УНИВЕРСИТЕТ ИТМО

Я ПРОФИ

СТУДЕНЧЕСКАЯ ОЛИМПИАДА Я — ПРОФЕССИОНАЛ

Машинное обучение: от базовых понятий до решения нестандартных задач

Лекция 4

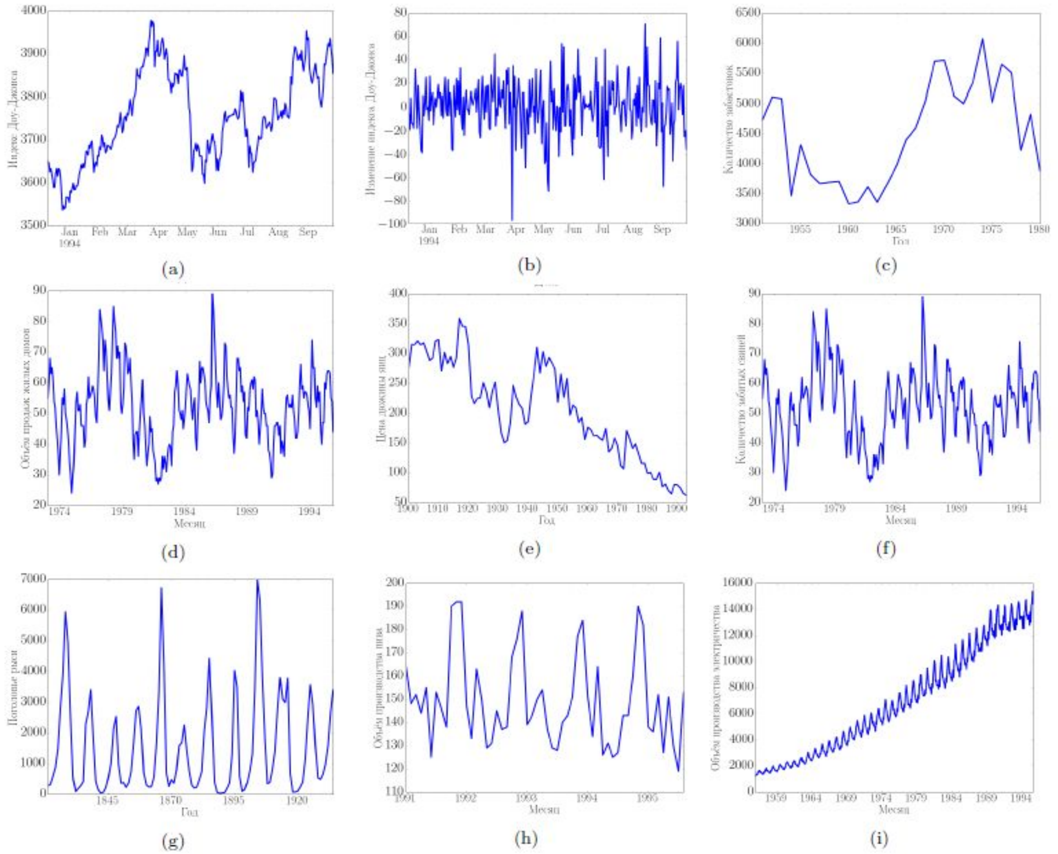
Временные ряды

План

- Временные ряды и их свойства
- Модель ARIMA
- Метрики точности прогноза
- Одномерное и многомерное прогнозирование
- Прогнозирование как задача машинного обучения
- Кросс-валидация на временных ряда
- Нейронные сети в зачах предсказания временных рядов

Примеры временных рядов

Временной ряд – это последовательность значений, описывающих протекающий во времени процесс, измеренных в последовательные моменты времени, обычно через равные промежутки.



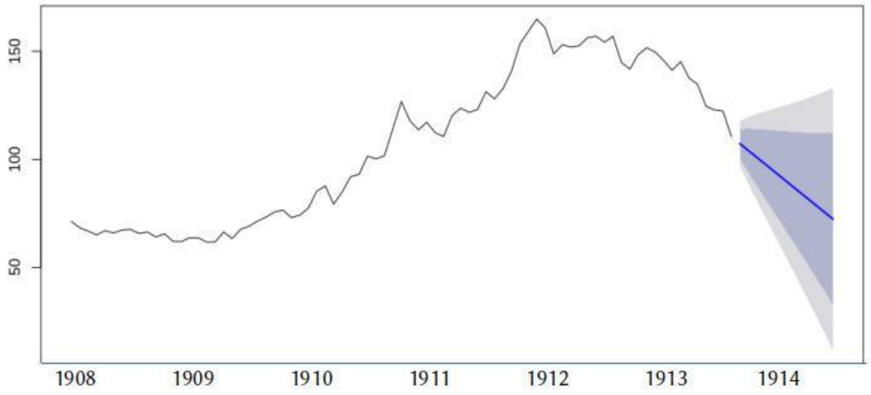
Анализ временных рядов

Свойства временных рядов:

