Содержание

1. В чем проблема классического подхода?
2. Переход к задаче обучения с учителем.
3. Добавление произвольных признаков.
4. Datetime признаки.
5. Скользящие статистики.
6. Обучение произвольной модели.
7. Кроссвалидация для временных рядов.

В чем проблема классического подхода?

- Необходимость ручного выбора параметров
- Ограниченность в выборе модели - линейная регрессия
- Нет возможности добавить произвольные признаки в модель
- Несовместимое с большинством современных пакетов API.

Что делать?

Ответ - переход к задаче обучения с учителем.

Что делать?

1. Выбрать тип решаемой задачи
2. Перевести временной ряд в матрицу объекты-признаки
3. Добавить дополнительные признаки по желанию
4. Обучить произвольную модель
5. Profit!

Выбор типа решаемой задачи

Предсказание одной точки вперед

$$(x_{ij}, \ldots, x_{nj}) -> y_j$$

Предсказание вектора точек

$$(x_{ij}, \ldots, x_{nj}) -> (y_{j1}, \ldots, y_{pj})$$

Переход к матрице лагов

Используем метод скользящего окна

$$(x_j; y_j) = (t_{j-1}, \ldots, t_{j-w}; t_j) \forall j \in (w, len(t))$$

w - ширина окна, t - исходный временной ряд

Переход к матрице лагов

2019-05-19 04:00:00	1.0
2019-05-19 05:00:00	6.0
2019-05-19 06:00:00	11.0
2019-05-19 07:00:00	16.0
2019-05-19 08:00:00	26.0
2019-05-19 09:00:00	38.0
2019-05-19 10:00:00	52.0
2019-05-19 11:00:00	81.0
2019-05-19 12:00:00	90.0
2019-05-19 13:00:00	25.0
2019-05-19 14:00:00	103.0
2019-05-19 15:00:00	87.0
2019-05-19 16:00:00	61.0
2019-05-19 17:00:00	48.0
2019-05-19 18:00:00	46.0
2019-05-19 19:00:00	33.0
2019-05-19 20:00:00	26.0
2019-05-19 21:00:00	8.0
2019-05-20 00:00:00	14.0
2019-05-20 01:00:00	6.0
2019-05-20 02:00:00	4.0
2019-05-20 03:00:00	14.0
2019-05-20 04:00:00	20.0
2019-05-20 05:00:00	30.0
2019-05-20 06:00:00	80.0

	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	season_lag	y
2019-05-19 11:00:00	52.0	38.0	26.0	16.0	11.0	6.0	1.0	1.0	81.0
2019-05-19 12:00:00	81.0	52.0	38.0	26.0	16.0	11.0	6.0	6.0	90.0
2019-05-19 13:00:00	90.0	81.0	52.0	38.0	26.0	16.0	11.0	11.0	25.0
2019-05-19 14:00:00	25.0	90.0	81.0	52.0	38.0	26.0	16.0	16.0	103.0
2019-05-19 15:00:00	103.0	25.0	90.0	81.0	52.0	38.0	26.0	26.0	87.0
2019-05-19 16:00:00	87.0	103.0	25.0	90.0	81.0	52.0	38.0	38.0	61.0
2019-05-19 17:00:00	61.0	87.0	103.0	25.0	90.0	81.0	52.0	52.0	48.0
2019-05-19 18:00:00	48.0	61.0	87.0	103.0	25.0	90.0	81.0	81.0	46.0
2019-05-19 19:00:00	46.0	48.0	61.0	87.0	103.0	25.0	90.0	90.0	33.0
2019-05-19 20:00:00	33.0	46.0	48.0	61.0	87.0	103.0	25.0	25.0	26.0
2019-05-19 21:00:00	26.0	33.0	46.0	48.0	61.0	87.0	103.0	103.0	8.0
2019-05-20 00:00:00	8.0	26.0	33.0	46.0	48.0	61.0	87.0	87.0	14.0
2019-05-20 01:00:00	14.0	8.0	26.0	33.0	46.0	48.0	61.0	61.0	6.0
2019-05-20 02:00:00	6.0	14.0	8.0	26.0	33.0	46.0	48.0	48.0	4.0
2019-05-20 03:00:00	4.0	6.0	14.0	8.0	26.0	33.0	46.0	46.0	14.0
2019-05-20 04:00:00	14.0	4.0	6.0	14.0	8.0	26.0	33.0	33.0	20.0
2019-05-20 05:00:00	20.0	14.0	4.0	6.0	14.0	8.0	26.0	26.0	30.0
2019-05-20 06:00:00	30.0	20.0	14.0	4.0	6.0	14.0	8.0	8.0	80.0

