APPLIED MACHINE LEARNING

BY
MESFIN DIRO

July 21, 2021

Bias Error

- Bias are the simplifying assumptions made by a model to make the target function easier to learn.
- In supervised machine learning an algorithm learns a model from training data.
- Generally, linear algorithms have a high bias making them fast to learn and easier to understand but generally less flexible.
- **Low Bias**: Less assumptions about the form of the target function.
- **High-Bias**: More assumptions about the form of the target function.

Variance Error

- Variance is the amount that the estimate of the target function will change if different training data was used.
- The target function is estimated from the training data by a machine learning algorithm, so we should expect the algorithm to have some variance.
- Machine learning algorithms that have a high variance are strongly influenced by the specifics of the training data.
- **Low Variance**: Small changes to the estimate of the target function with changes to the training dataset: Linear Regression and Logistic Regression.
- **High Variance**: Large changes to the estimate of the target function with changes to the training dataset.
- Nonlinear machine learning algorithms that have a lot of flexibility have a high variance: decision trees, Support Vector Machine, Random forest.

Bias-Variance Trade-Off

- Supervised machine learning algorithms can best be understood through the lens of the bias-variance trade-off.
- In supervised machine learning an algorithm learns a model from training data.
- The prediction error for any machine learning algorithm can be broken down into three parts:
  - Bias Error
  - Variance Error
  - Irreducible Error.
- The irreducible error cannot be reduced regardless of what algorithm is used.

Bias-Variance Trade-Off

- The goal of any supervised machine learning algorithm is to achieve low bias and low variance. In turn the algorithm should achieve good prediction performance.
  - **Linear** machine learning algorithms often have a high bias but a low variance.
  - **Nonlinear** machine learning algorithms often have a low bias but a high variance.
- The parameterization of machine learning algorithms is often a battle to balance out bias and variance.
- There is no escaping the relationship between bias and variance in machine learning.
  - Increasing the bias will decrease the variance.
  - Increasing the variance will decrease the bias.

Overfitting and Underfitting

- The cause of poor performance in machine learning is either overfitting or underfitting the data.
- **Overfitting** refers to a model that models the training data too well.
- Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function.
- Underfitting refers to a model that can neither model the training data nor generalize to new data.
- There are two important techniques that you can use when evaluating machine learning algorithms to limit overfitting

  1. Use a resampling technique to estimate model accuracy.
  2. Hold back a validation dataset.

Model Validation

- Parameters which define the model architecture are referred to as hyper-parameters:
  - degree of polynomial features in regression model
  - maximum depth of decision tree
  - number of trees in random forest
  - number of neurons in neural network etc
- Model parameters are learned during training when we optimize a loss function using something like gradient descent.
- However, hyper-parameters are not model parameters and they cannot be directly trained from the data.
- Thus, the process of searching for the ideal model architecture is referred to as **hyper-parameter tuning**.

Model Validation

- The basic recipe for applying a supervised machine learning model are:
  - Choose a class of model
  - Choose model hyper-parameters
  - fit the model to the training data
  - use the model to predict labels for new unseen data
- The choice of model and choice of hyper-parameters - are perhaps the most important part of using these tools and techniques effectively

Model Validation

- Validation is a process of deciding whether the numerical results quantifying hypothesized relationships between variables, are acceptable as descriptions of the data, is known as **validation**.
- Generally, an error estimation for the model is made after training, better known as evaluation of residuals.
- In this process, a numerical estimate of the difference in predicted and original responses is done, also called the training error.
- However, this only gives us an idea about how well our model does on data used to train it.
- To avoid overfitting, it is common practice when performing a supervised machine learning experiment to hold out part of the available data as a **test set.**
- Getting evaluation technique for our model to generalize to an independent/ unseen data set is known as Cross Validation.

Cross Validation

- **Cross-validation(CV)** is a technique for evaluating a machine learning model and testing its performance.
- **CV** is commonly used in applied **ML** tasks to evaluate machine learning models on a limited data sample.
- It helps to compare and select an appropriate model for the specific predictive modeling problem.
- There are a lot of different techniques that may be used to **cross-validate** a model:
  - Hold-out
  - K-folds
  - Leave-one-out etc

Hold-out Method

- This is the simplest evaluation method and is widely used in Machine Learning projects.
- Here the entire dataset(population) is divided into 2 sets – train set and test set.
- The data can be divided into 70-30 or 60-40, 75-25 or **80-20**, or even 50-50 depending on the use case.
- As a rule, the proportion of training data has to be larger than the test data.

