ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

Отчет об исследовательском проекте на тему:

Применение машинного обучения в задаче прогнозирования погодных условий

Выполнил:

Студент группы БПМИ251
Кокурин Алексей Викторович

Приняли руководители проекта:

Ратников Федор Дмитриевич
Ведущий научный сотрудник
Факультет компьютерных наук НИУ ВШЭ

Бугаев Егор Петрович
Стажер-исследователь
Факультет компьютерных наук НИУ ВШЭ

Москва 2025

Содержание

1 Введение **3**

2 Цель **3**

3 Практическая часть **4**

3.1 Линейная зависимость без шума 4

3.2 Линейная зависимость с добавлением шума 4

3.3 Прогноз значения и неопределённости следующего шага 4

3.4 Нелинейная зависимость и нейросетевая модель 5

3.5 Нелинейная зависимость и предсказание (x_{i+1}, σ_{i+1}) только из x_i 7

4 Обзор литературы **7**

1 Введение

Прогнозирование погодных условий относится к классу задач прогнозирования временных рядов: температуры, давления, влажности и др. Все эти погодные параметры зависят как от закономерных компонент, так и от случайных факторов (шума).

Особенность предсказания погоды заключается в том, что часто требуется получить прогноз на несколько шагов вперёд: через час, через 5 часов, через сутки. При этом модель, обученная на прогноз одного шага вперёд, может применяться итеративно, образуя цепочку предсказаний. Такая модель называется авторегрессионной и предполагает, что текущее значение связано с предыдущим:

$$x_t = f(x_{t-1}) + \varepsilon(x_{t-1})$$

Однако в данном случае возникает проблема — ошибка и неопределённость модели накапливаются. Поэтому ключевая цель моей работы заключается не только в построении модели прогнозирования, но и в разработке метода оценки уверенности модели на каждом шаге с учётом нового шума (случайных факторов), а также неопределённости предыдущего шага модели.

2 Цель

Целью данной работы является исследование возможностей применения моделей машинного обучения для одношагового прогнозирования, описывающих динамику погодных параметров. Также в рамках работы предполагается разработка алгоритма итеративного многошагового прогнозирования, при котором одношаговая модель используется последовательно для получения прогноза на несколько временных шагов вперёд. Особое внимание уделяется анализу и оценке неопределенности модели в своих прогнозах на каждом шаге, включая как влияние случайных факторов (шума), так и накопление ошибок и неопределённости, передающихся от предыдущих шагов модели. Конечной целью является построение модели, позволяющей не только получать прогнозные значения, но и сопровождать их корректной оценкой уверенности, что особенно важно для практических задач прогнозирования.

3 Практическая часть

Практическую часть работы я начал с экспериментов на игрушечных данных. Такой подход позволяет заранее задать истинную зависимость и уровень шума, а значит — точно проверить, насколько модель восстанавливает параметры и оценивает неопределённость.

3.1 Линейная зависимость без шума

Данные генерируются по формуле

$$x_{i+1} = f(x_i) = \alpha x_i.$$

Обучающая выборка формируется как пары $(x_0, y_0 = \alpha x_0)$. Обучение модели проводится с помощью линейной регрессии. Точность получилась равной 100%, поскольку в этом случае присутствует строгая линейная зависимость и отсутствует шум.

3.2 Линейная зависимость с добавлением шума

$$x_{i+1} = \alpha x_i + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma)$$

Таким образом, каждое наблюдение содержит случайное отклонение, имитирующее шум измерений или влияние неучтённых факторов.

Обучающая выборка формируется аналогично предыдущей задаче, однако к целевой переменной добавляется шум с заданным стандартным отклонением σ. Для обучения используется линейная регрессия. Точность в этой задаче составила 98%, то есть модель научилась достаточно точно прогнозировать не только среднее значение, но и корректно оценивать шум в данных.

3.3 Прогноз значения и неопределённости следующего шага

В третьей задаче модель обучается прогнозировать не только следующее значение временного ряда, но и его неопределённость. Входными данными являются пары (x_i, σ_i), а выходом модели — пары (x_{i+1}, σ_{i+1}).

Генерация данных основана на линейной зависимости с шумом, при этом дисперсия

следующего шага вычисляется по формуле:

$$\sigma_{i+1}^2 = \alpha^2 \sigma_i^2 + q^2,$$

где q — новый шум, добавляемый на текущем шаге. Также эта модель использовалась для итеративного прогнозирования: предсказанные значения и дисперсии подаются обратно на вход модели. При увеличении длины цепочки неопределённость и ошибка прогноза накапливаются, что соответствует теоретическим ожиданиям.

