Применение методов машинного обучения для калибровки газового датчика

Козьмин Артём Дмитриевич

Новосибирский национальный исследовательский государственный университет, Новосибирск

24 октября 2022 г.

Содержание

Введение

Мотивация

Область применения электрохимических датчиков:

- Мониторинг окружающего воздуха.
- Химическая промышленность.
- Бытовые системы вентиляции.

Эл.-хим. датчик $NH3$.

Мотивация

Область применения электрохимических датчиков:

- Мониторинг окружающего воздуха.
- Химическая промышленность.
- Бытовые системы вентиляции.

Эл.-хим. датчик $NH3$.

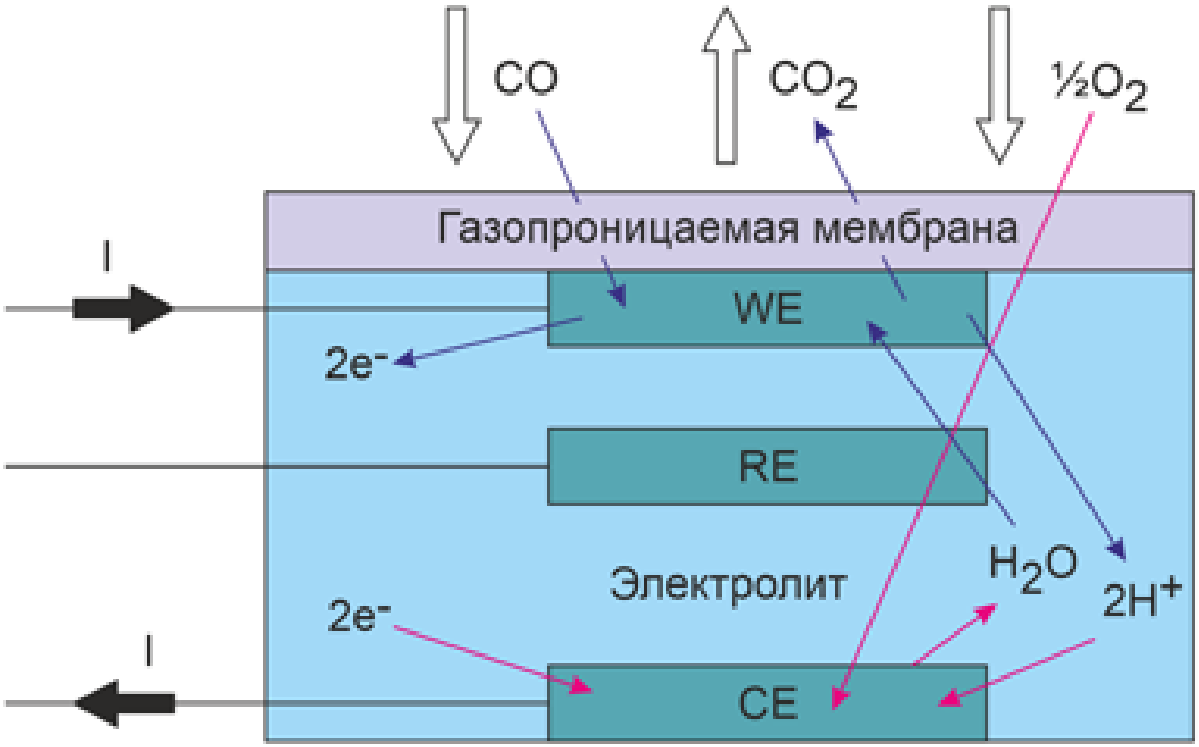
Преимущества:

- Низкая стоимость
- Компактность
- Доступность

Недостатки:

- Неявная зависимость
- Перекрёстная чувствительность
- Подвержены влиянию окружающей среды

Принцип работы



Электрохимический датчик CO.

- Газ попадает внутрь датчика через мембрану.
- На электроде WE реакция окисления.
- На CE реакция восстановления.
- Ток пропорционален концентрации газа.

Математическая постановка задачи

Применение методов машинного обучения для восстановления концентрации угарного газа CO по выходным данным электрохимического датчика.

Задача регрессии:
$Y = f(X) + \varepsilon$, где $f-$ функция регрессии, $\varepsilon-$ случайный шум.

