In [1]:
```
import IPython.display
IPython.display.display_latex(IPython.display.Latex(filename="../macros.tex"))
```

Недообучение переобучение Underfitting and Overfitting

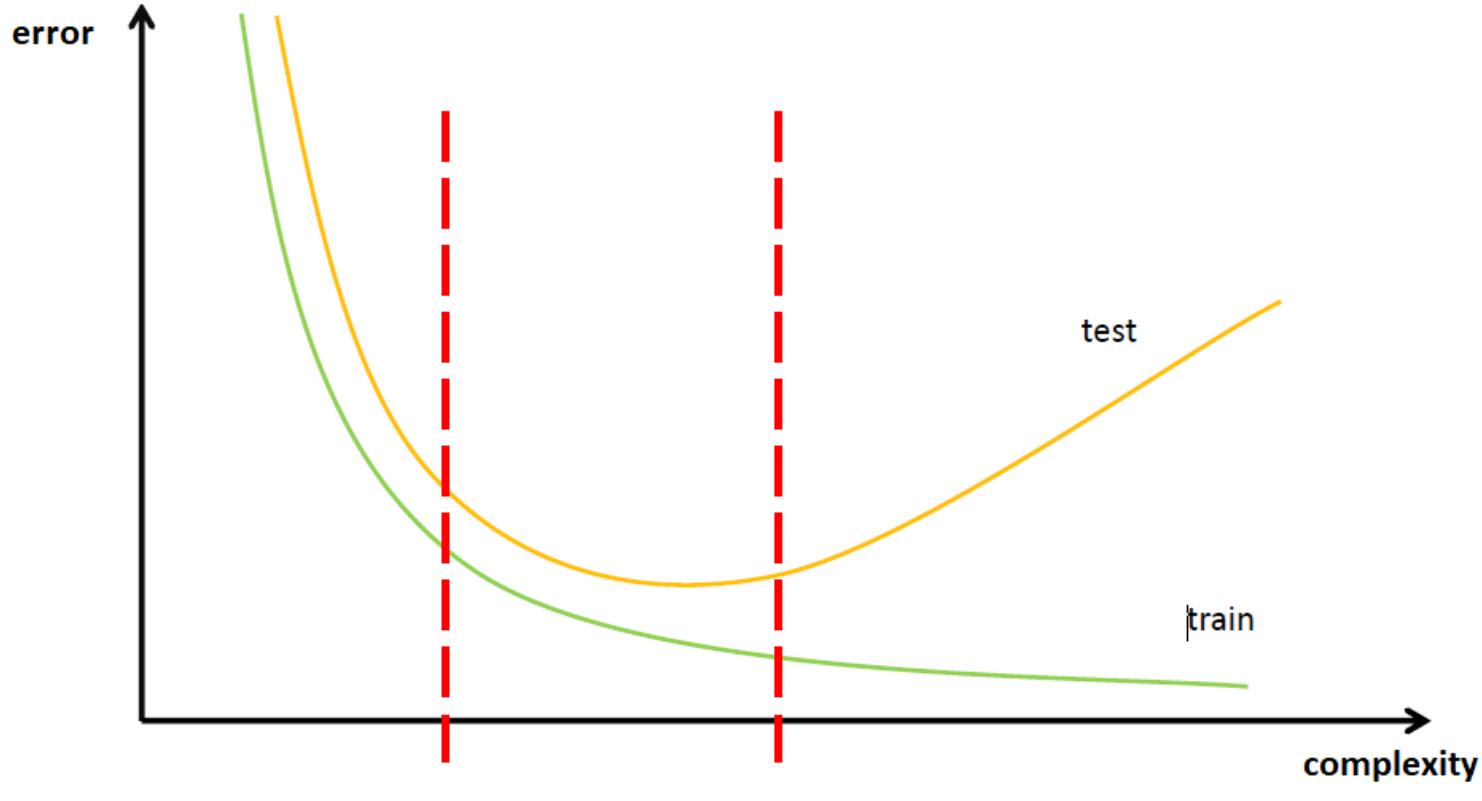
Переобучение ситуация когда вероятность ошибки алгортима на тестовой выборке существенно выше чем средняя ошибка на тренировочной выборке. Вы можете переобучиться когда используете слишком сложные модели.

