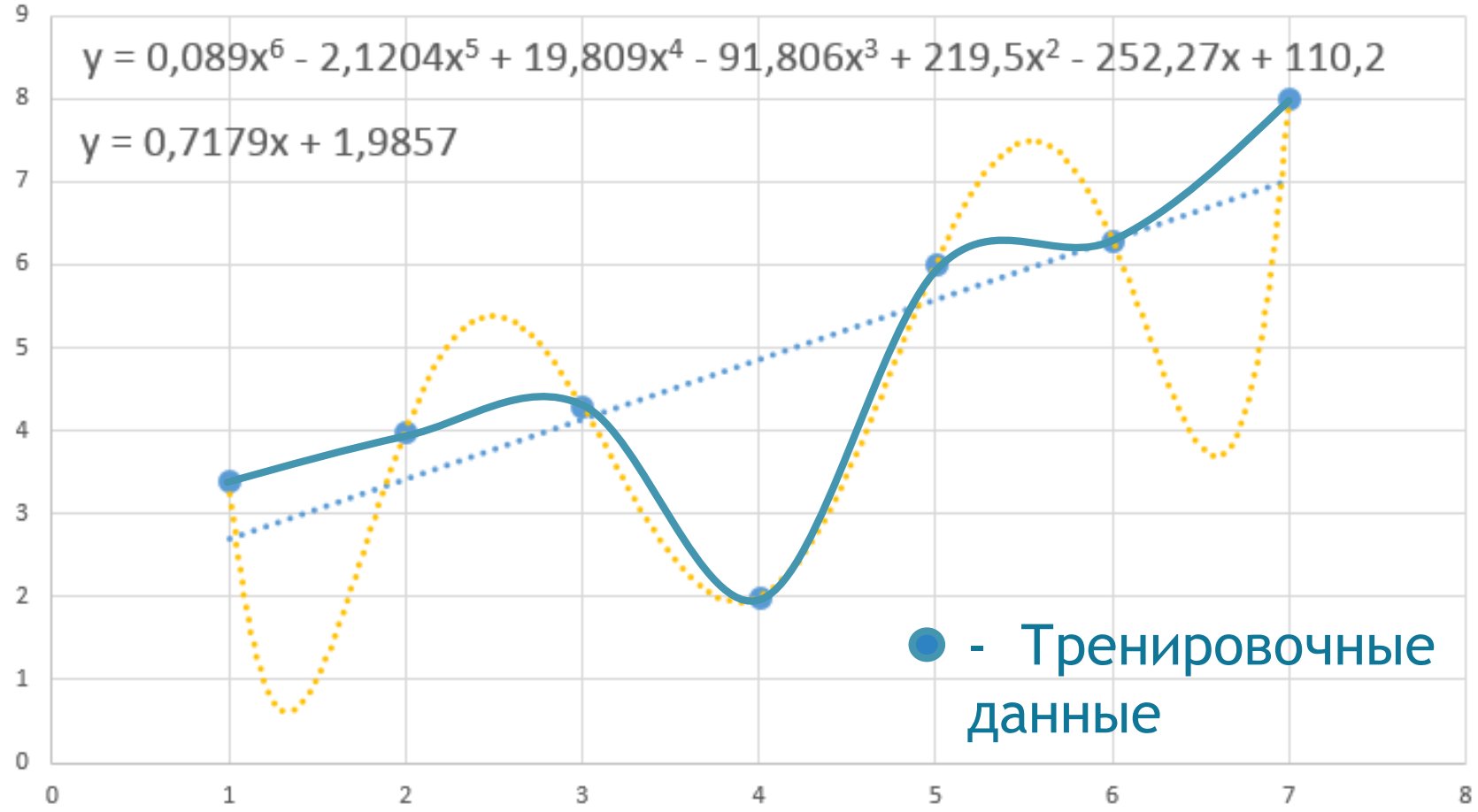
Переобучение, регуляризация

Пояснение на непрерывной переменной
По 7ми точкам процесса построили две модели

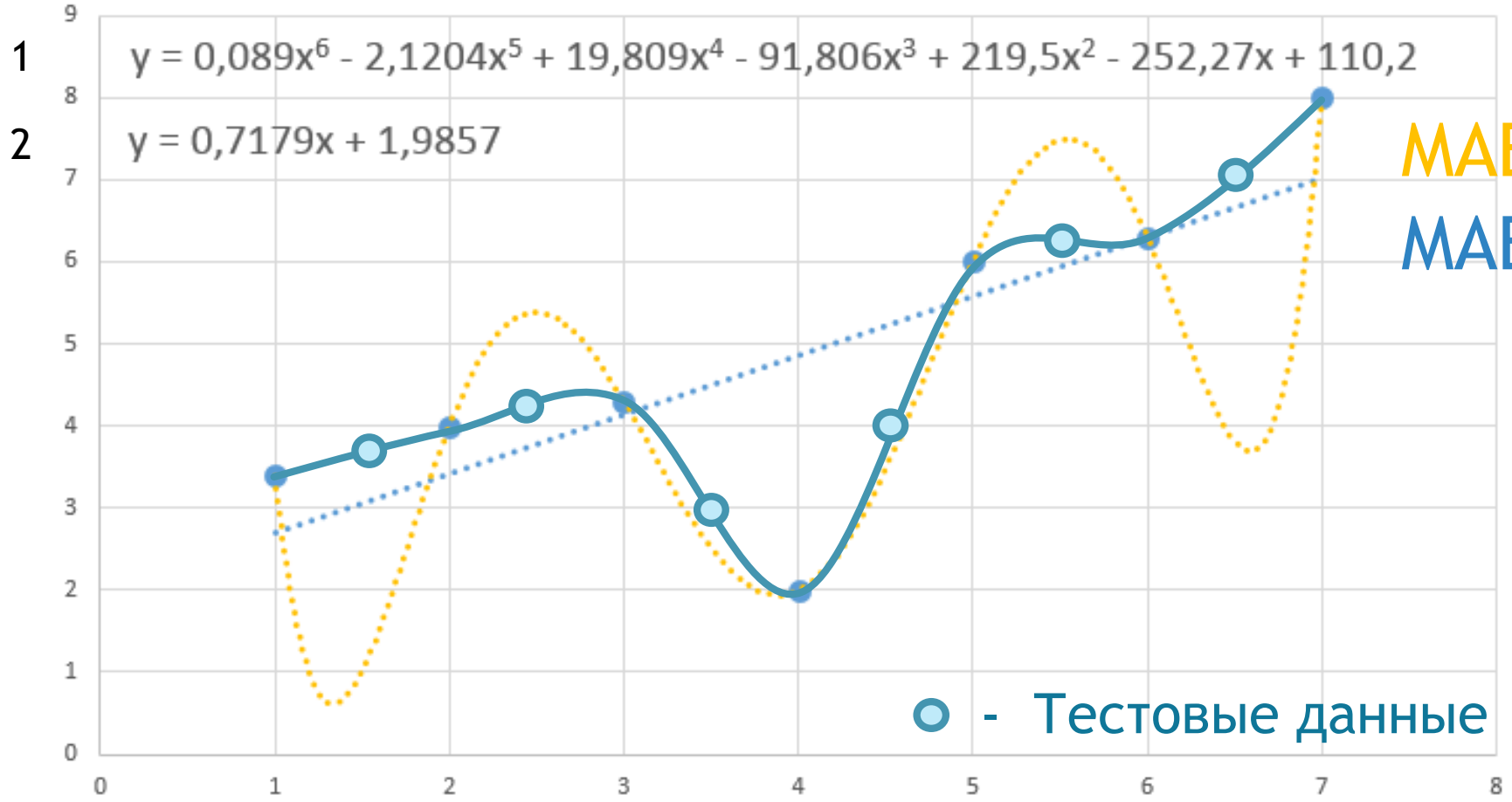
1

2



- Модель 1 – сложная
- Модель 2 - простая

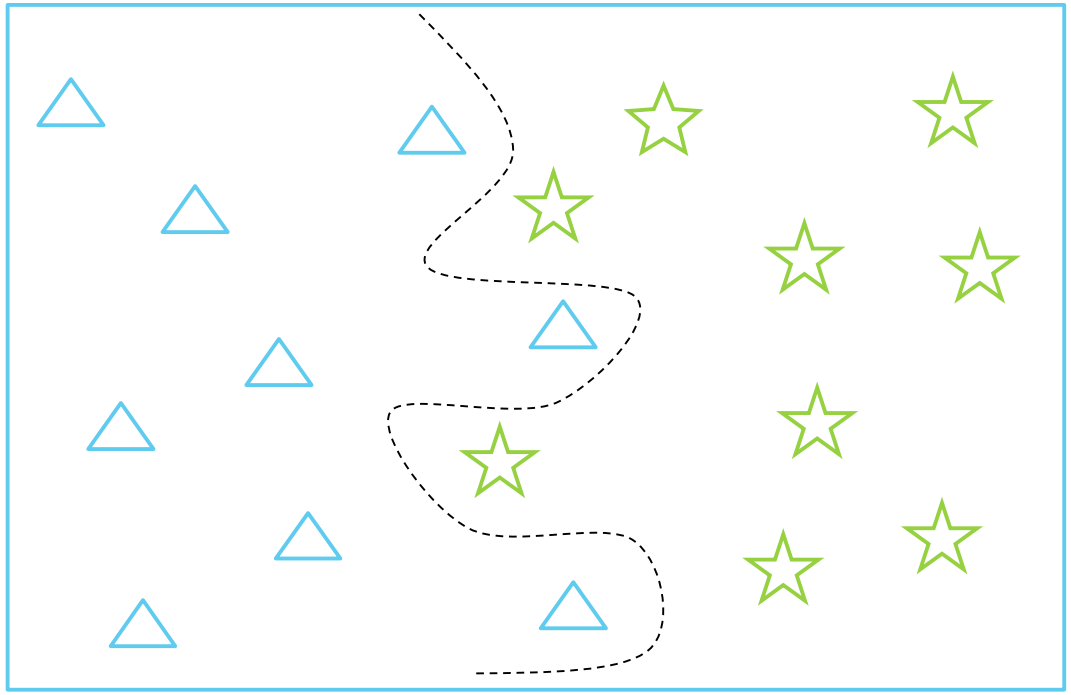
Пояснение на непрерывной переменной Прогноз f(x) по x на основе моделей



- [3;5] – прогнозы будут лучше у сложной модели