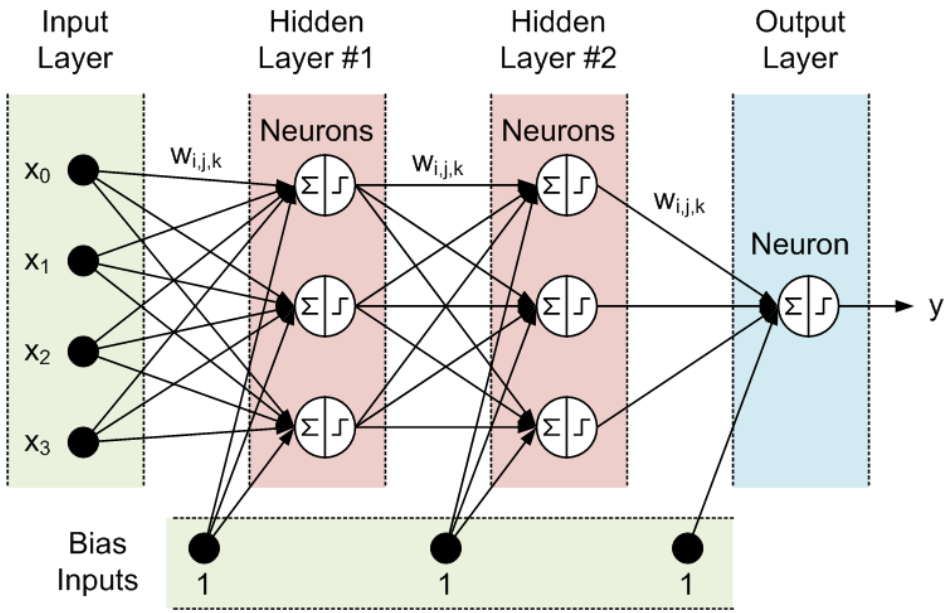
Математическая постановка задачи

Применение методов машинного обучения для восстановления концентрации угарного газа CO по выходным данным электрохимического датчика.

Задача регрессии:
$Y = f(X) + \varepsilon$, где $f-$ функция регрессии, $\varepsilon-$ случайный шум.

Методы машинного обучения:

- Линейная регрессия:
  $f_\omega = \omega_0 + \sum_{i=1}^{p} \omega_i x_i \equiv x^T \omega$
- Полиномиальная регрессия 2 степени:
  $f_\omega = \omega_0 + \sum_{i=1}^{p} \omega_i x_i + \sum_{i,j=1}^{p,p} \omega_{ij} x_i x_j$
- Нейронные сети с прямой связью.



Нейронная сеть с прямой связью.

Исследуемые данные

Набор исследуемых данных:

- Мультисенсорное устройство, разработано Pirelli Labs
- 5 Датчиков CO, $NMHC$, NO_2, O_3, NO_x
- Температура воздуха T и влажность RH

Исследуемые данные

Набор исследуемых данных:

- Мультисенсорное устройство, разработано Pirelli Labs
- 5 Датчиков CO, $NMHC$, NO_2, O_3, NO_x
- Температура воздуха T и влажность RH
- Центр итальянского города
- С марта 2004 года по апрель 2005 года
- Одно усреднённое измерение в час
- 7344 ненулевых измерений
- Целевые концентрации с эталонного анализатора

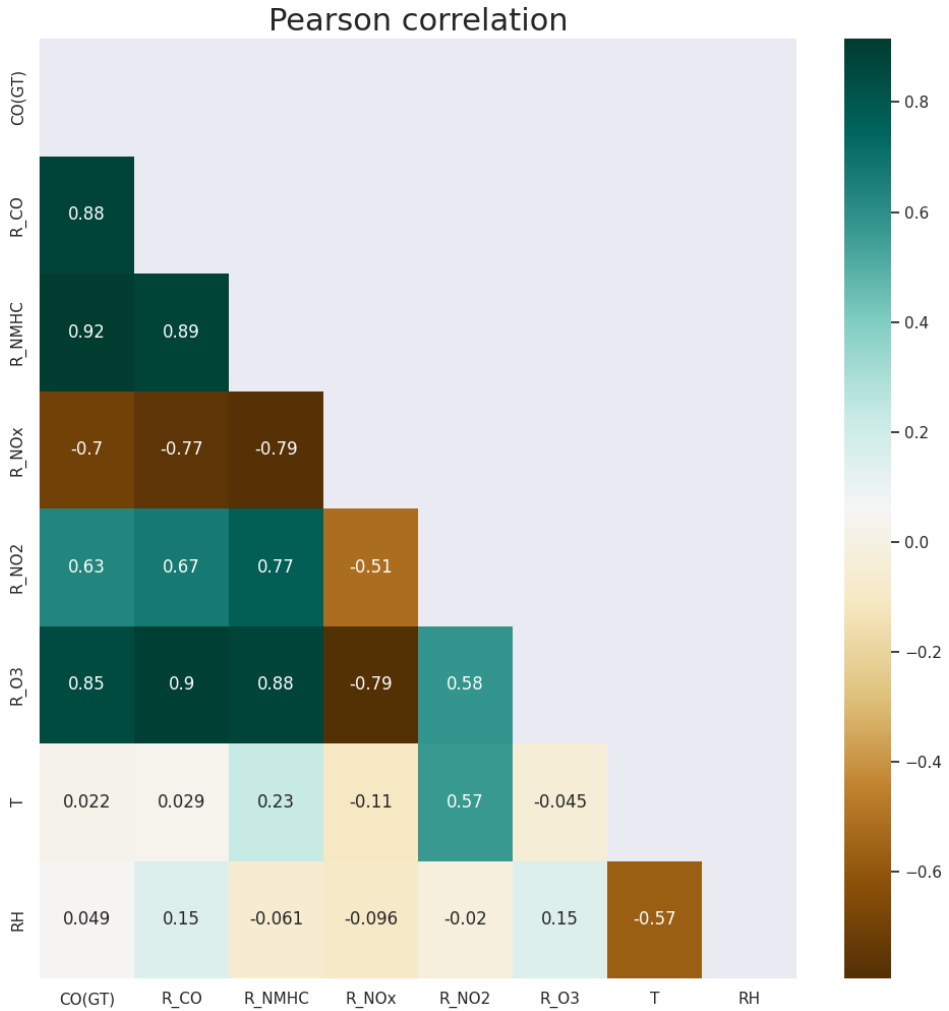
Предобработка данных

Коэффициент корреляции r — Пирсона между двумя величинами X_1, X_2:

$$r_{X_1X_2} = \frac{\mathbf{cov}_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}}$$

