.А. Выбор мультимоделей в задачах классификации (научный руководитель В.В. **Стрижов**). Московский физико-технический ...