Model complexity.

Victor Kitov
v.v.kitov@yandex.ru

Yandex School of Data Analysis

Hyperparameters selection

- Using CV we can select hyperparameters of the model[1]
- Each model has hyperparameter, corresponding to model complexity.
- Model complexity - ability to reproduce training set.
- Examples:
  - regression: # of features d, e.g. $x, x^2, ... x^d$
  - K-NN: number of neighbors K

[1] can we use CV loss in this case as estimation for future losses?

Underfitted and overfitted models[2]

Too simple (underfitted) model

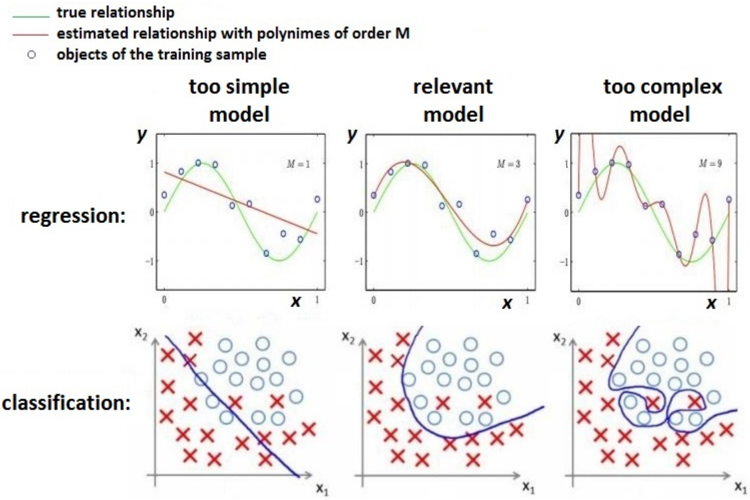
Model that oversimplifies true relationship $\mathcal{X} \to \mathcal{Y}$.

Too complex (overfitted) model

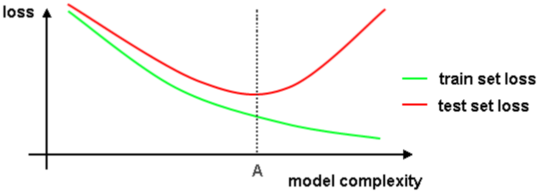
Model that is too tuned on particular peculiarities (noise) of the training set instead of the true relationship $\mathcal{X} \to \mathcal{Y}$.

[2] In fact most models overfit, meaning that empirical risk<expected risk. Underfitted models just have lower difference than overfitted ones.

Examples of overfitted / underfitted models



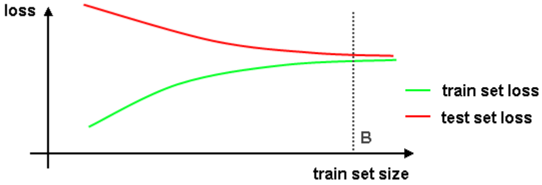
Loss vs. model complexity



Comments:

- expected loss on test set is always higher than on train set.
- left to A: model too simple, underfitting, high bias
- right to A: model too complex, overfitting, high variance

Loss vs. train set size



Comments:

- expected loss on test set is always higher than on train set.
- right to B there is no need to further increase training set size
  - useful to limit training set size when model fitting is time consuming