- Тренд
- Сезонность
- Цикл(ы)
- Ошибки (шум)
- Стационарность

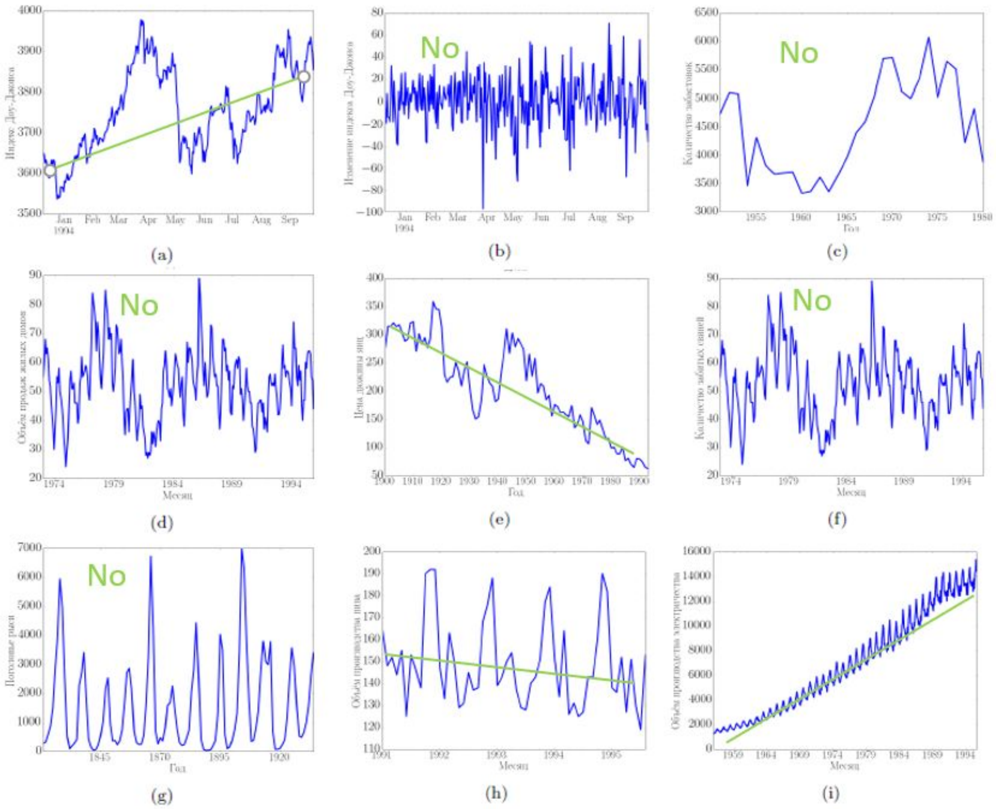
Задачи:

- Поиск аномалий
- Поиск локальных трендов
- (локальные) максимумы и минимумы
- Корреляция с внешними характеристиками (новости, внешние переменные, стоимость валюты и т. д.)
- **ПРОГНОЗИРОВАНИЕ**



Свойства временных рядов: тренд

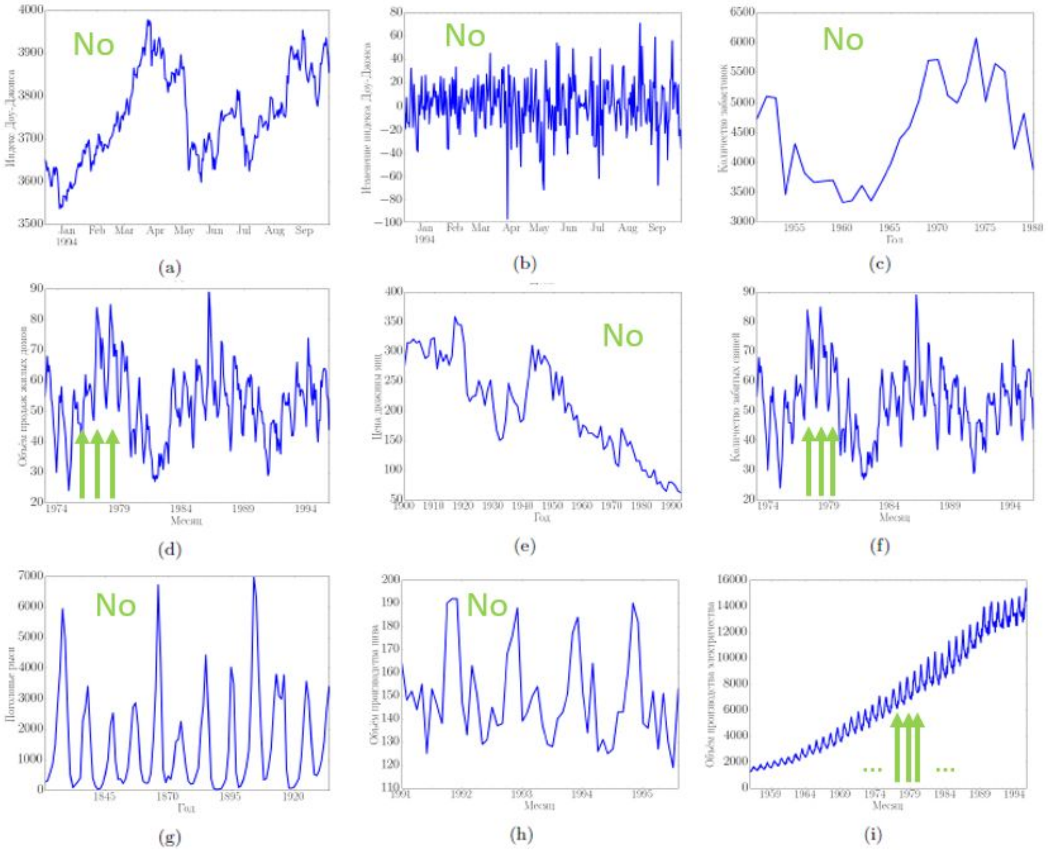
Тренд - плавная длительная смена уровня ряда.



Свойства временных рядов: сезонность

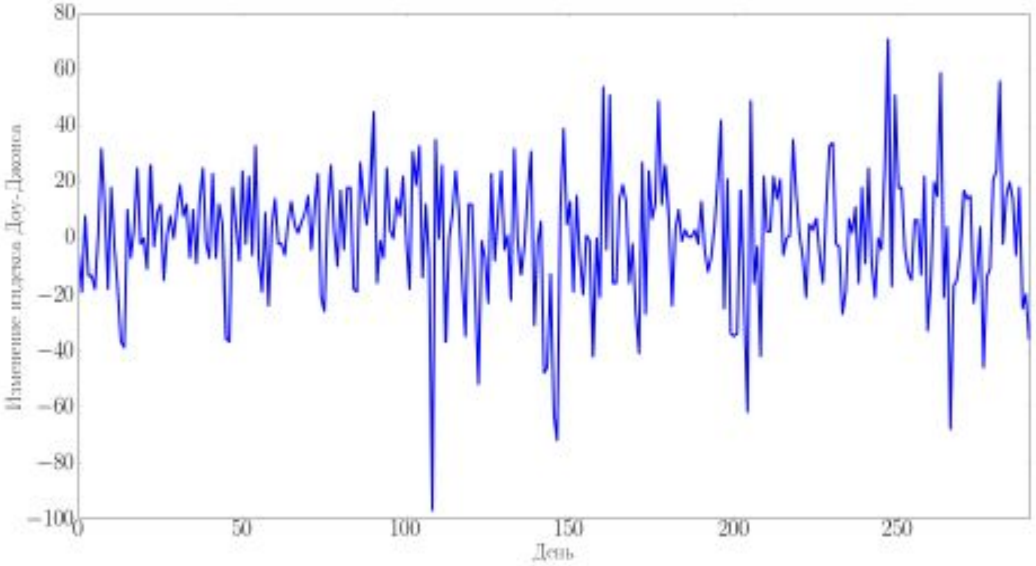
Сезонность – это циклические изменения уровня ряда с постоянным периодом.