3.4 Нелинейная зависимость и нейросетевая модель

В данной задаче линейная зависимость заменяется нелинейной функцией:

$$x_{i+1} = e^{\alpha x_i} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma)$$

что делает невозможным точное предсказание с помощью линейных методов. Для решения задачи используется нейросеть, которая по входным данным (x_i, σ_i) предсказывает (x_{i+1}, σ_{i+1}).

Архитектура модели состоит из общего блока и двух выходных голов: одна отвечает за прогноз значения, другая — за оценку неопределённости. Для того чтобы дисперсия оставалась неотрицательной, применяется функция активации Softplus. Функция потерь в этой модели — среднеквадратичная ошибка для обоих выходов.

Многошаговое прогнозирование показывает, что при увеличении числа итераций точность снижается, а неопределённость возрастает, что демонстрирует эффект накопления ошибок в нелинейных авторегрессионных моделях.

На рисунке 3.1 видно, что на первых шагах модель достаточно точно предсказывает значение и его разброс. По мере увеличения числа шагов предсказания начинают отклоняться от истинных значений, а область неопределённости расширяется.

На рисунке 3.2 модель допускает существенно большую ошибку при оценке дисперсии по сравнению с ошибкой прогнозирования самого значения. Это связано с особенностями генерации обучающих данных: дисперсия в процессе растёт быстрее, чем само значение функции. Поэтому при итеративном прогнозировании модель уже на сравнительно малых шагах сталкивается с диапазонами дисперсии, которые были слабо представлены или отсутствовали в обучающей выборке. В результате ошибка оценки неопределённости накапливается и увеличивается быстрее, чем ошибка среднего значения.

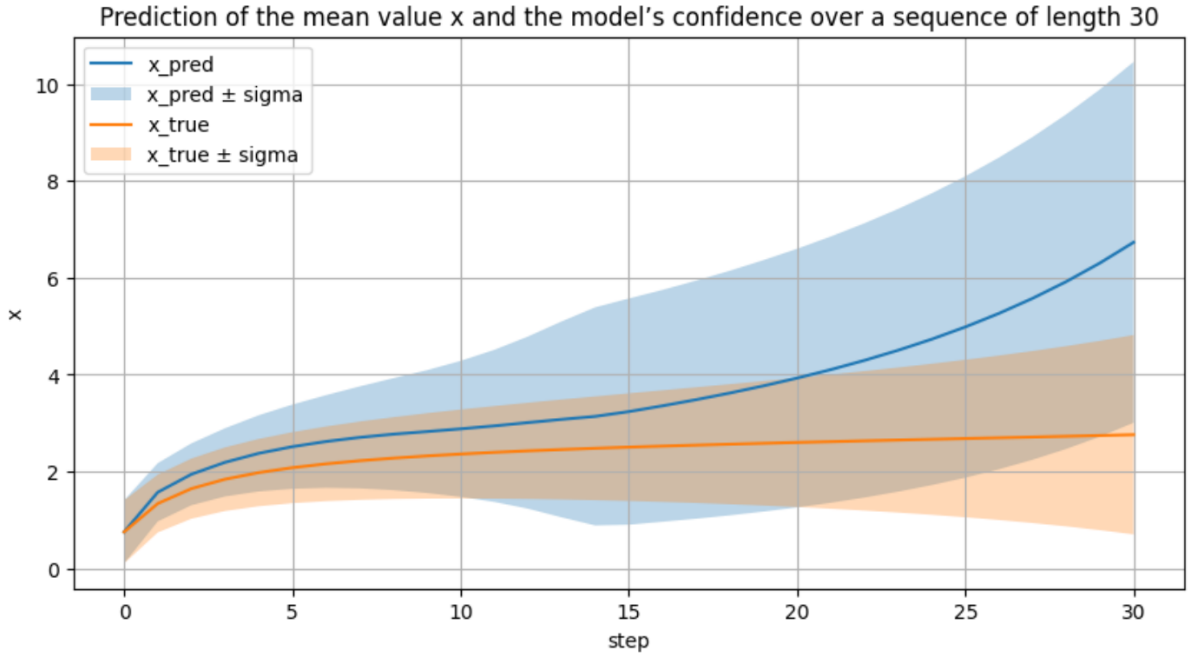

Рис. 3.1: Сравнение истинного и предсказанного среднего значения временного ряда с оценкой неопределённости в задаче многошагового авторегрессионного прогнозирования