Добавление произвольных признаков

Произвольные признаки - любая внешняя по отношению к самому ряду информация. Фаза луны, номер счета и т.д.

$$f(t) : timestamp -> value$$

Временные признаки

Datetime признаки - возможность добавить информацию о дне недели, месяца, времени суток и так далее.

Во многих случаях решает проблему автоматического выбора периода сезонности.

Временные признаки

	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	season_lag	y
2019-05-19 11:00:00	52.0	38.0	26.0	16.0	11.0	6.0	1.0	1.0	81.0
2019-05-19 12:00:00	81.0	52.0	38.0	26.0	16.0	11.0	6.0	6.0	90.0
2019-05-19 13:00:00	90.0	81.0	52.0	38.0	26.0	16.0	11.0	11.0	25.0
2019-05-19 14:00:00	25.0	90.0	81.0	52.0	38.0	26.0	16.0	16.0	103.0
2019-05-19 15:00:00	103.0	25.0	90.0	81.0	52.0	38.0	26.0	26.0	87.0
2019-05-19 16:00:00	87.0	103.0	25.0	90.0	81.0	52.0	38.0	38.0	61.0
2019-05-19 17:00:00	61.0	87.0	103.0	25.0	90.0	81.0	52.0	52.0	48.0
2019-05-19 18:00:00	48.0	61.0	87.0	103.0	25.0	90.0	81.0	81.0	46.0
2019-05-19 19:00:00	46.0	48.0	61.0	87.0	103.0	25.0	90.0	90.0	33.0
2019-05-19 20:00:00	33.0	46.0	48.0	61.0	87.0	103.0	25.0	25.0	26.0
2019-05-19 21:00:00	26.0	33.0	46.0	48.0	61.0	87.0	103.0	103.0	8.0
2019-05-20 00:00:00	8.0	26.0	33.0	46.0	48.0	61.0	87.0	87.0	14.0
2019-05-20 01:00:00	14.0	8.0	26.0	33.0	46.0	48.0	61.0	61.0	6.0
2019-05-20 02:00:00	6.0	14.0	8.0	26.0	33.0	46.0	48.0	48.0	4.0
2019-05-20 03:00:00	4.0	6.0	14.0	8.0	26.0	33.0	46.0	46.0	14.0
2019-05-20 04:00:00	14.0	4.0	6.0	14.0	8.0	26.0	33.0	33.0	20.0
2019-05-20 05:00:00	20.0	14.0	4.0	6.0	14.0	8.0	26.0	26.0	30.0
2019-05-20 06:00:00	30.0	20.0	14.0	4.0	6.0	14.0	8.0	8.0	80.0