Циклы – это изменение уровня ряда с переменным периодом.



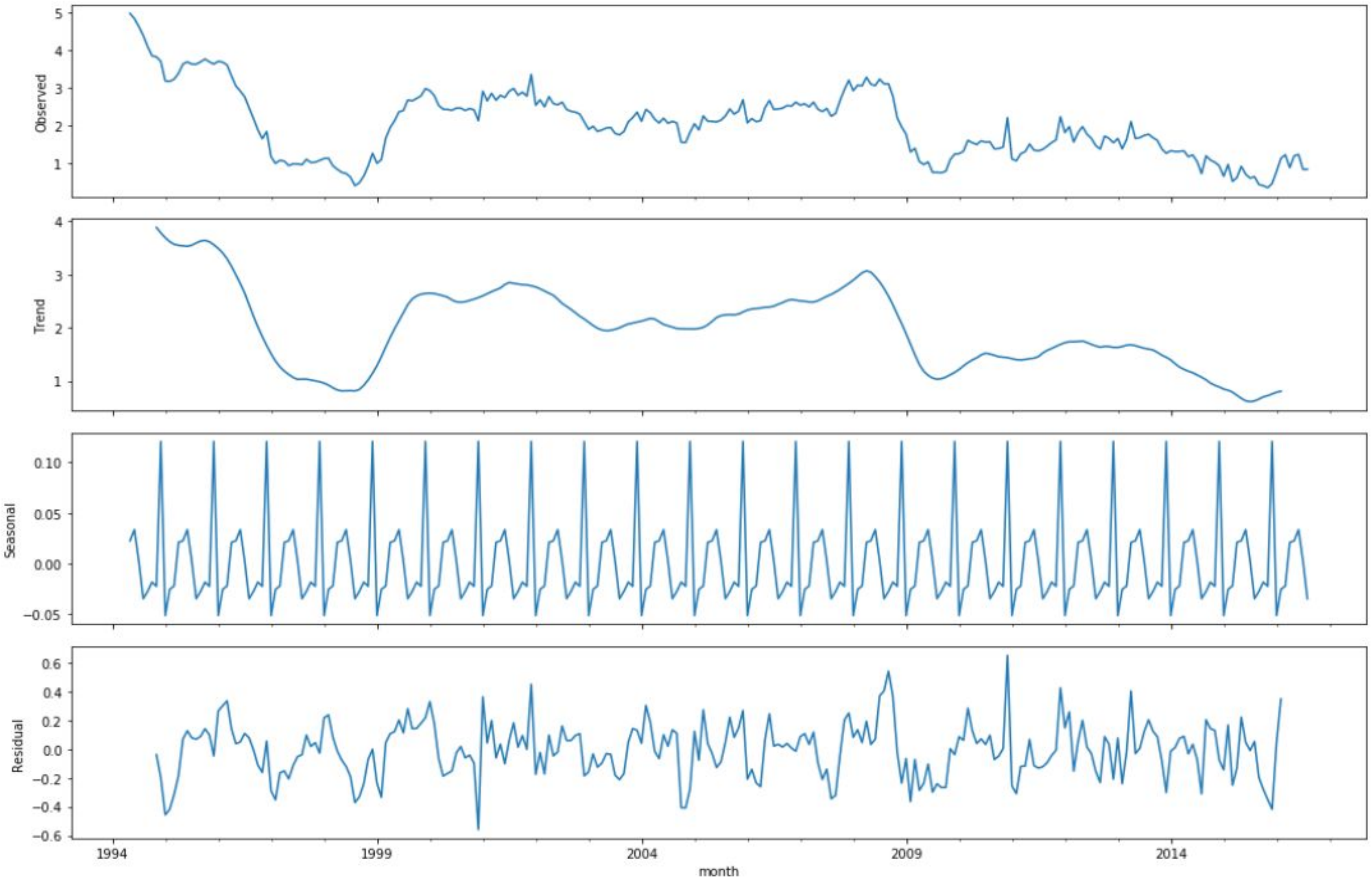
Свойства временных рядов: шум

Шум – это непредсказуемый случайный компонент временных рядов.



- несистематическое поведение: нет тренда, нет сезонности, нет циклов…
- случайная составляющая;
- ~ небольшие отклонения;

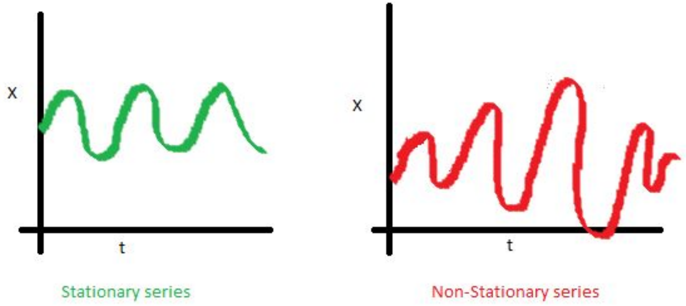
Компоненты временных рядов



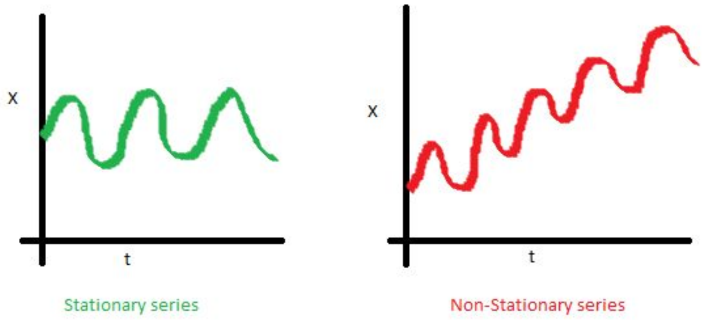
Свойства временных рядов: стационарность

Стационарность – это свойство процесса не менять своих статистических характеристик с течением времени, а именно постоянство матожидания, постоянство дисперсии (гомоскедастичность) и независимость ковариационной функции от времени (должна зависеть только от расстояния между наблюдениями).