- [1;3],[5;7] – у простой

Переобучение

- Хорошая подгонка под тренировочные данные
  - Слишком сложная модель
  - Начинает отражать характеристики шума в данных, нежели исходное распределение



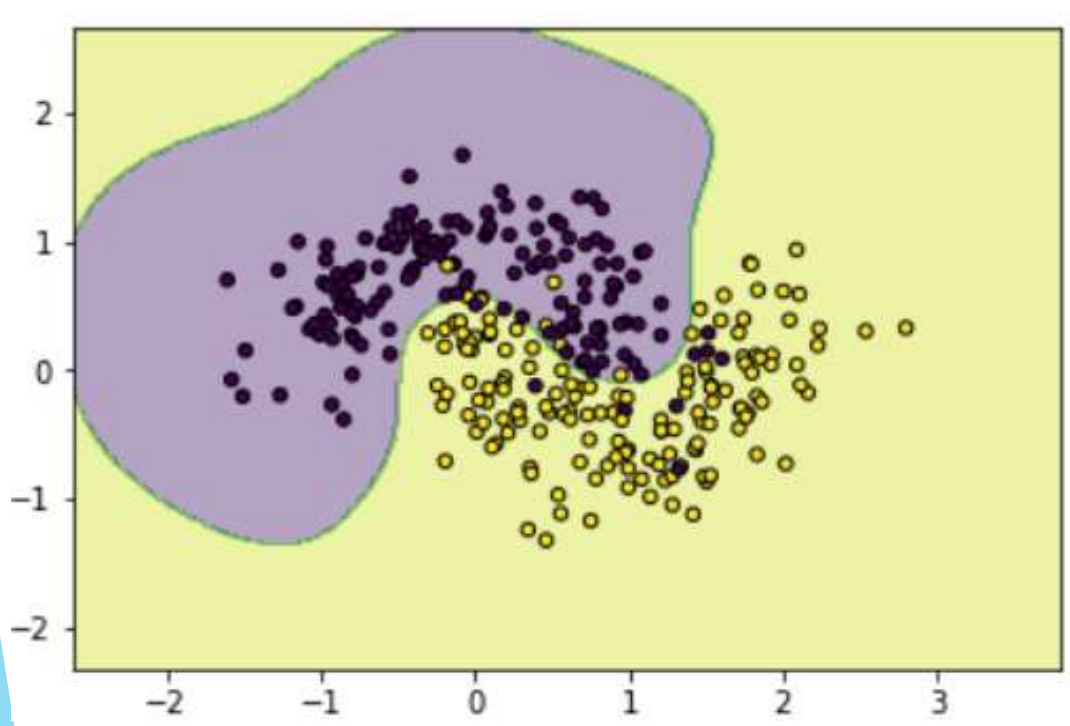
Переобученная модель (overfitted)
Имеется высокая
дисперсия (variance)

Одни и те же образцы обрабатываются по разному, в зависимости от очередности прохода.