Интерпретация:

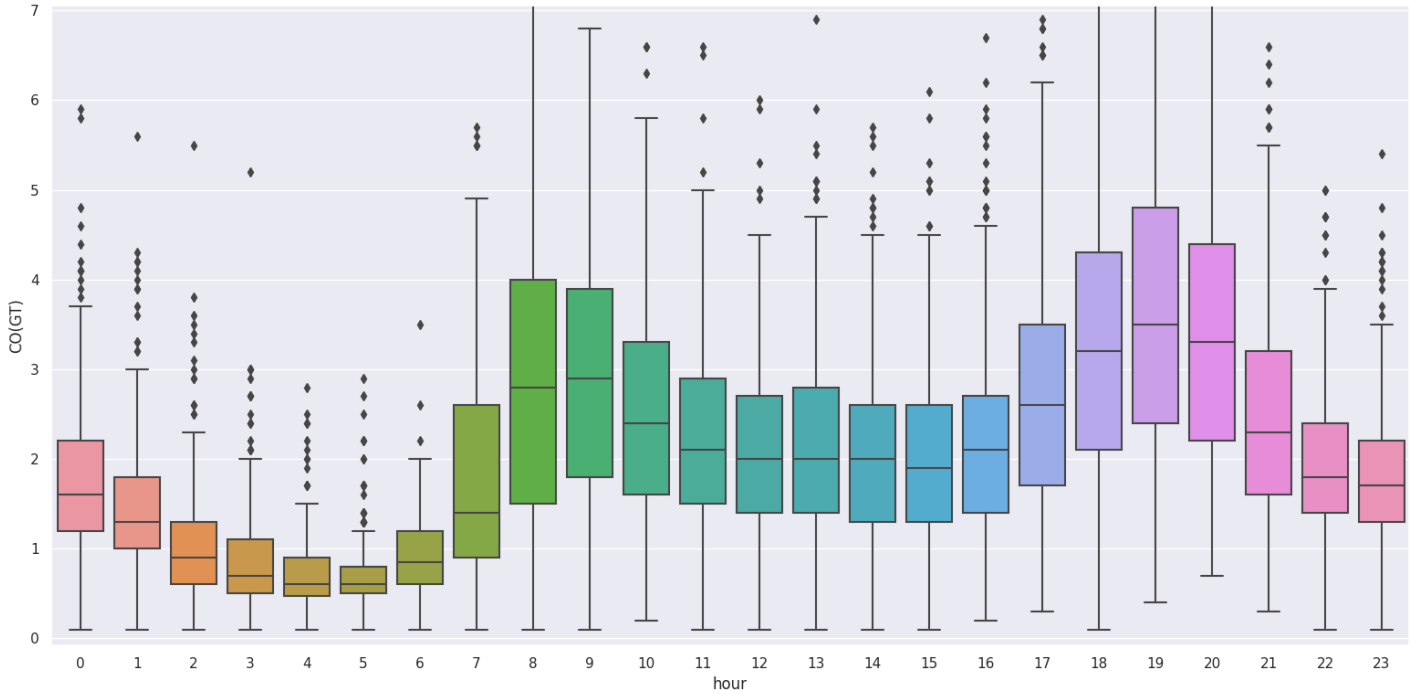
- $r \approx -1$ - отрицательная зависимость
- $r \approx 0$ - отсутсвие линейной зависимости
- $r \approx 1$ - положительная зависимость



Корреляционная матрица Пирсона.

Зависимость концентрации CO от часа измерения

Ящичная диаграмма концентрации газа CO от часа измерения.



- Нижняя и верхняя граница ящика - первый и третий квартили
- Линия в середине ящика - медиана
- Концы усов - края статистически значимой выборки
- Точки - данные выходящие за границы усов (выбросы)