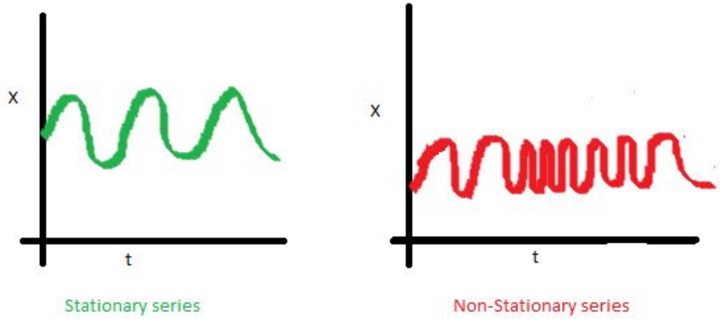
Изменение дисперсии



Изменение матожидания



Непостоянство ковариаций

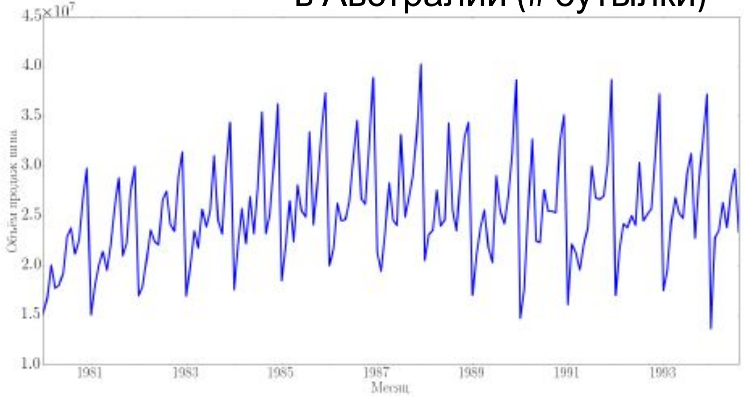


Источник

Автокорреляция (I)

Автокорреляция – это статистическая взаимосвязь между последовательностями величин одного ряда, взятыми со сдвигом.

Ежемесячный объем продаж вина в Австралии (# бутылки)



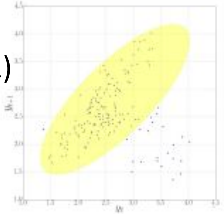
Автокорреляционная функция для лага τ:

$$r_\tau = \frac{\sum_{t=1}^{T-\tau}(y_t - \bar{y})(y_{t+\tau} - \mathbb{E}y))}{\sum_{t=1}^{T-\tau}((y_t - \bar{y}))^2}.$$

Корреляционная функция Пирсона между значением временного ряда в момент времени (t) и $(t+\tau)$.

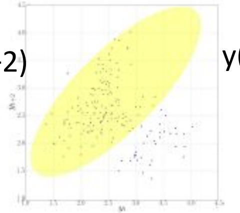
Зависимость значений от предыдущих шагов

y(t+1)



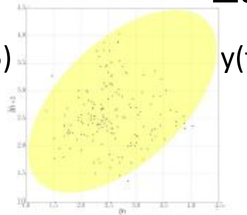
y(t)

y(t+2)



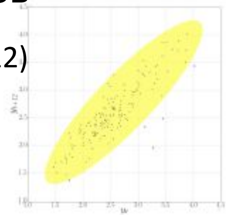
y(t)

y(t+5)



y(t)

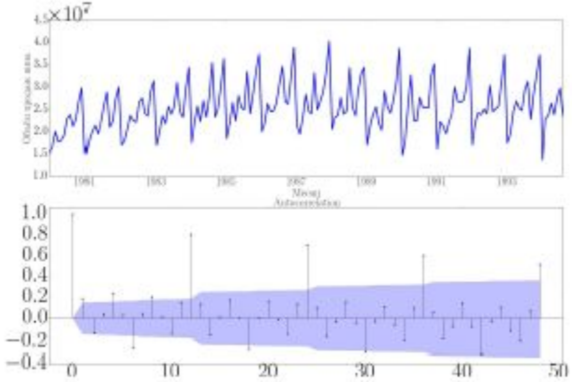
y(t+12)

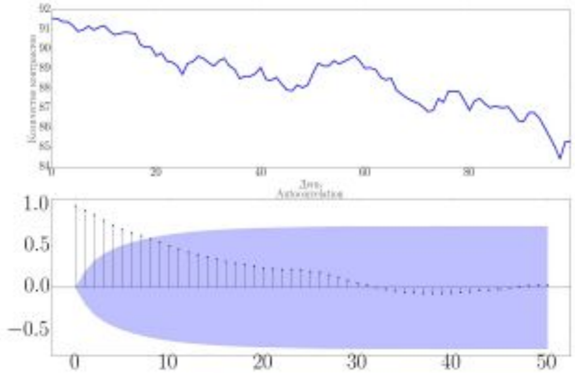


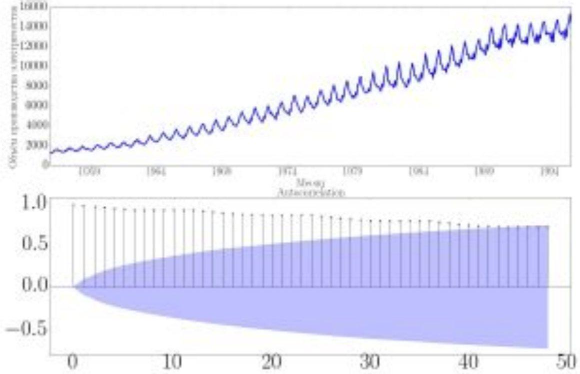
y(t)