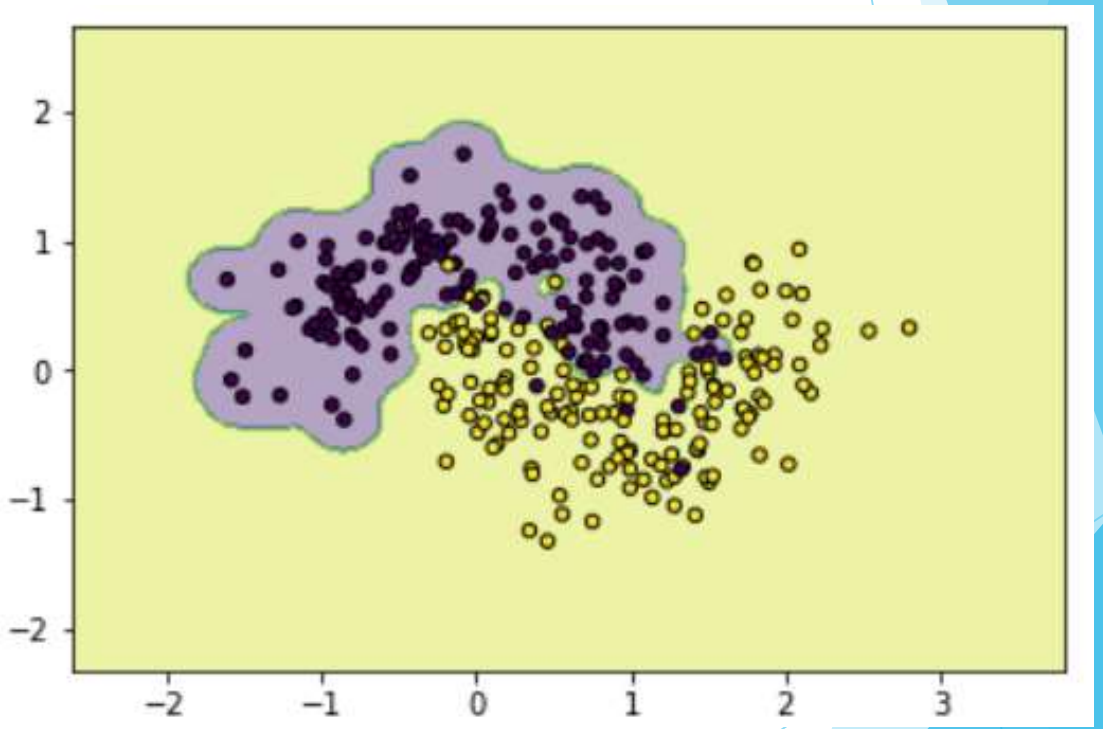
Не происходит систематизации знания.

Модель чувствительна к случайности в тренировочных данных

Классификация SVC, kernel='rbf'