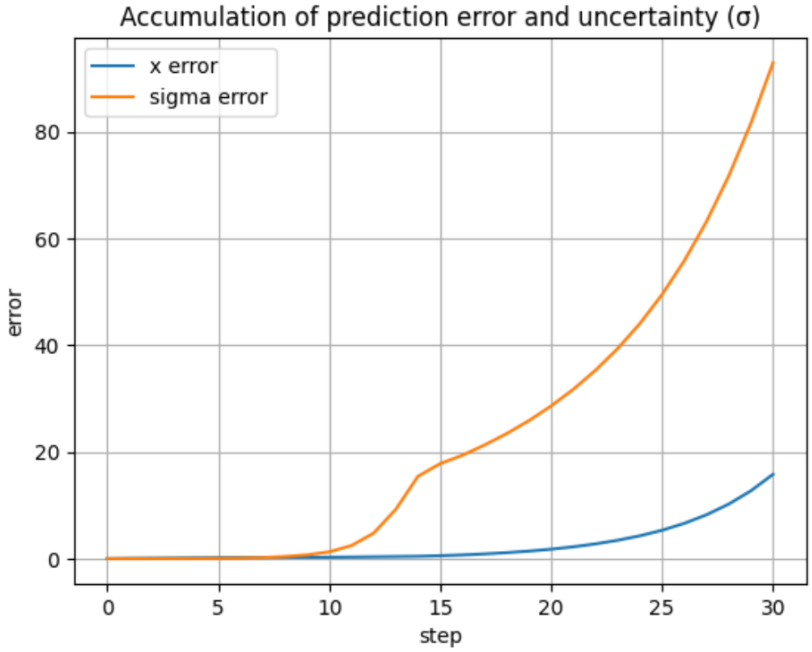

Рис. 3.2: Рост ошибки прогноза среднего значения и ошибки оценки неопределённости в процессе многошагового авторегрессионного прогнозирования

3.5 Нелинейная зависимость и предсказание (x_{i+1}, σ_{i+1}) только из x_i

В пятой задаче используется вероятностная постановка задачи, в которой нейросеть обучается напрямую предсказывать параметры распределения следующего шага. Предполагается, что значение x_{i+1} имеет нормальное распределение с математическим ожиданием $\mu(x_i)$ и стандартным отклонением $\sigma(x_i)$, которые предсказываются моделью.

Обучение проводится путём минимизации отрицательного логарифма правдоподобия нормального распределения:

$$\mathcal{L} = \frac{(x_{i+1} - \mu)^2}{2\sigma^2} + \log \sigma$$

Использование отрицательного логарифма правдоподобия в качестве функции потерь позволяет модели одновременно обучаться точному прогнозу значения и корректной оценке неопределённости. В результате модель не переоценивает свою уверенность и может явно сигнализировать о снижении точности при росте горизонта прогнозирования.

Как и в предыдущей задаче, результаты показывают, что при увеличении количества шагов ошибка и неопределённость накапливаются, что является фундаментальным ограничением одношаговых авторегрессионных моделей.

4 Обзор литературы

При выполнении практической части работы и реализации моделей машинного обучения потребовались знания теории вероятностей и математической статистики. В частности, для построения и обучения вероятностных моделей потребовалось изучение таких тем, как случайные величины и их распределения, плотность распределения вероятности, нормальное и многомерное нормальное распределения, а также правдоподобие. Эти методы использовались для оценки параметров распределений и генерации обучающих выборок модели.

Теоретической базой исследования послужили фундаментальные работы по теории вероятностей и математической статистике [1, 2, 3], а также современные источники по статистическому выводу и машинному обучению [4, 5].

Также я прочитал несколько статей. Одна из них — FourCastNet [6] — это модель машинного обучения для глобального прогноза погоды, разработанная NVIDIA. Ее основная идея заключается в том, что вместо решения физических уравнений FourCastNet обучается на реальных метеоданных и предсказывает погодные условия как задачу операторного обучения. FourCastNet стала одной из первых моделей, показавших, что методы машинного обучения могут конкурировать с лучшими физическими моделями прогноза погоды.

Другая статья, которую я прочитал, — WeatherGFT [7] — это новая модель для прогноза погоды, предложенная как гибридная система, сочетающая физику и машинное обучение. Ее основная цель — научиться делать прогнозы не только на большие интервалы (6 часов), но и на гораздо более детальные временные масштабы (30 минут), даже в случаях, когда такие данные отсутствуют в обучающем наборе.

Список литературы (или источников)

[1] А. Н. Ширяев. *Вероятности и случайные процессы.* МЦНМО, Москва, 2019.

[2] Е. С. Вентцель. *Теория вероятностей.* Высшая школа, Москва, 2006.

[3] М. Х. Дойч. *Введение в математическую статистику.* Наука, Москва, 1984.

[4] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference.* Springer, New York, 2004.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Springer, New York, 2006.

[6] Jaideep Pathak, Sanjeev Subramanian, Peter Harrington, Santhosh Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Anima Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[7] Weathergft: Learning weather forecasting at arbitrary temporal resolutions via physics-guided fine-tuning. 2024.