Вейвлет преобразование

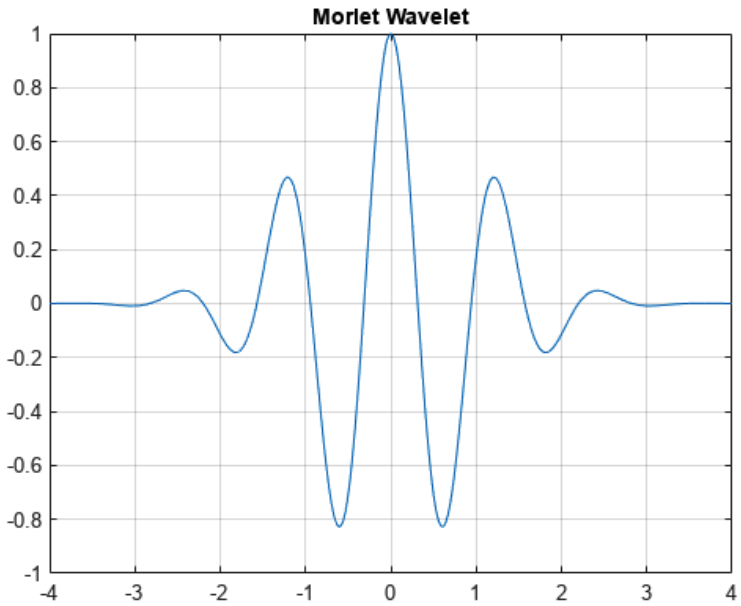
- Вейвлет преобразование - свёртка вейвлет функции $\psi(t)$ с сигналом $f(t)$:

$$[W_\psi f](a,b) = |a|^{-1/2} \int_{-\infty}^{\infty} f(t)\psi^*\left(\frac{t-b}{a}\right)dt,$$

где a - масштабный коэффициент, b - параметр сдвига.

- Переход из временного представления в частотно временное.
- Вейвлет Морле (Morlet):

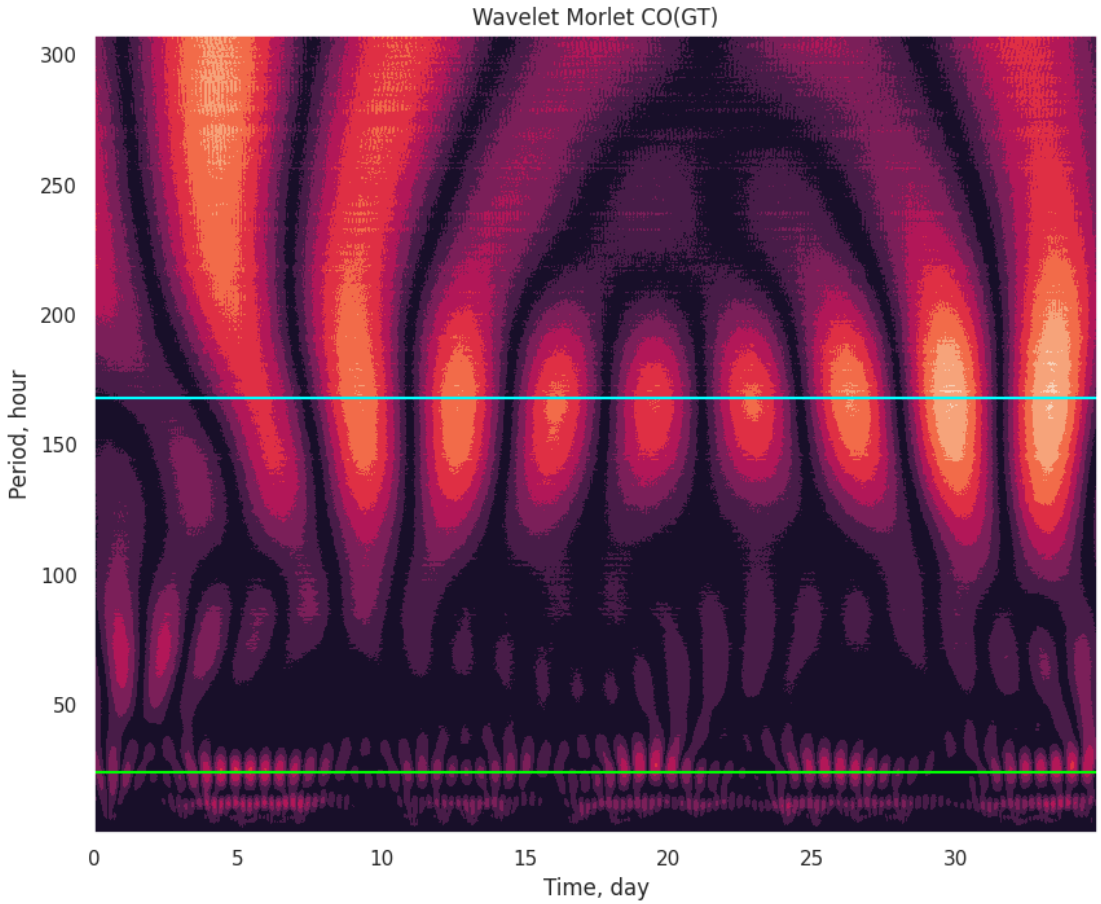
$$\psi(t) = e^{-t^2/2}cos(5t)$$



Вейвлет Морле.

Периодическая структура в данных

Картина вейвлет коэффициентов временного ряда концентрации целевого газа CO(GT).