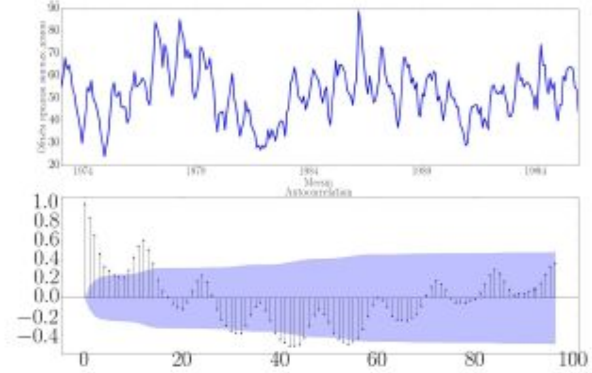
Автокорреляция (II)

Примеры:









Операции с временными рядами

- Дифференцирование (derivative):

$$y' = y_t - y_{t-1}.$$

- Сезонное дифференцирован
Seasonal derivative:

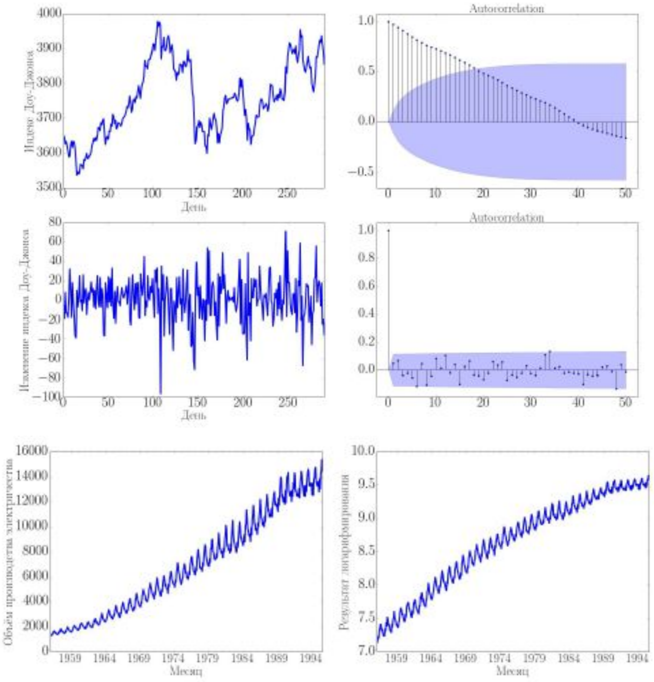
$$y'_t = y_t - y_{t-s}.$$

- Нормализация дисперсии (преобразование Бокса-Кок

$$y'_t = \begin{cases} \ln y_t, & \lambda = 0, \\ \left(y_t^\lambda - 1\right)/\lambda, & \lambda \neq 0. \end{cases}$$

- Тест на стационарность (Критерий Дики-Фуллера):

H_0 – non-stationarity
H_1 – stationarity



Модель ARIMA (I)

autoregressive integrated moving average

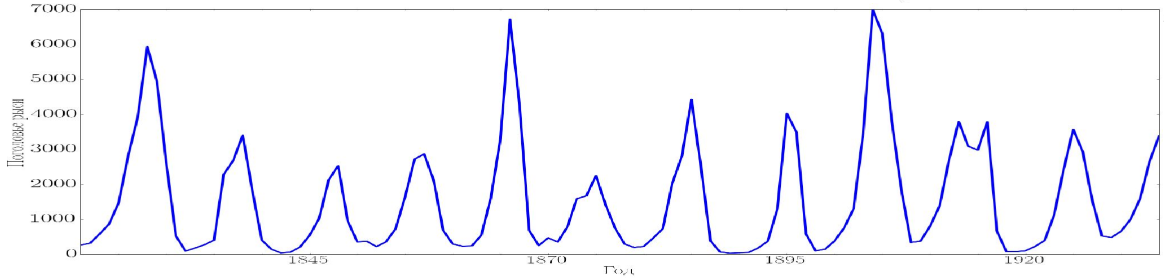
- Показывает хорошие результаты в прогнозировании авторегрессионных временных рядов с сильной сезонностью;
- Необходима индивидуальная тонкая настройка для каждого нового примера.

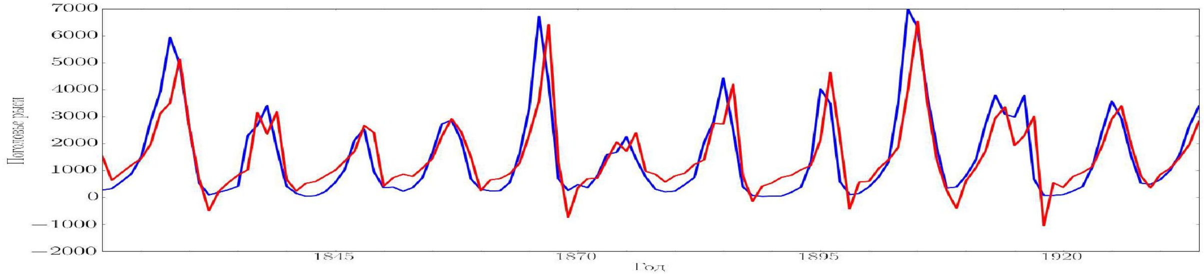
Компоненты:

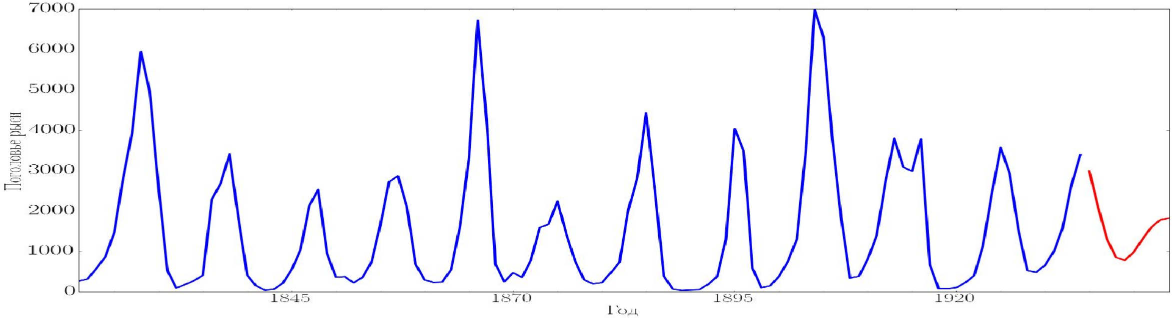
- AR(p), авторегрессионная компонента: $y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t.$
- MA(q), компонента скользящего среднего: $y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$
- ARMA(p,q): $y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}.$

Модель ARIMA (II)

ARMA(2,2)







Модель ARIMA (III)

УНИВЕРСИТЕТ ИТМО

Wold's theorem:
Каждый стационарный временной ряд может быть аппроксимирован моделью ARMA (p, q) с заданной точностью.