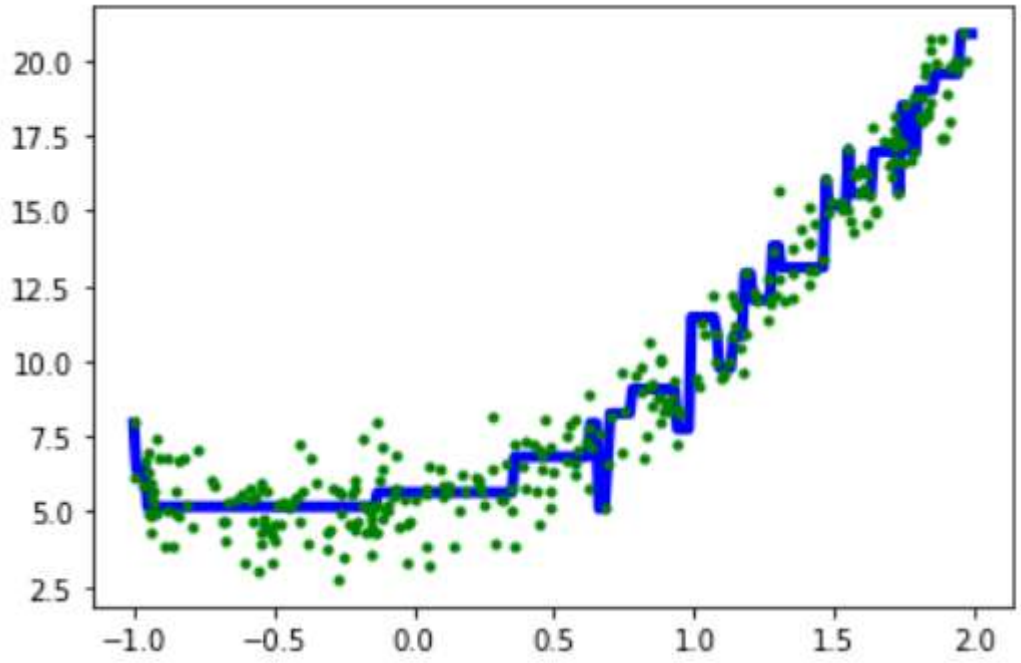


SVC(kernel='rbf', gamma=2) = 0.93

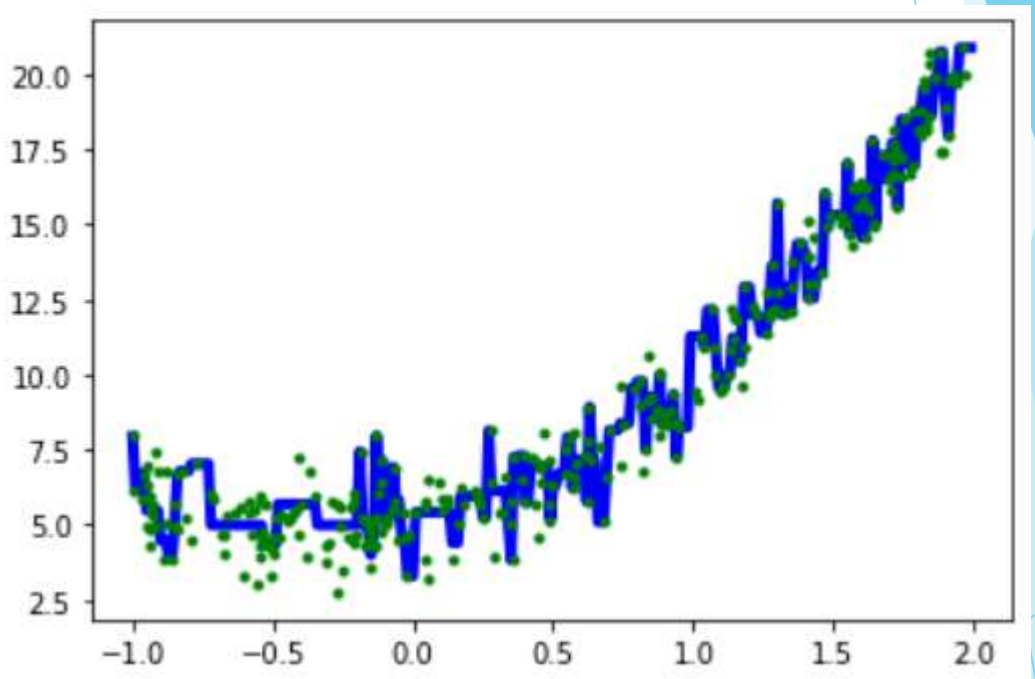


SVC(kernel='rbf', gamma=50) = 0.88

Регрессия DecisionTreeRegressor



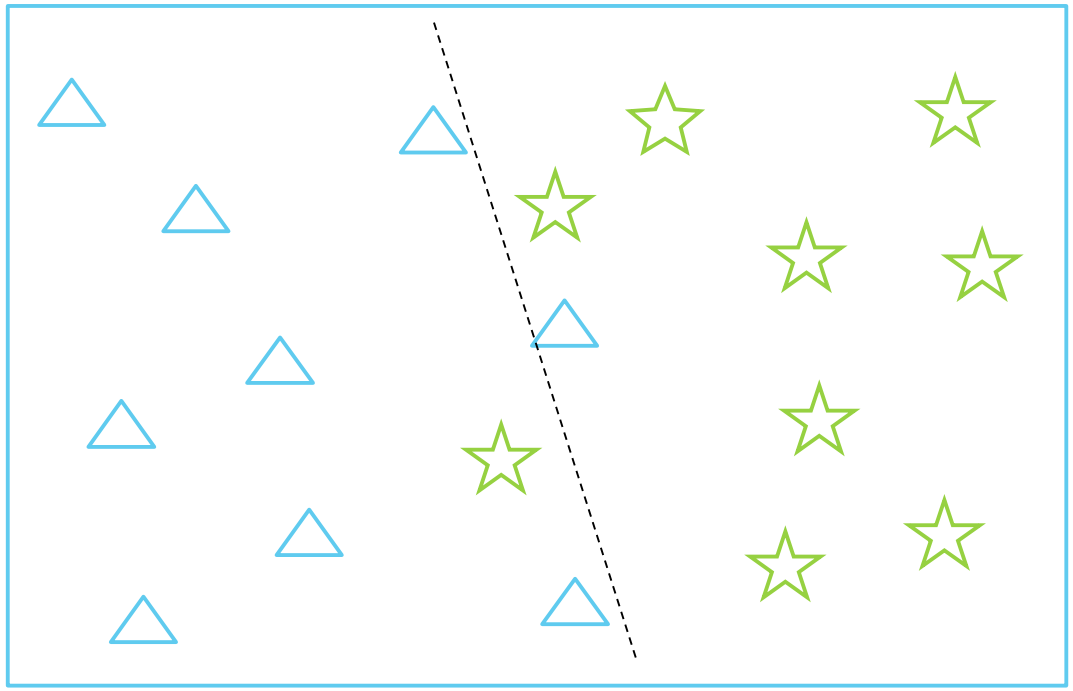
max_depth=5, MSE=1.42



max_depth=10, MSE=1.89

Недообучение

- Плохо обобщаются тренировочные данные