- Зелёная линия - период 24 часа
- Голубая линия - период 7 дней

Линейная регрессия

Метрики

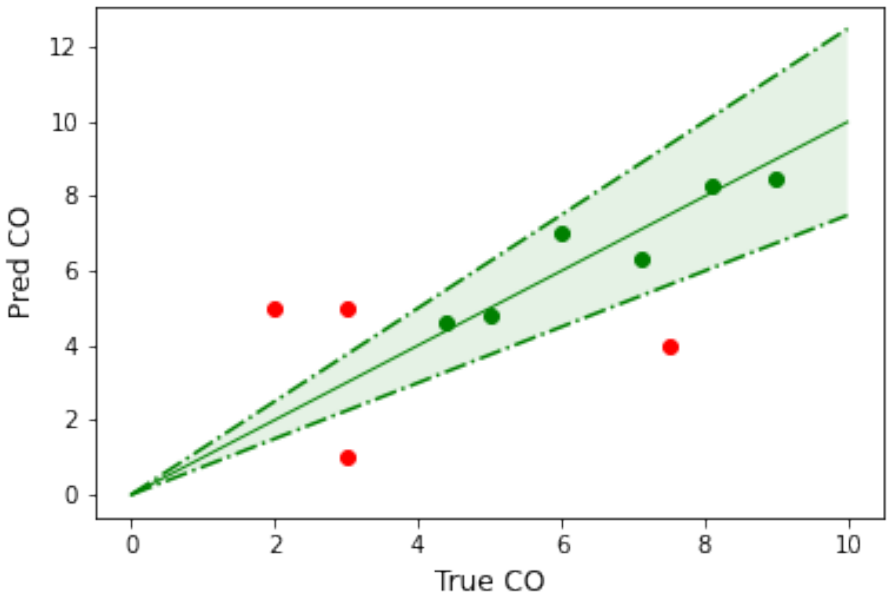
Используемые метрики:

- MSE - среднеквадратичная ошибка:
  $MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$
- MAPE - средняя абсолютная процентная ошибка:
  $MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\frac{|Y_i - \hat{Y}_i|}{|Y_i|}$
- GRE - процент прогнозов лежащих за 25% порогом от истинного значения концентрации

Метрики

Используемые метрики:

- MSE - среднеквадратичная ошибка:
  $MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$
- MAPE - средняя абсолютная процентная ошибка:
  $MAPE = \frac{100\%}{N}\sum_{i=1}^{N}\frac{|Y_i - \hat{Y}_i|}{|Y_i|}$
- GRE - процент прогнозов лежащих за 25% порогом от истинного значения концентрации

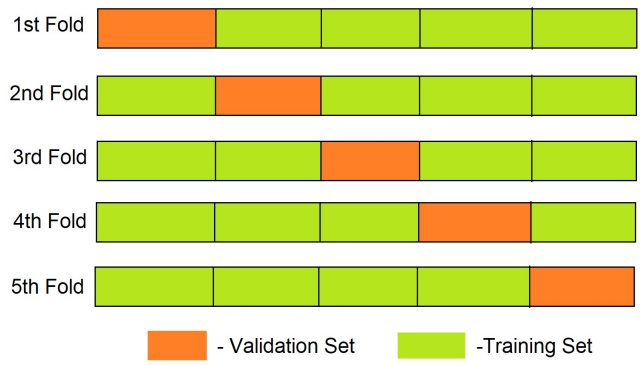


Метрика GRE = 40%.

Результаты

Разделение данных:

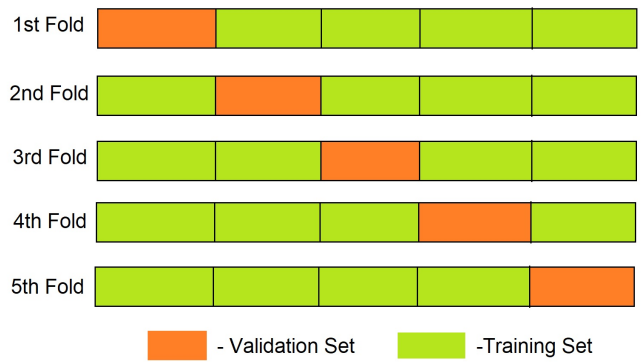
- 2000 часов - тренировка
- 5344 часов - тест
- K-Fold кросс-валидация для моделей с регуляризацией



Результаты

Разделение данных:

- 2000 часов - тренировка
- 5344 часов - тест
- K-Fold кросс-валидация для моделей с регуляризацией



Линейная регрессия без регуляризации.

Модель	$MAPE, \%$	$MSE, (mg/m^3)^2$	$GRE, \%$
R_{CO}	36.7	0.57	41.2
R_{CO}, R_{NM}, T	31.2	0.30	28.3
R_{CO}, R_{NM}, T, CO_h	**31.1**	**0.30**	**26.4**

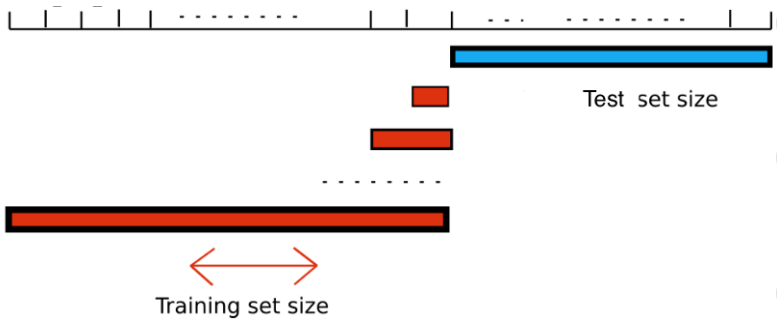
Полиномиальная регрессия с L_1 регуляризацией.