Временной ряд должен быть **стационарен:**

- Преобразование Бокса-Кокса (log)
- Дифференцирование (одношаговое или сезонное)

$\Rightarrow$ ARIMA(p,d,q) – модель ARMA для временных рядов, где d-порядок дифференцирования (взятия последовательной разности)

Сезонность $+\phi_S y_{t-S} + \phi_{2S} y_{t-2S} + \cdots + \phi_{PS} y_{t-PS}$ + P components with period S

$+\theta_S \varepsilon_{t-S} + \theta_{2S} \varepsilon_{t-2S} + \cdots + \theta_{PS} \varepsilon_{t-QS}.$ + Q components with period S

$\Downarrow$

SARMA(p,q)x(P,Q)

Модель ARIMA (IV)

Необходимо найти значения (P,Q,p,q).

Минимизация информационного критерия Акаике (Akaike info criterion):   AIC = 2 lnL + 2k

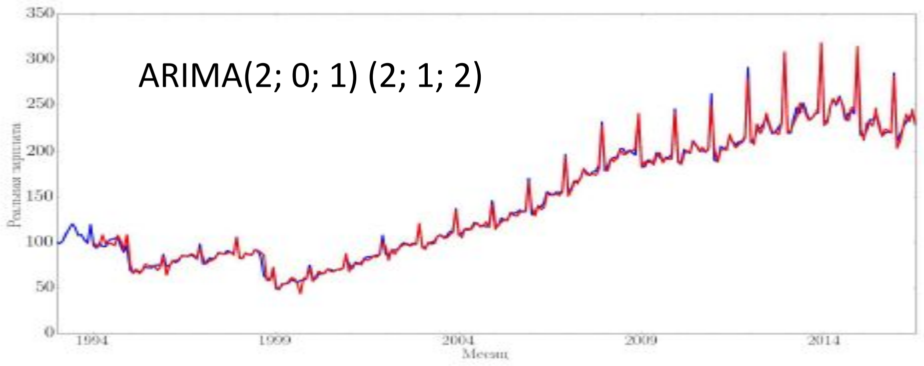
L - Функция правдоподобия

k = P + Q + p + q + 1 – число параметров модели

Лучшая модель - модель ARIMA(p,q)x(P,Q) с минимальным значением AIC.

SARMA(p,q)x(P,Q)
+ d – порядок дифференцирования
+ D – порядок сезонного дифференцирования

= модель **SARIMA(p,d,q)x(P,D,Q)**



Метрики точности прогноза

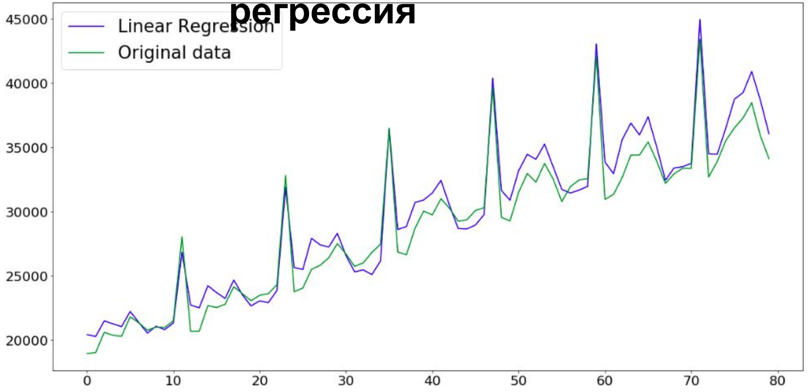
Пример. Сравним две модели:

- линейная регрессия
- скользящее среднее значение.

График ниже иллюстрирует результат прогнозирования моделей на тестовом наборе данных.

Скользящее среднее

Линейная регрессия



Метрики точности прогноза

Метрики оценки точности прогноза:

- R^2
- MSE (RMSE) – mean squared error – среднеквадратичная ошибка
- MAE – mean absolute error – средняя абсолютная ошибка
- MAPE – mean absolute percentage error – средняя абсолютная ошибка в %
- SMAPE – symmetric mean absolute percentage error – симметричная средняя абсолютная ошибка в %

Метрики точности прогноза: R^2

"R квадрат" или коэффициент детерминации – это доля дисперсии зависимой переменной, которую можно спрогнозировать на основе независимых переменных.

- Обычно используется для моделей линейной регрессии
- $0 \ \le \ R^2 \le 1$
- Чем выше значение, тем лучше.

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$$

$$TSS = \sum_{t=1}^{n} (y_t - \bar{y})^2$$

$$\bar{y} = \frac{1}{n} \sum_{t=1}^{n} y_t$$

```python
from sklearn.metrics import r2_score

print("Linear Regression R^2:",  round(r2_score(y, y_pred_lr), 3))
print("SMA R^2:",  round(r2_score(y , y_sma), 3))
```

```
Linear Regression R^2: 0.942
SMA R^2: 0.822
```

Метрики точности прогноза: MSE

Среднеквадратичная ошибка (MSE) измеряет среднее значение квадратов ошибок, то есть среднеквадратичную разность между прогнозируемыми и фактическими значениями.

- Всегда неотрицательна.
- Значения ближе к нулю лучше.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2$$

```python
from sklearn.metrics import mean_squared_error

print("Linear Regression MSE:", round(mean_squared_error(y, y_pred_lr), 3))
print("SMA MSE:",  round(mean_squared_error(y , y_sma), 3))
```

```
Linear Regression MSE: 1882343.713
SMA MSE: 5774211.042
```

Метрики точности прогноза: RMSE

Среднеквадратичная ошибка - это корень из среднего квадрата разности между прогнозируемыми и фактическими значениями.