  - Модель недостаточно сложна
  - Недостаточно гибка для учета всех нюансов выборки

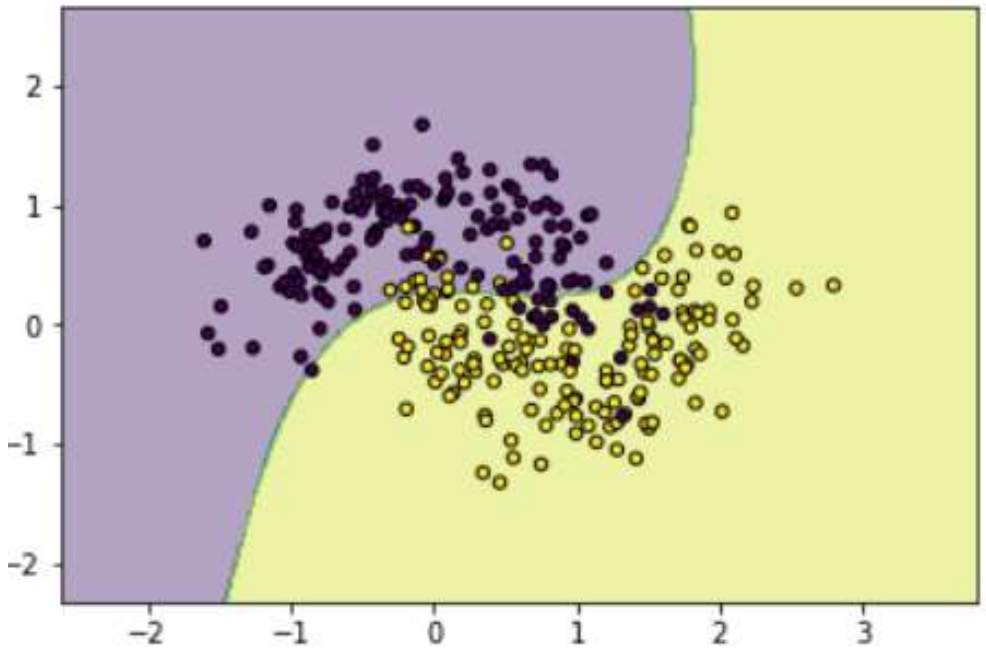


Прогнозы находятся далеко от правильных значений в целом

Является мерой систематической ошибки

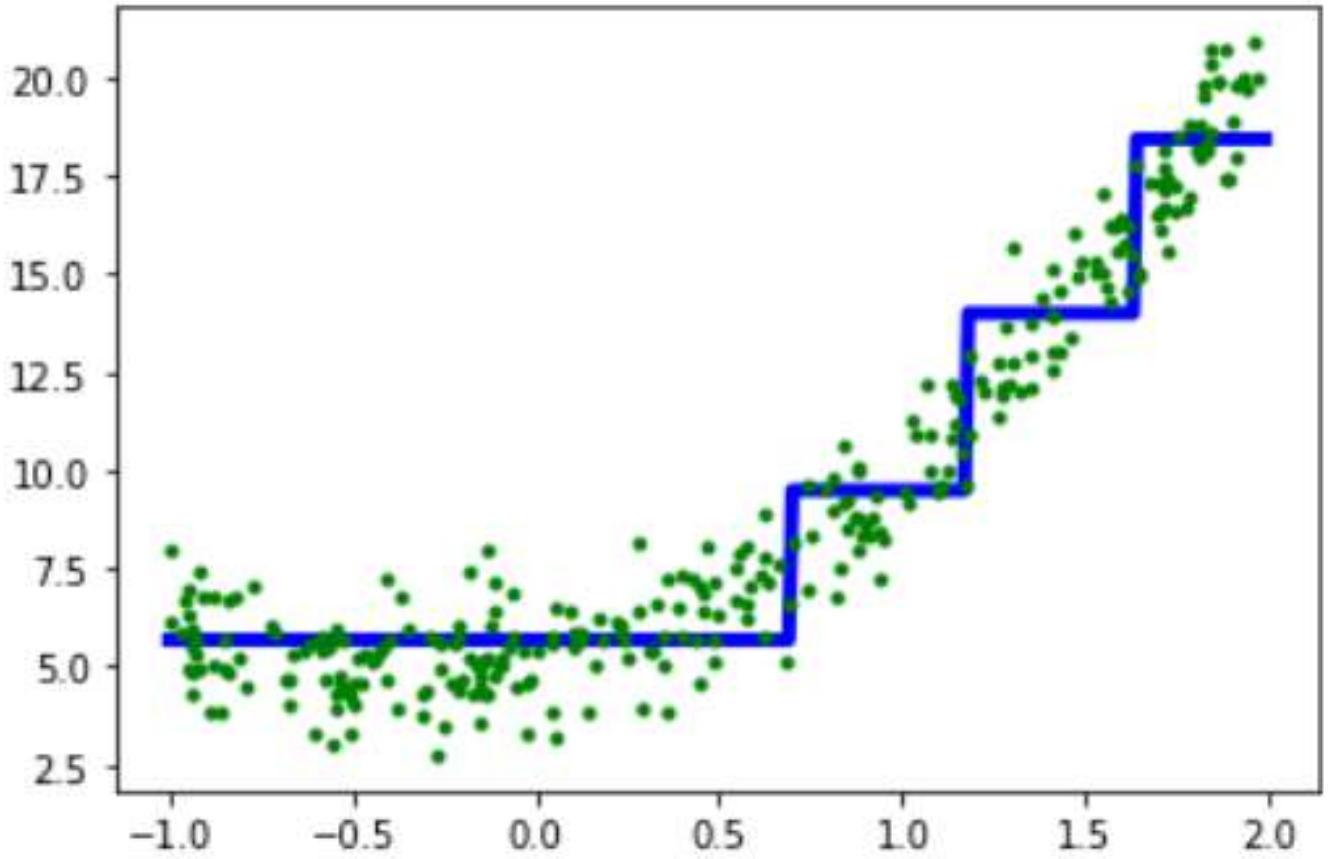
Недообученная модель (underfitted)
Имеется значительное
смещение (bias)

