Model selection and regularization

Generalization error: refers to how well the method generalizes, what is the error on unseen data.

Types of errors:

Training error: refers to the error computed during the training phase. It *measures how well the model performs on the training data* and indicates how effectively the model has learned the patterns and relationships within the training set. It's used to tune the parameters.

Validation error: computed on a separate validation set that is not used during the training set, *it serves as an estimate of how well the model generalizes to unseen data. The validation error helps to tune the hyperparameter.*

Test error: calculated on an independent test set that is separate from both training and validation data. I*t provides an estimate of the model's performance on unseen data.*

Feature selection

Reducing the number of features by selecting a smaller subset may have two positive effects: increase interpretability and robustness.

using the best subset selection algorithm:

train all 2^p possible models and select the best one

What to do if the results are bad

Try to fix the problem of high variance by:

Getting more training examples

Trying smaller sets of features

▼ Use regularization

A strategy used to prevent overfitting by decreasing the degrees of freedom

Increasing gama in regularization

Apply early stopping

Try to fix the problem of high bias by:

Adding features

Adding transformed features, e.g., power of features as in poly regression

Decreasing gama in regularization