	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	season_lag	y	weekday	monthday	is_weekend	month	hour
2019-05-19 11:00:00	52.0	38.0	26.0	16.0	11.0	6.0	1.0	1.0	81.0	6	19	0	5	11
2019-05-19 12:00:00	81.0	52.0	38.0	26.0	16.0	11.0	6.0	6.0	90.0	6	19	0	5	12
2019-05-19 13:00:00	90.0	81.0	52.0	38.0	26.0	16.0	11.0	11.0	25.0	6	19	0	5	13
2019-05-19 14:00:00	25.0	90.0	81.0	52.0	38.0	26.0	16.0	16.0	103.0	6	19	0	5	14
2019-05-19 15:00:00	103.0	25.0	90.0	81.0	52.0	38.0	26.0	26.0	87.0	6	19	0	5	15
2019-05-19 16:00:00	87.0	103.0	25.0	90.0	81.0	52.0	38.0	38.0	61.0	6	19	0	5	16
2019-05-19 17:00:00	61.0	87.0	103.0	25.0	90.0	81.0	52.0	52.0	48.0	6	19	0	5	17
2019-05-19 18:00:00	48.0	61.0	87.0	103.0	25.0	90.0	81.0	81.0	46.0	6	19	0	5	18
2019-05-19 19:00:00	46.0	48.0	61.0	87.0	103.0	25.0	90.0	90.0	33.0	6	19	0	5	19
2019-05-19 20:00:00	33.0	46.0	48.0	61.0	87.0	103.0	25.0	25.0	26.0	6	19	0	5	20
2019-05-19 21:00:00	26.0	33.0	46.0	48.0	61.0	87.0	103.0	103.0	8.0	6	19	0	5	21
2019-05-20 00:00:00	8.0	26.0	33.0	46.0	48.0	61.0	87.0	87.0	14.0	0	20	1	5	0
2019-05-20 01:00:00	14.0	8.0	26.0	33.0	46.0	48.0	61.0	61.0	6.0	0	20	1	5	1
2019-05-20 02:00:00	6.0	14.0	8.0	26.0	33.0	46.0	48.0	48.0	4.0	0	20	1	5	2
2019-05-20 03:00:00	4.0	6.0	14.0	8.0	26.0	33.0	46.0	46.0	14.0	0	20	1	5	3
2019-05-20 04:00:00	14.0	4.0	6.0	14.0	8.0	26.0	33.0	33.0	20.0	0	20	1	5	4
2019-05-20 05:00:00	20.0	14.0	4.0	6.0	14.0	8.0	26.0	26.0	30.0	0	20	1	5	5
2019-05-20 06:00:00	30.0	20.0	14.0	4.0	6.0	14.0	8.0	8.0	80.0	0	20	1	5	6

Скользящие статистики

Среднее, медиана, стандартное отклонение и прочие статистики, вычисленные на лагах каждого объекта обучающей выборки.

Помогают “стабилизировать” матожидание и дисперсию предсказаний.

Скользящие статистики

	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	season_lag	y
2019-05-19 11:00:00	52.0	38.0	26.0	16.0	11.0	6.0	1.0	1.0	81.0
2019-05-19 12:00:00	81.0	52.0	38.0	26.0	16.0	11.0	6.0	6.0	90.0
2019-05-19 13:00:00	90.0	81.0	52.0	38.0	26.0	16.0	11.0	11.0	25.0
2019-05-19 14:00:00	25.0	90.0	81.0	52.0	38.0	26.0	16.0	16.0	103.0
2019-05-19 15:00:00	103.0	25.0	90.0	81.0	52.0	38.0	26.0	26.0	87.0
2019-05-19 16:00:00	87.0	103.0	25.0	90.0	81.0	52.0	38.0	38.0	61.0
2019-05-19 17:00:00	61.0	87.0	103.0	25.0	90.0	81.0	52.0	52.0	48.0
2019-05-19 18:00:00	48.0	61.0	87.0	103.0	25.0	90.0	81.0	81.0	46.0
2019-05-19 19:00:00	46.0	48.0	61.0	87.0	103.0	25.0	90.0	90.0	33.0
2019-05-19 20:00:00	33.0	46.0	48.0	61.0	87.0	103.0	25.0	25.0	26.0
2019-05-19 21:00:00	26.0	33.0	46.0	48.0	61.0	87.0	103.0	103.0	8.0
2019-05-20 00:00:00	8.0	26.0	33.0	46.0	48.0	61.0	87.0	87.0	14.0
2019-05-20 01:00:00	14.0	8.0	26.0	33.0	46.0	48.0	61.0	61.0	6.0
2019-05-20 02:00:00	6.0	14.0	8.0	26.0	33.0	46.0	48.0	48.0	4.0
2019-05-20 03:00:00	4.0	6.0	14.0	8.0	26.0	33.0	46.0	46.0	14.0
2019-05-20 04:00:00	14.0	4.0	6.0	14.0	8.0	26.0	33.0	33.0	20.0
2019-05-20 05:00:00	20.0	14.0	4.0	6.0	14.0	8.0	26.0	26.0	30.0
2019-05-20 06:00:00	30.0	20.0	14.0	4.0	6.0	14.0	8.0	8.0	80.0

	lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	season_lag	y	weekday	monthday	is_weekend	month	hour	mean	std
2019-05-19 11:00:00	52.0	38.0	26.0	16.0	11.0	6.0	1.0	1.0	81.0	6	19	0	5	11	21.428571	18.365340
2019-05-19 12:00:00	81.0	52.0	38.0	26.0	16.0	11.0	6.0	6.0	90.0	6	19	0	5	12	32.857143	26.585890
2019-05-19 13:00:00	90.0	81.0	52.0	38.0	26.0	16.0	11.0	11.0	25.0	6	19	0	5	13	44.857143	31.029172
2019-05-19 14:00:00	25.0	90.0	81.0	52.0	38.0	26.0	16.0	16.0	103.0	6	19	0	5	14	46.857143	28.858439
2019-05-19 15:00:00	103.0	25.0	90.0	81.0	52.0	38.0	26.0	26.0	87.0	6	19	0	5	15	59.285714	31.925509
2019-05-19 16:00:00	87.0	103.0	25.0	90.0	81.0	52.0	38.0	38.0	61.0	6	19	0	5	16	68.000000	29.563491
2019-05-19 17:00:00	61.0	87.0	103.0	25.0	90.0	81.0	52.0	52.0	48.0	6	19	0	5	17	71.285714	26.824829
2019-05-19 18:00:00	48.0	61.0	87.0	103.0	25.0	90.0	81.0	81.0	46.0	6	19	0	5	18	70.714286	27.341752
2019-05-19 19:00:00	46.0	48.0	61.0	87.0	103.0	25.0	90.0	90.0	33.0	6	19	0	5	19	65.714286	28.329692
2019-05-19 20:00:00	33.0	46.0	48.0	61.0	87.0	103.0	25.0	25.0	26.0	6	19	0	5	20	57.571429	28.377557
2019-05-19 21:00:00	26.0	33.0	46.0	48.0	61.0	87.0	103.0	103.0	8.0	6	19	0	5	21	57.714286	28.188143
2019-05-20 00:00:00	8.0	26.0	33.0	46.0	48.0	61.0	87.0	87.0	14.0	0	20	1	5	0	44.142857	25.491362
2019-05-20 01:00:00	14.0	8.0	26.0	33.0	46.0	48.0	61.0	61.0	6.0	0	20	1	5	1	33.714286	19.189531
2019-05-20 02:00:00	6.0	14.0	8.0	26.0	33.0	46.0	48.0	48.0	4.0	0	20	1	5	2	25.857143	17.324632
2019-05-20 03:00:00	4.0	6.0	14.0	8.0	26.0	33.0	46.0	46.0	14.0	0	20	1	5	3	19.571429	15.873008
2019-05-20 04:00:00	14.0	4.0	6.0	14.0	8.0	26.0	33.0	33.0	20.0	0	20	1	5	4	15.000000	10.785793
2019-05-20 05:00:00	20.0	14.0	4.0	6.0	14.0	8.0	26.0	26.0	30.0	0	20	1	5	5	13.142857	7.904188
2019-05-20 06:00:00	30.0	20.0	14.0	4.0	6.0	14.0	8.0	8.0	80.0	0	20	1	5	6	13.714286	9.050125

Обучение произвольной модели

- Linear regression
- Decision trees
- Gradient boosting decision trees
- Classical neural networks
- GRU, LSTM

Отличие от обычной supervised задачи

Необходимость кроссвалидации по скользящему окну