Классификация SVC, kernel='rbf'



SVC(kernel='rbf', gamma=0.3) = 0.87

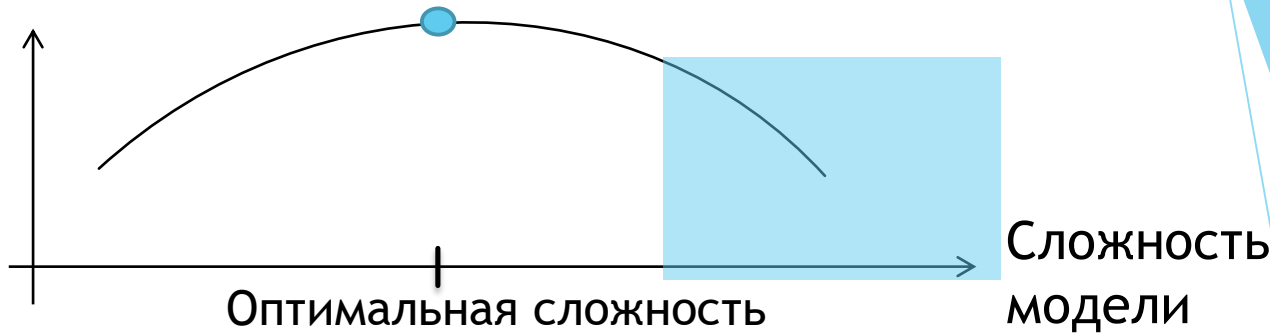
Регрессия DecisionTreeRegressor

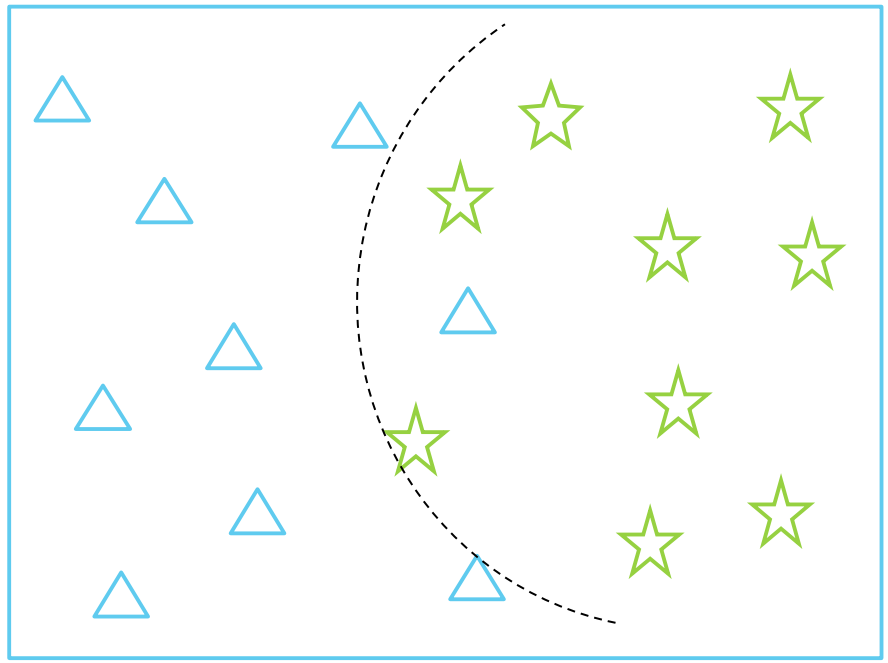


n_trees=2, MSE=2.05

Компромисс

Эффективность модели на проверочных данных

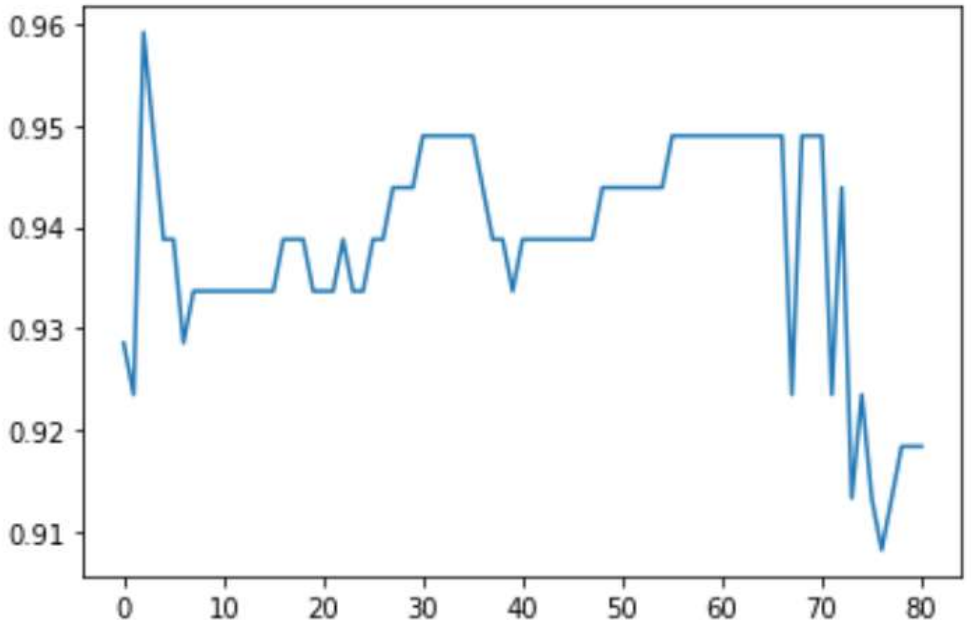




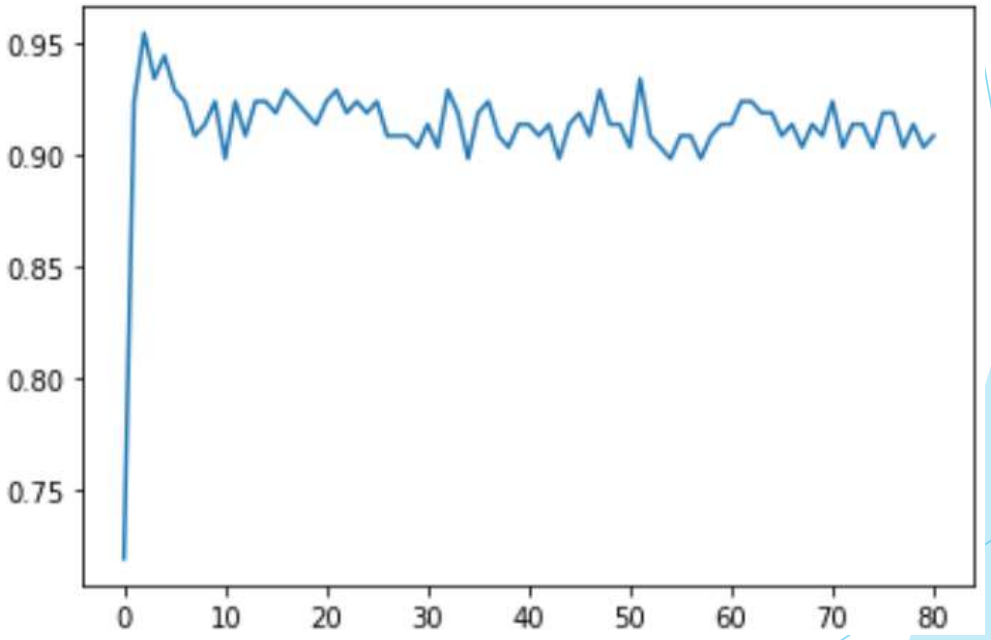
- Упростить модель

Сложность модели/точность из ДЗ (определение pH)

kNN



Tree



Пояснение на непрерывной переменной
Прогноз f(x) по x на основе моделей

$y = -0{,}000746x^6 - 0{,}00915x^5 - 0{,}0254x^4 - 0{,}0063x^3 + 0{,}02413x^2 - 0{,}019x + 3{,}756$

- [3;5] – прогнозы будут лучше у сложной модели
- [1;3],[5;7] – у простой

Регуляризация

- Добавление некоторых ограничений – штраф за сложность модели

$$S(\mathbf{w}) = \frac{1}{2}\sum_i \left(y^{(i)} - \phi(z^{(i)})\right)^2 + \lambda R(w)$$

- L1 регуляризация – приводит к обнулению весов некоторых признаков (см. геометрическую интерпретацию) (LASSO, Least Absolute Shrinkage and Selection Operator)

$$R(w) = \|w\|_1 = \sum_i |w_i|$$

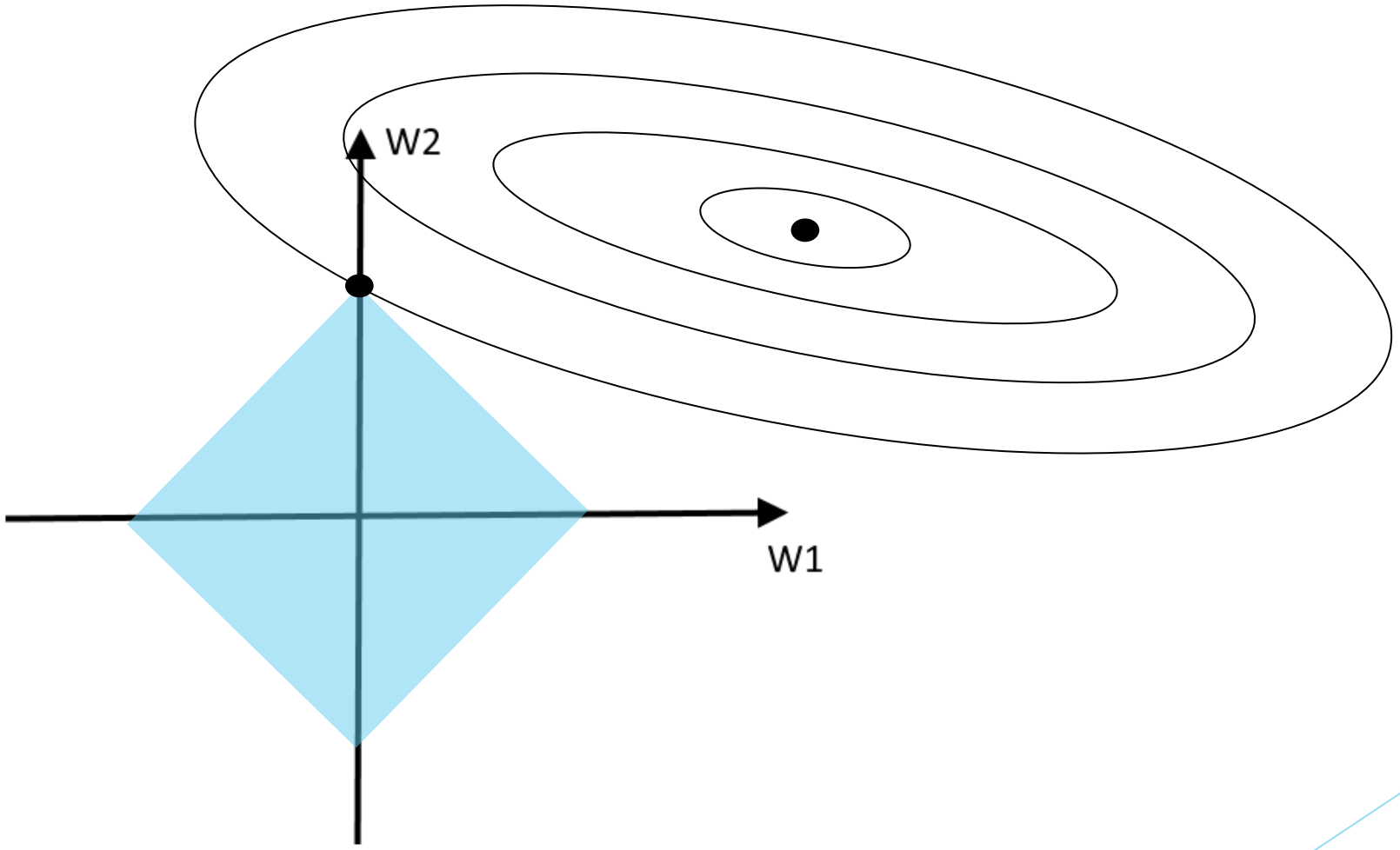
  - отбираются наиболее важные признаки, которые влияют больше всего

- L2 регуляризация – штрафы уменьшают коэффициенты, но не обнуляют их. Шумы будут влиять в меньшей степени. (Ridge – гребневая регрессия)