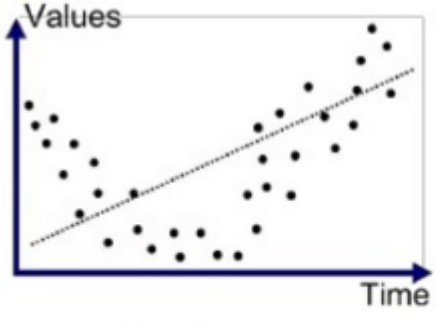
Недообучение ситуация когда ошибка на обучающей выборке очень большая. Не получаеся настроить алгоритм. Это может случится если использовать слишком простую модель.

$$f(x) = w_3 x^3 + w_2 x^2 + w_1 x + w_0 + \epsilon$$

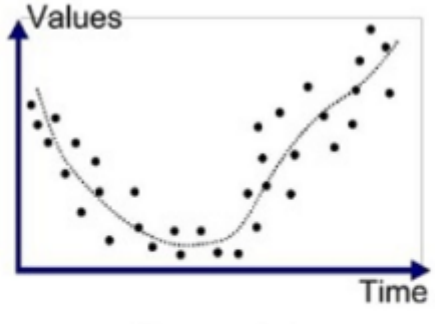
$$\alpha_1(x) = \theta_1 x + \theta_0$$

$$\alpha_1(x) = \sum_{i=1}^{8} \theta_i x^i$$

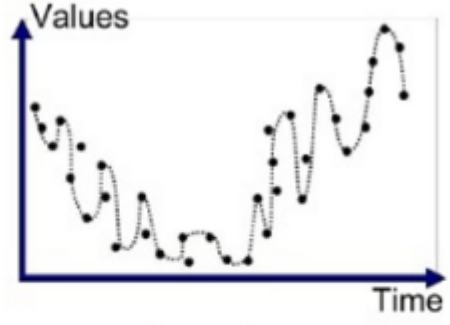

error
test
train
complexity



Underfitted



Good Fit/Robust



Overfitted

Сложность:

- Количество параметров
- Разброс возможных значений параметров

Пример:

- глубина дерева
- количество итераций в бустинге
- ...

Если у модели много параметров, она может "запомнить" примеры из тренировочной выборки.

Переобучение по гиперпараметрам:

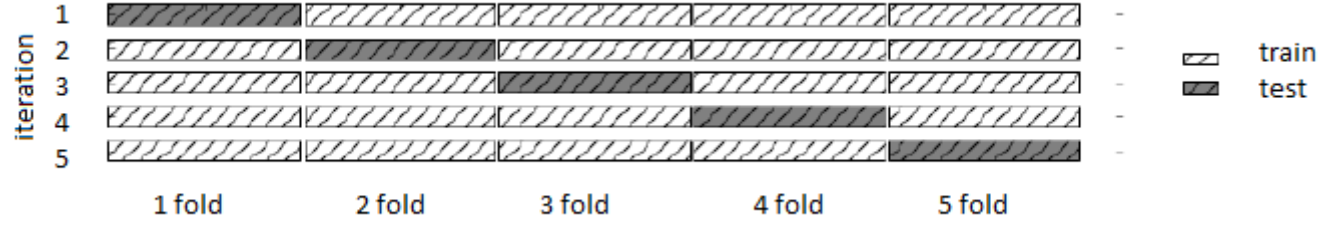
мы выбираем "лучшие" гиперпараметры по тестовой выборке, но на реальных объектах получаем плохое предсказание.

Регуляризация - процесс ввода дополнительной информации для предотвращения переобучения.

Кросс валидация, перекрестная проверка - оценка обобщающей способности + проверка на переобучение.

При перекрестной проверке мы несколько раз разбиваем образец на тренировочную и тестовую часть и обучаем несколько моделей.

k-fold cross-valdation



Вместо 1 оценки получаем k. Так что это лучшая оценка обобщающей способности, и мы проверяем переобучение.

Хороший подход определить гиперпараметры: *grid search + kfold cv*

отложенная выборка, финальный тест

Мы выбираем финальную тестовую выборку (случайно) перед началом моделирования и не используем его до конца моделирования. Мы используем его только один раз для проверки переобучения.

Если мы имеем большую разницу между оценочной ошибкой и ошибкой в конечном тесте, это означает, что мы ошибаемся в логике моделирования.