Модель	$MAPE, \%$	$MSE, (mg/m^3)^2$	$GRE, \%$
$Pol_2(R_{CO}, R_{NM}, T)$	25.7	0.26	23.1
$Pol_2(...) + CO_h$	**25.5**	**0.24**	**20.5**

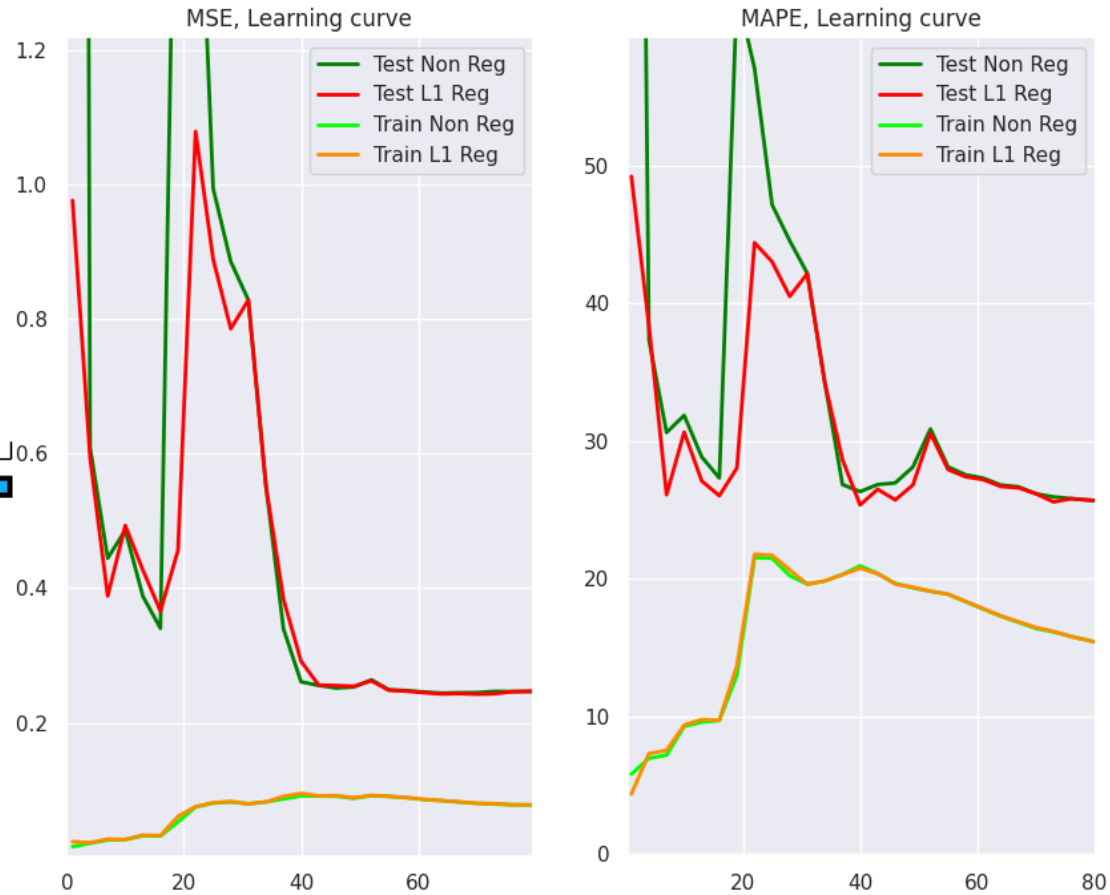
Кривые обучения

Разделение данных:

- Тест - с июня 2004 года по апрель 2005 года, 5344 измерений.
- Тренировка - от 1 и до 80 дней.