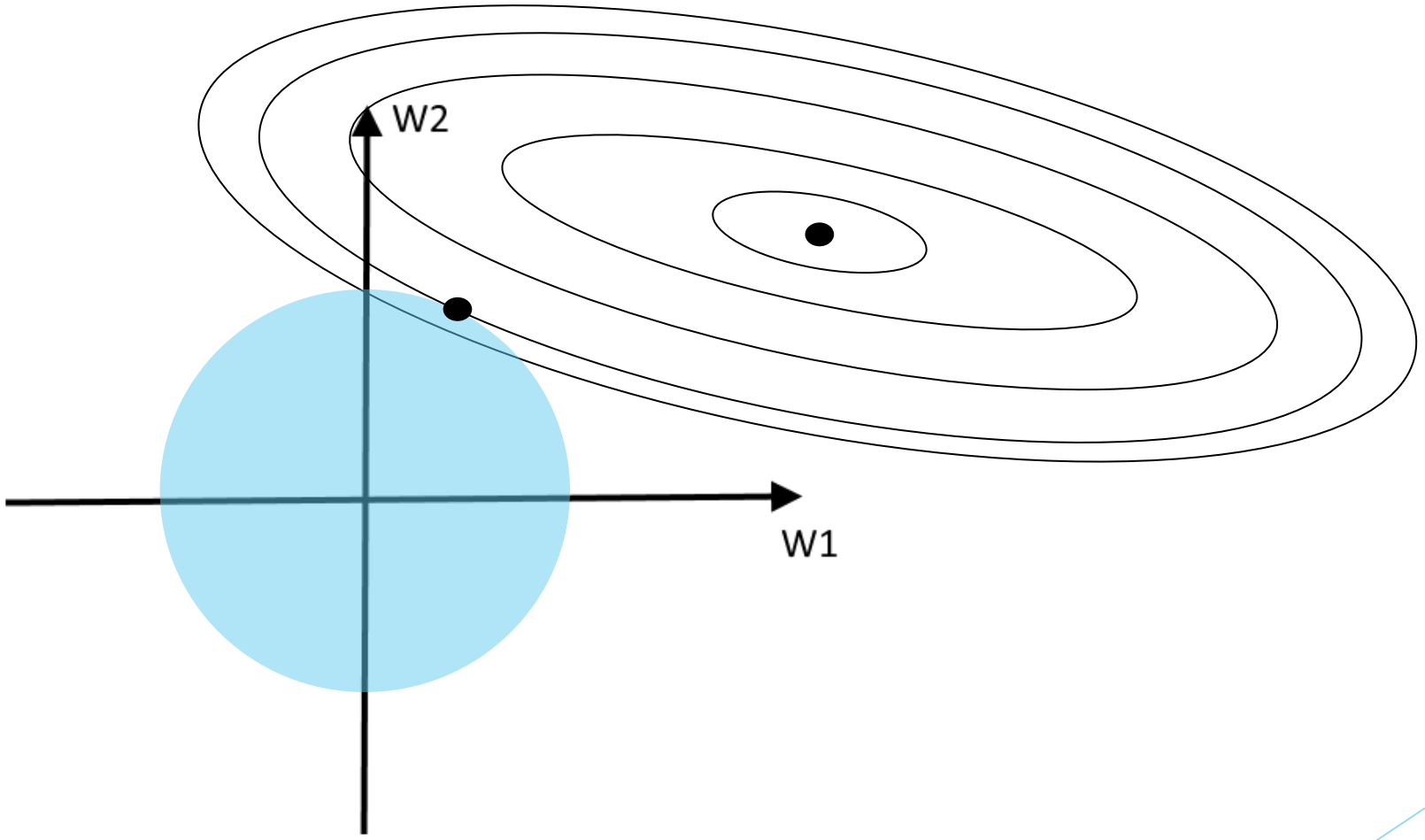
$$R(w) = \|w\|_2 = \sum_i w_i^2$$

  - Запрет на непропорционально большие веса

Графическая интерпретация L1



Графическая интерпретация L2



Параметры логистической регрессии

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
           intercept_scaling=1, max_iter=100, multi_class='multinomial',
           n_jobs=None, penalty='l2', random_state=None, solver='lbfgs',
           tol=0.0001, verbose=0, warm_start=False)
```

- C - инверсия параметра регуляризации

$$C = \frac{1}{\lambda}$$

$$S(\mathbf{w}) = \frac{1}{2} C \sum_i \left(y^{(i)} - \phi(z^{(i)}) \right)^2 + R(w)$$

- Penalty - тип регуляризации L1 или L2
- ElasticNet - регуляризация L1+L2

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

Значения весов

- Веса коэффициентов при C = 100, acc=0.947

```
1 np.set_printoptions(suppress=True)
2 clfr.coef_
```

```
array([[ 3.85125967, -0.74804975,  3.01098093,  1.53451617, -0.09567634,
         5.06697689, -6.89399177, -2.14745713,  1.17011734, -0.29895194,
        -5.03757078,  0.53469636,  4.86902231, -6.29910995,  0.4353784 ,
        -1.83450406,  1.55110284, -3.28826201,  1.40550043,  4.62693129,
        -4.93576998, -1.33331074, -1.58421024, -5.39770826,  0.09179964,
         0.86185534, -0.59651454,  0.21178445, -1.65902189, -3.19372922]])
```

- C=0.01, acc=0.953