- Всегда неотрицательна.
- Значения ближе к нулю лучше.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

```python
from sklearn.metrics import mean_squared_error

print("Linear Regression RMSE:", round(np.sqrt(mean_squared_error(y, y_pred_lr)), 3))
print("SMA RMSE:",  round(np.sqrt(mean_squared_error(y , y_sma)), 3))
```

```
Linear Regression RMSE: 1371.985
SMA RMSE: 2402.959
```

Метрики точности прогноза: MAE

Средняя абсолютная ошибка - это среднее расстояние по вертикали между каждой прогнозируемой точкой и фактической линией.

$$MAE = \frac{1}{n}\sum_{t=1}^{n} | y_t - \hat{y}_t |$$

```python
from sklearn.metrics import mean_absolute_error

print("Linear Regression MAE:",  round(mean_absolute_error(y, y_pred_lr), 3))
print("SMA MAE:", round(mean_absolute_error(y , y_sma), 3))
```

```
Linear Regression MAE: 1148.816
SMA MAE: 1341.285
```

Метрики точности прогноза: MAPE

Средняя абсолютная процентная ошибка (MAPE) показывает среднюю долю ошибки относительно фактического значения. MAPE обычно выражает точность в процентах.

Нельзя использовать, если есть нулевые значения, потому что будет деление на ноль.
Для слишком низких прогнозов процентная ошибка не может превышать 100%, но для слишком высоких прогнозов нет верхнего предела процентной ошибки.

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right|$$

```
def mean_absolute_percentage_error(y_true, y_pred):
    return round(np.mean(np.abs((y_true - y_pred) / y_true)) * 100, 3)

print("Linear Regression MAPE:",  mean_absolute_percentage_error(y, y_pred_lr))
print("SMA MAPE:", mean_absolute_percentage_error(y , y_sma))

```

```
Linear Regression MAPE: 4.0
SMA MAPE: 22.493
```

Метрики точности прогноза: SMAPE

УНИВЕРСИТЕТ ИТМО

Симметричная средняя абсолютная ошибка в процентах - это показатель точности, основанный на процентах.

Абсолютная разница между фактическим значением и прогнозируемым значением делится на половину суммы абсолютных значений фактического значения и прогнозируемого значения. Значение этого вычисления суммируется для каждой подобранной точки t и снова делится на количество подобранных точек n.

$$SMAPE = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|\hat{y}_t - y_t|}{\frac{1}{2} * (|y_t| + |\hat{y}_t|)}$$

```
def smape(y_true, y_pred):
    return round(np.mean(np.abs((y_pred - y_true))/((np.abs(y_true)) + np.abs((y_pred)) / 2)), 3 )

print("Linear Regression SMAPE:",  smape(y, y_pred_lr))
print("SMA SMAPE:", smape(y , y_sma))
```

```
Linear Regression SMAPE: 0.026
SMA SMAPE: 0.147
```

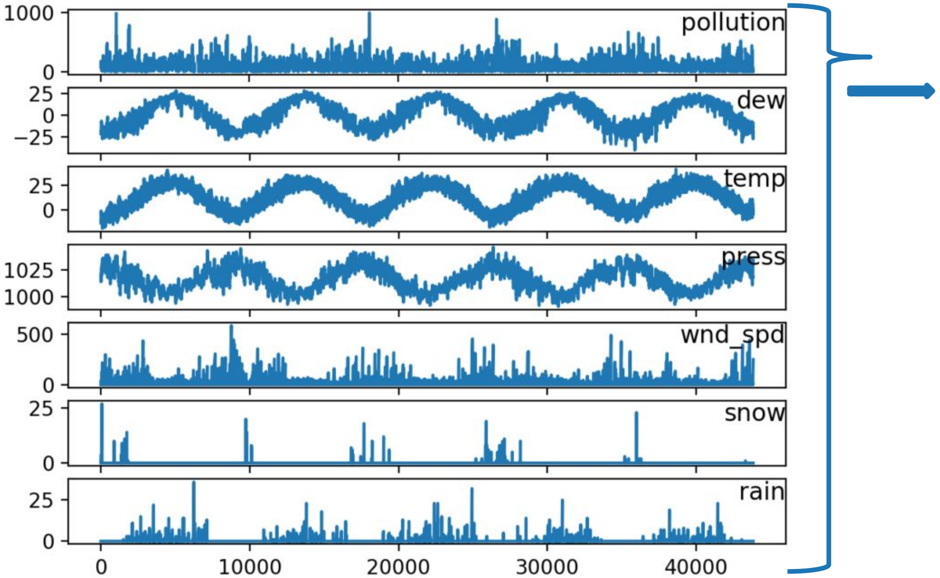
Одномерное и многомерное прогнозирование

Одномерный (Univariate):

- Один целевой временной ряд
- Прогнозирование только на его основе

Многомерное (Multivariate):

- Один целевой временной ряд
- Несколько характеристик за один и тот же период времени, которые могут повлиять на результат (курс валюты, температура, уровень безработицы и др.)
- Прогноз на основе полных данных



Прогнозирование как задача машинного обучения

УНИВЕРСИТЕТ ИТМО

Прогнозирование на один шаг вперед. Задача обучения с учителем.

Необходимые данные:

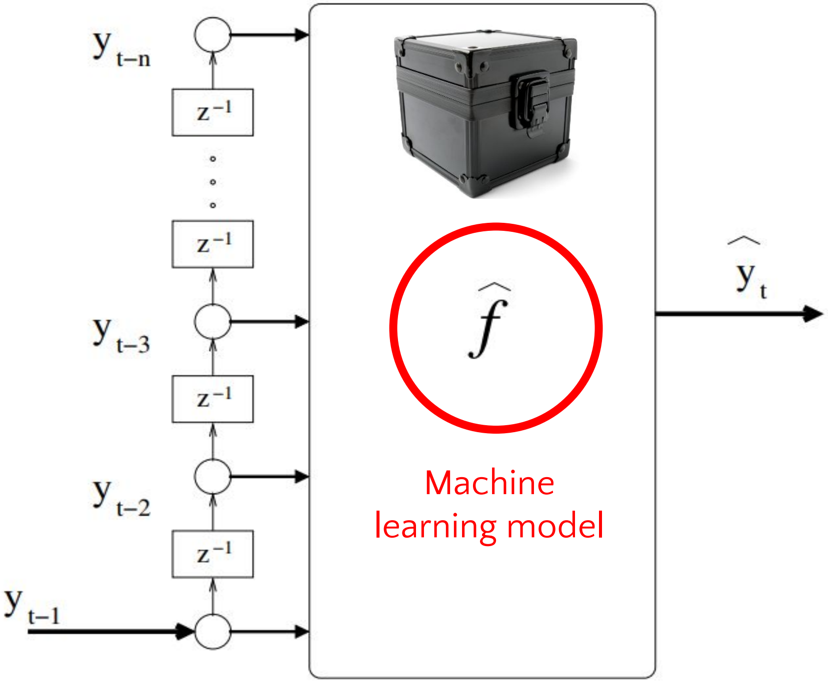
- обучающий набор (входы)
- метки (выходам)
- и тестовый набор