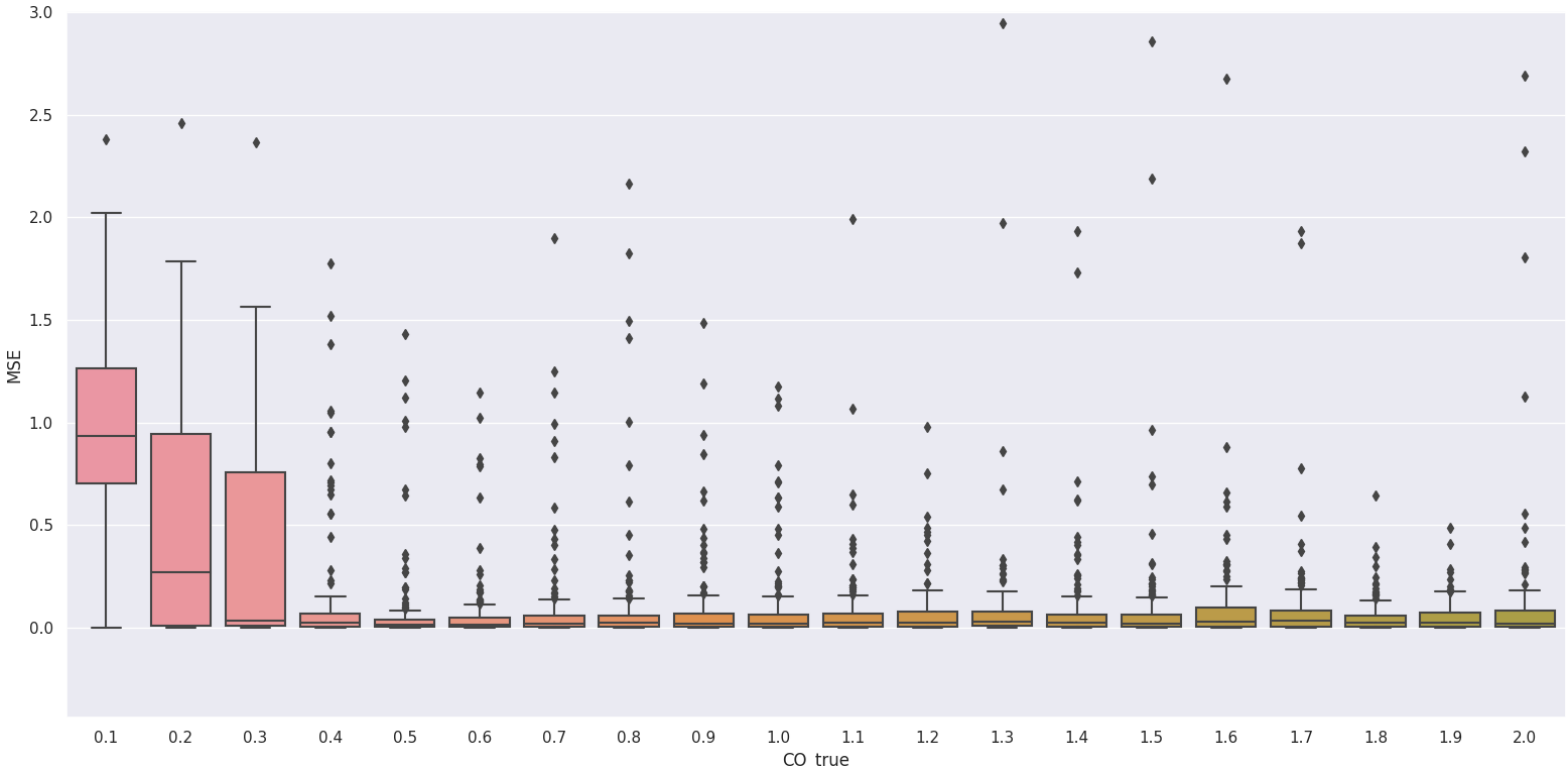
Необходимое время калибровки от 10 до 40 дней.



Кривые обучения $Pol_2(R_{CO}, R_{NM}, T) + CO_h$

Анализ ошибок

Анализ некорректных прогнозов CO в модели полиномиальной регрессии с признаками $Pol_2(R_{CO}, R_{NM}, T) + CO_h$ в случае L_1 регуляризации.



Ящичная диаграмма MSE от целевой концентрации CO
Основной вклад при концентрациях целевого газа ниже $0.4\ mg/m^3$

Нейронные сети

Описание нейронной сети

Полносвязная нейронная сеть с прямой связью:

- 4 входных нейрона: R_{CO}, R_{NM}, T, CO_h
- 1 скрытый слой с 10 нейронами
- Функция активации $tanh$
- 1 выходной нейрон, функция активации $linear$
- Алгоритм оптимизации $ADAM$
- Функция потерь MAE

Описание нейронной сети

Полносвязная нейронная сеть с прямой связью:

- 4 входных нейрона: R_{CO}, R_{NM}, T, CO_h
- 1 скрытый слой с 10 нейронами
- Функция активации $tanh$
- 1 выходной нейрон, функция активации $linear$
- Алгоритм оптимизации $ADAM$
- Функция потерь MAE

Результаты FFNN для различных функций потерь.

Функция потерь	Формула	$MAPE$, %	MSE, $(mg/m^3)^2$	GRE, %
MSE	$\frac{1}{N}\sum(Y_i - \hat{Y}_i)^2$	24.7±0.2	0.27±0.003	20.6 ±1.1
MAE	$\frac{1}{N}\sum\lvert Y_i - \hat{Y}_i\rvert$	$\mathbf{23.9 \pm 0.2}$	$\mathbf{0.28 \pm 0.007}$	$\mathbf{18.2 \pm 0.6}$
$MAPE$	$\frac{100\%}{N}\sum\frac{\lvert Y_i - \hat{Y}_i\rvert}{\lvert Y_i\rvert}$	25.1±0.4	0.37±0.007	21.9 ±1.0

Результаты

Настройка гиперпараметров - метод $GridSearchCV$ поиска по сетке.