```
1 clfr2.coef_
```

```
array([[-0.12646764, -0.13553344, -0.12498955, -0.11628545, -0.05999642,
        -0.04771833, -0.11080519, -0.12247111, -0.02181388,  0.04881202,
        -0.09997278, -0.00126052, -0.08049474, -0.08140629,  0.00775021,
         0.02889708, -0.0196549 , -0.04340753,  0.0352802 ,  0.05042594,
        -0.14112672, -0.15057784, -0.13418831, -0.12366993, -0.10006021,
        -0.06475056, -0.10754904, -0.13399401, -0.09388958, -0.03824308]])
```

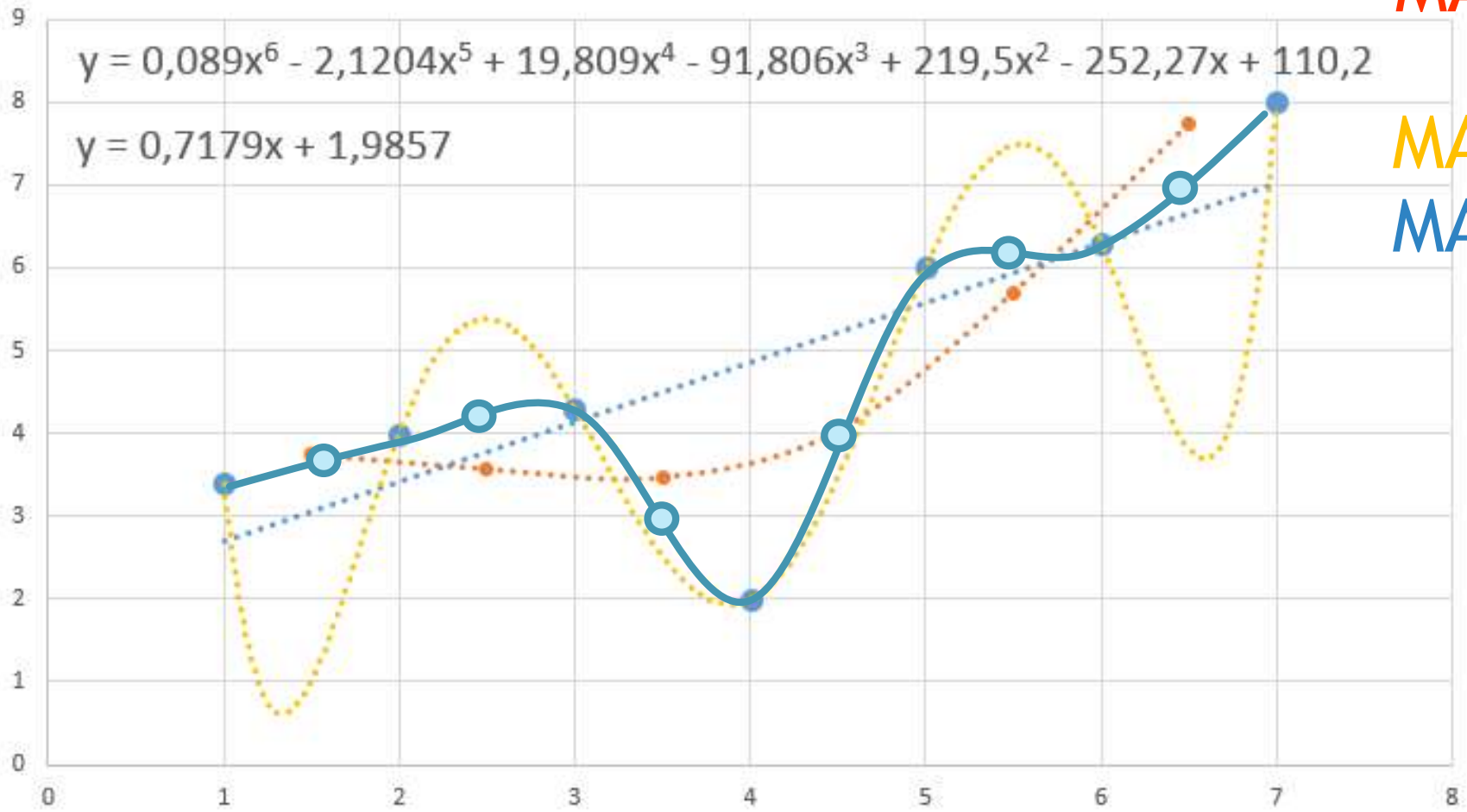
- Снижая параметр C увеличиваем силу регуляризации

lection_5_regularization_logistic_regression

Пояснение на непрерывной переменной
Прогноз f(x) по x на основе моделей

$y = -0{,}000746x^6 + 0{,}00915x^5 - 0{,}0254x^4 - 0{,}0063x^3 + 0{,}02413x^2 + 0{,}019x + 3{,}756$

MAE=0,69



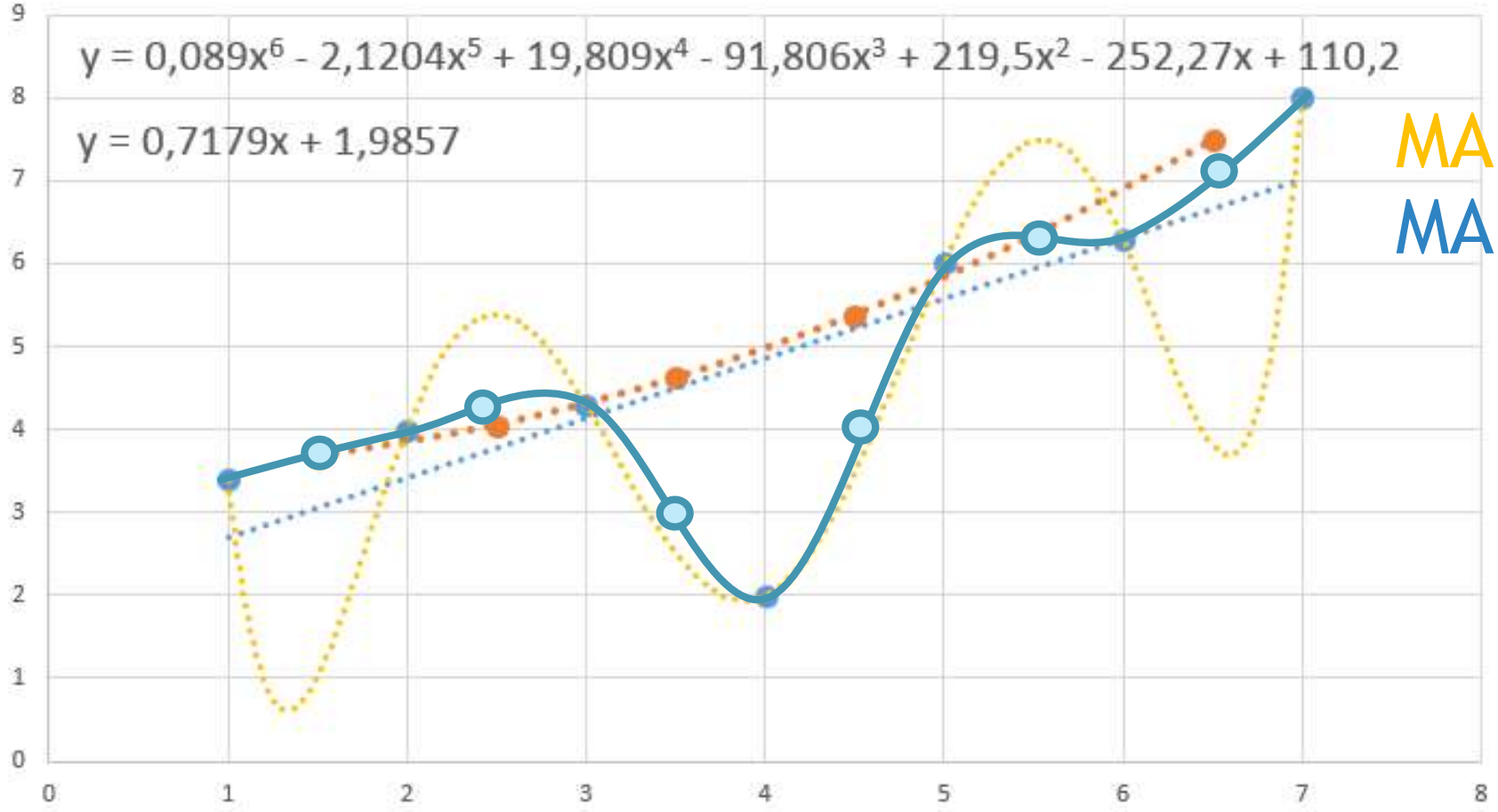
MAE=2,07

MAE=0,94

Пояснение на непрерывной переменной
Прогноз f(x) по x на основе моделей

$y = 0{,}0967x^2 - 0{,}0099x + 3{,}469$

MAE=0,75



MAE=2,07

MAE=0,94