Временной ряд: **S: [$y_0, y_1, \ldots\ldots y_{t-2}, y_{t-1}$]**

Предсказываем **<y_t>**

$$X = \begin{bmatrix} y_{N-1} & y_{N-2} & \cdots & y_{N-n-1} \\ y_{N-2} & y_{N-3} & \cdots & y_{N-n-2} \\ \vdots & \vdots & \vdots & \vdots \\ y_n & y_{n-1} & \cdots & y_1 \end{bmatrix} \quad Y = \begin{bmatrix} y_N \\ y_{N-1} \\ \vdots \\ y_{n+1} \end{bmatrix}$$

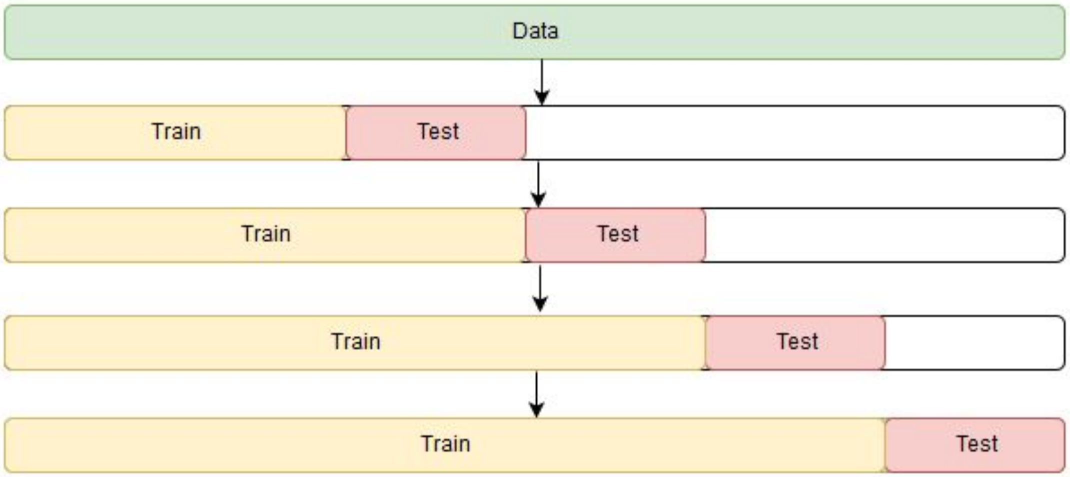
входы выходы



Кросс-валидация на временных рядах

Временной ряд имеет временную структуру, поэтому случайно перемешивать в фолдах значения всего ряда без сохранения этой структуры нельзя, так как в процессе потеряются все взаимосвязи наблюдений.

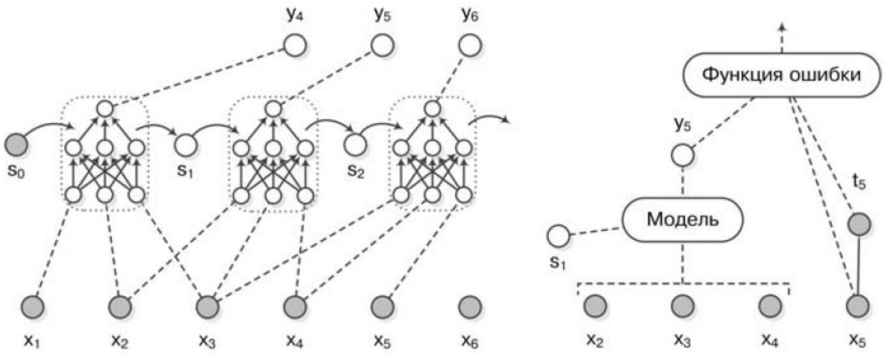
Для этого модель обучается и тестируется на последовательных интервалах данных



Нейронные сети в зачах предсказания временных рядов

- Модель должна «помнить» элементы последовательности с целью использовать их в дальнейшем;
- Необходимо фиксировать зависимости с большим временным окном.

Рекуррентные нейронные сети - это вид нейронных сетей, где связи между элементами образуют направленную последовательность. Благодаря этому появляется возможность обрабатывать серии событий во времени или последовательные пространственные цепочки.



Список дополнительной литературы

1. Афанасьев В.Н., Юзбашев М.М. Анализ временных рядов и прогнозирование — М.: Финансы и статистика, 2001. — 228 с.:
2. Портал https://machinelearningmastery.com/
3. Статья: https://habr.com/ru/company/ods/blog/327242/

Вопросы для самоконтроля

1) Для оценки точности прогноза с нулевыми значениями в фактических данных нельзя использовать:

1. R^2
2. MAPE
3. MSE

2) Что не относится к операциям, которые используются для преобразования временного ряда к стационарному:

1. Дифференцирование
2. Масштабирование
3. Преобразование Бокса-Кокса

Ответы на вопросы можно найти в файле «ответы.docx»

УНИВЕРСИТЕТ ИТМО

Я ПРОФИ
СТУДЕНЧЕСКАЯ
ОЛИМПИАДА
Я — ПРОФЕССИОНАЛ

https://yandex.ru/profi/

Спасибо за внимание!

Материалы подготовлены
преподавателями Факультета цифровых трансформаций
и сотрудниками Национального центра когнитивных разработок Университета ИТМО
для участников Студенческой олимпиады «Я – профессионал» (направление «Машинное обучение»)
dx.itmo.ru, actcognitive.org