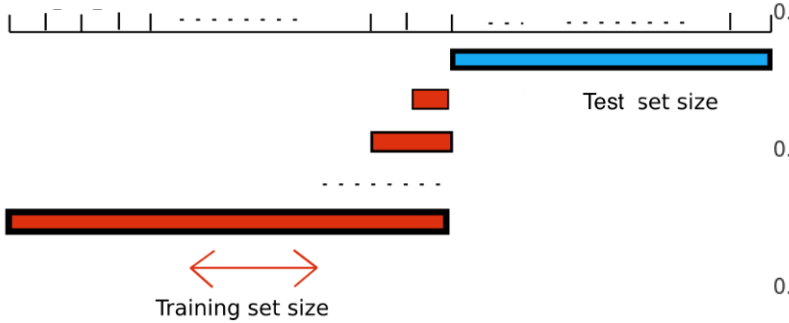
Результаты нейронных сетей

Модель	Рег.	$MAPE, \%$	$MSE, (mg/m^3)^2$	$GRE, \%$
R_{CO}, R_{NM}, T	нет	25.8±0.3	0.30±0.009	24.6±1.0
R_{CO}, R_{NM}, T, CO_h	нет	24.0±0.2	0.28±0.006	18.3±0.4
R_{CO}, R_{NM}, T	L_1	25.1±0.2	0.29±0.001	20.9±0.3
R_{CO}, R_{NM}, T, CO_h	L_1	$\mathbf{24.0 \pm 0.2}$	$\mathbf{0.27 \pm 0.002}$	$\mathbf{16.8 \pm 0.5}$

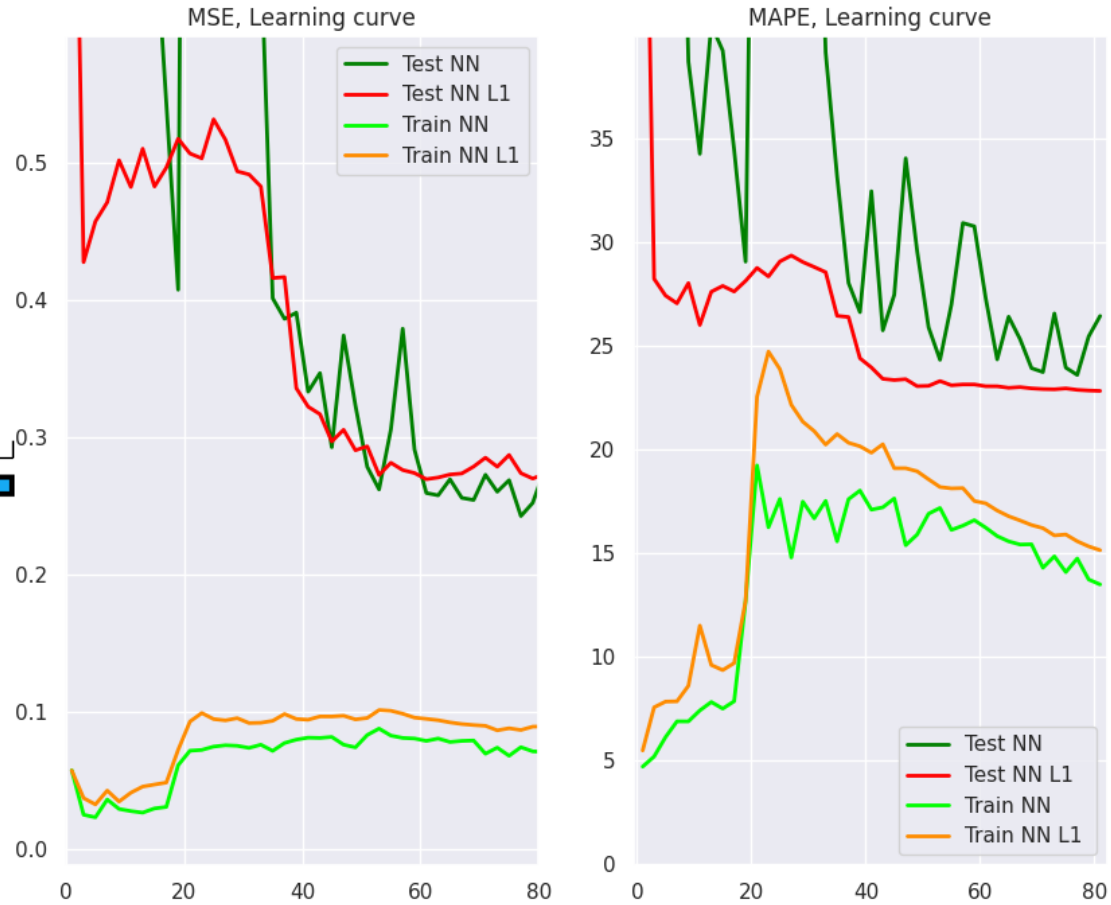
Кривые обучения

Разделение данных:

- Тест - с июня 2004 года по апрель 2005 года, 5344 измерений.
- Тренировка - от 1 и до 80 дней.



Необходимое время калибровки от 3 дней.



Кривые обучения FFNN R_{CO}, R_{NM}, T, CO_h

Заключение

Заключение

- Выполнено сравнение результатов моделей FFNN и MLR
- Выявлена переодичность целевой концентрации CO
- Новый признак позволил улучшить качество моделей
- Простые архитектуры FFNN показывают превосходство над MLR
- Время калибровки для FFNN - от 3 дней
- Время калибровки для MLR - от 10 дней

Спасибо за внимание!