

Методы машинного обучения в задаче предсказания погоды

Елена Волжина

руководитель Е.Г. Михайлова

СПбГУ, мат-мех, кафедра ИАС

31 мая 2018г

Введение

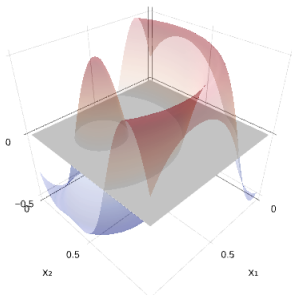
- ▶ Погода – состояние нижнего слоя атмосферы
- ▶ Температура, погодное явление, влажность, давление, ветер, ...
- ▶ Метеостанции, метеорадары, спутники
- ▶ Численный прогноз погоды
- ▶ Глобальные модели: GFS, JMA, ECMWF, CMC; региональная: WRF
- ▶ Машинное обучение: The Weather Company (IBM), Яндекс.Погода

Постановка задачи

- ▶ В момент *gentime* прогнозируем погоду в момент *time*
- ▶ $X = \{x_i = (f_i^1, \dots, f_i^M), i \in \overline{1..N}\}, y = \{y_i, i \in \overline{1..N}\}$
- ▶ f_i : прогнозы поставщиков, климатические данные, вспомогательные переменные
- ▶ **Задача 1:** для модели прогноза температуры добавить показания ближайших станций в момент *gentime* (на 7 часов вперёд)
- ▶ Целевые значения – *temperature_delta*
- ▶ **Задача 2:** обучить модель для предсказания интенсивности осадков (если они есть)
- ▶ Целевые значения – агрегированные по времени и географии значения с метеорологических радаров

Постановка задачи

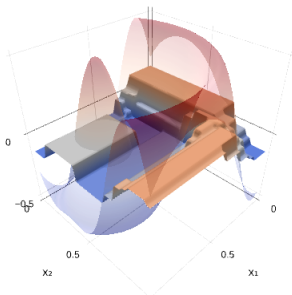
- ▶ Градиентный бустинг над решающими деревьями



Градиентный бустинг: 0 деревьев

Постановка задачи

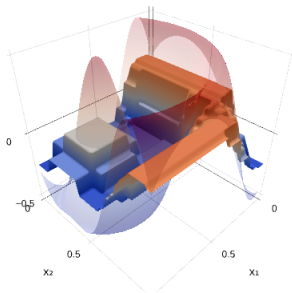
- ▶ Градиентный бустинг над решающими деревьями



Градиентный бустинг: 1 дерево глубины 6

Постановка задачи

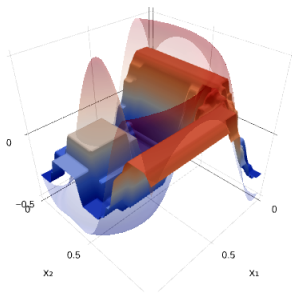
- ▶ Градиентный бустинг над решающими деревьями



Градиентный бустинг: 2 дерева глубины 6

Постановка задачи

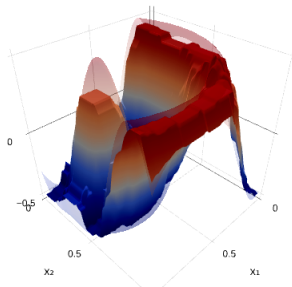
- ▶ Градиентный бустинг над решающими деревьями



Градиентный бустинг: 3 дерева глубины 6

Постановка задачи

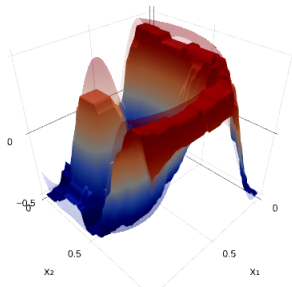
- ▶ Градиентный бустинг над решающими деревьями



Градиентный бустинг: 10 деревьев глубины 6

Постановка задачи

- ▶ Градиентный бустинг над решающими деревьями

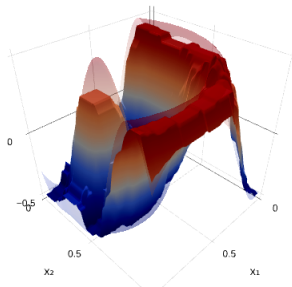


Градиентный бустинг: 10 деревьев глубины 6

- ▶ В Яндексе: Matrixnet, CatBoost

Постановка задачи

- ▶ Градиентный бустинг над решающими деревьями

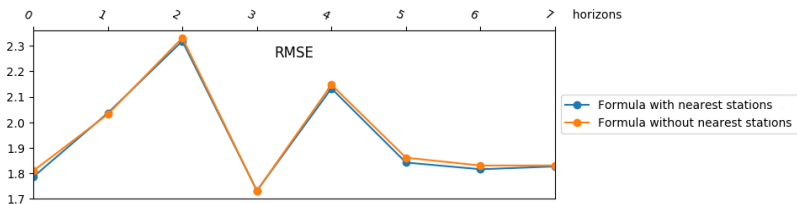


Градиентный бустинг: 10 деревьев глубины 6

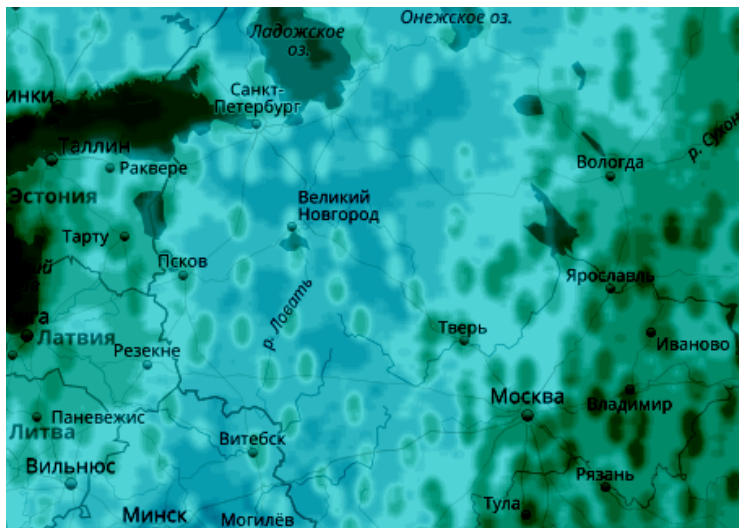
- ▶ В Яндексе: Matrixnet, CatBoost
- ▶ Метрика: $RMSE(y, \hat{y}) = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$

Эксперименты с температурой: начальное решение

Модели, обученные на всех краткосрочных горизонтах



Эксперименты с температурой: только первые N часов

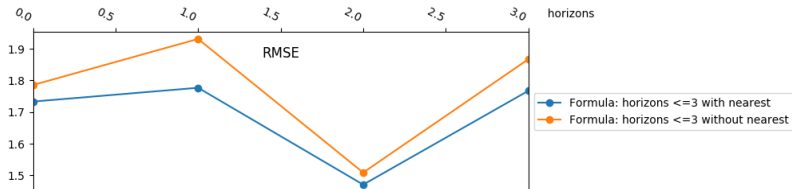


Эксперименты с температурой: сглаживание

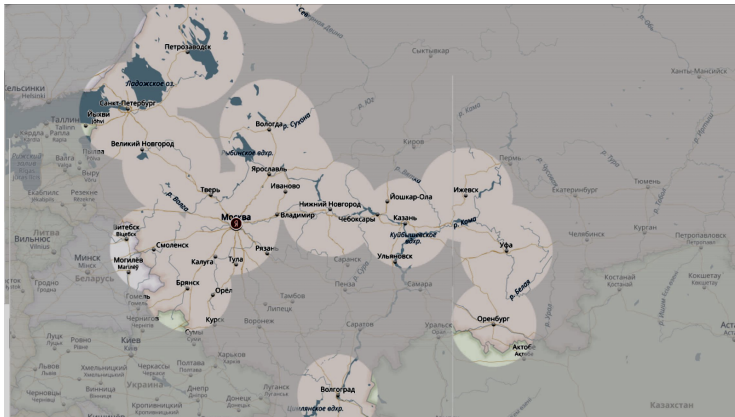
smoothed_fact_temperature_delta :=

$$\begin{cases} fact \cdot \left(1 - \frac{dist}{max_distance}\right) + forecast \cdot \frac{dist}{max_distance}, & \text{if } dist \leq max_distance \\ forecast, & \text{if } dist > max_distance \end{cases},$$

где *fact* – исходное значение с одной из ближайших станций,
dist – расстояние до неё, *forecast* – прогноз поставщика



Эксперименты с осадками



Обработка радаров:

- ▶ склейка
- ▶ фильтрация
- ▶ агрегация

Эксперименты с осадками: начальное решение

- ▶ один месяц, точки станций
- ▶ агрегация по времени за час и по географии в радиусе 3 километров
- ▶ пороги для отсеечения шумовых значений

Результаты неудовлетворительные, по RMSE модель не выигрывает прогнозы поставщиков, быстрое переобучение.

Эксперименты с осадками: проверки данных

- ▶ согласованность радаров: предыдущий час в признаки
- ▶ согласованность поставщиков: предсказывание одного по другим
- ▶ упрощение задачи: классификация

Эксперименты с осадками: улучшение сбора данных

- ▶ три месяца, точки сетки
- ▶ агрегация по времени за три часа и по географии в радиусе 30 километров
- ▶ дисбаланс в значениях:
 - ▶ логарифмирование
 - ▶ отдельно $(0, 1]$ и $(1, +\infty)$
 - ▶ искусственное балансирование данных

CatBoost

Эксперименты с осадками: промежуточные результаты

	model RMSE	provider RMSE
simple regression, any target	0.5386	0.8431
log precipitation, any target	0.5482	0.8431
simple regression @ (0, 1]	0.1890	0.7165
log precipitation @ (0, 1]	0.1898	0.7165

обычная обучающая выборка

	model RMSE	provider RMSE
simple regression	1.6231	1.8527
log precipitation	1.6977	1.8527
simple regression + weights	1.6695	1.8527
log precipitation + weights	1.7330	1.8527

сбалансированная обучающая выборка

Заключение

Улучшение прогнозов температуры

- ▶ данные с ближайших станций наиболее полезны первые 3 часа
- ▶ потребовалось сглаживание
- ▶ модель используется в Яндекс.Погоде

Прогноз интенсивности осадков

- ▶ агрегация данных с радаров
- ▶ модель готова для использования