K-Fold Cross Validation

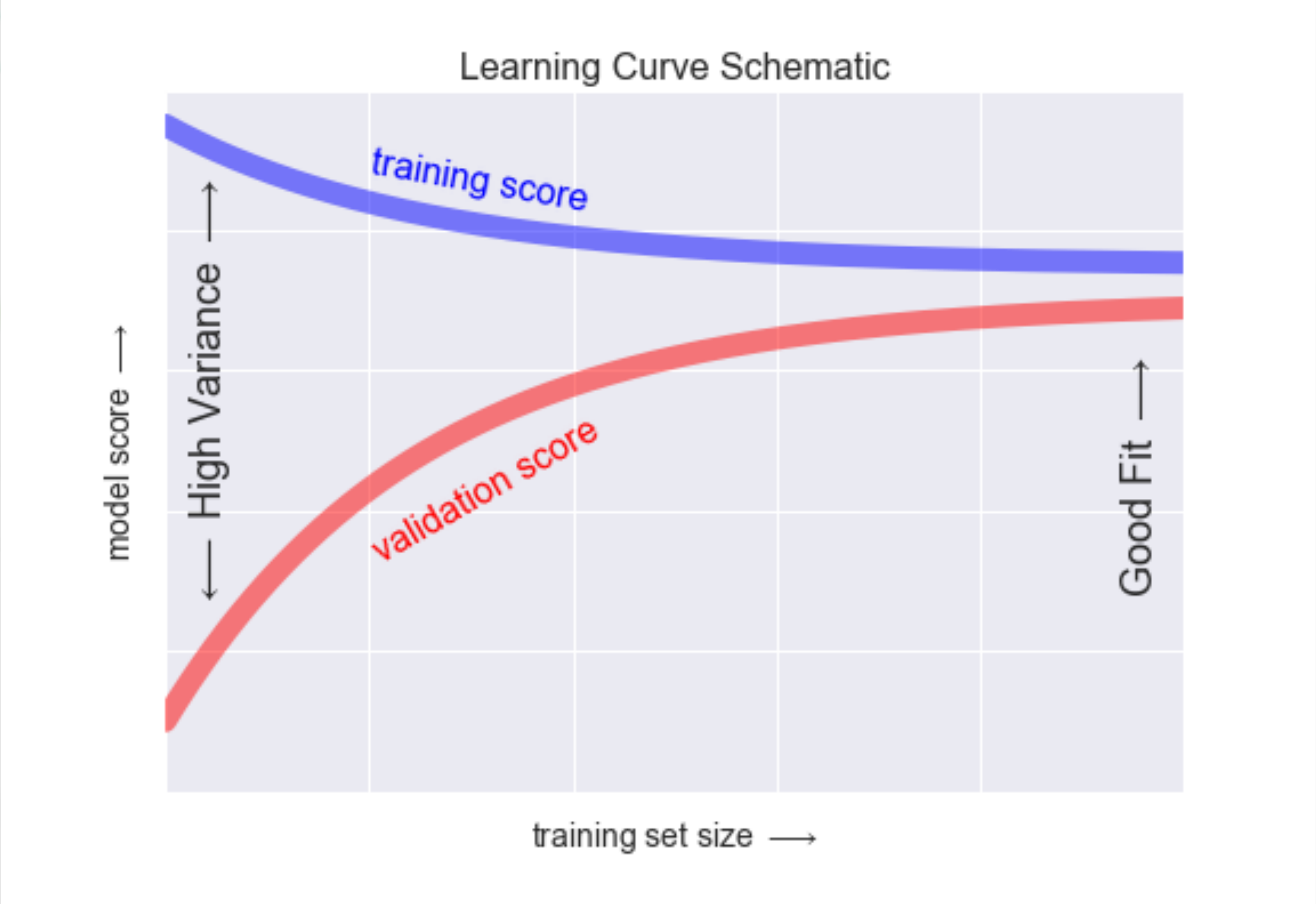
- The dataset should be as large as possible to train the model and removing considerable part of it for validation poses a problem of losing valuable portion of data that we would prefer to be able to train
- The k-fold cross validation is a procedure used to estimate the skill of the model on new data.

Learning Curves

- A learning curve is a plot of model learning performance over experience or time..
- Learning curves are a widely used diagnostic tool in machine learning for algorithms that learn from a training dataset incrementally.
- Generally, a learning curve is a plot that shows time or experience on the x-axis and learning or improvement on the y-axis.
- **Learning Curve**: Line plot of learning (y-axis) over experience (x-axis)
  - **Train Learning Curve**: Learning curve calculated from the training dataset that gives an idea of how well the model is learning.
  - **Validation Learning Curve**: Learning curve calculated from a hold-out validation dataset that gives an idea of how well the model is generalizing.
- There are three common dynamics that you are likely to observe in learning curves; they are:
  - Underfit: cannot learn the training dataset
  - Overfit: random fluctuations in the training dataset
  - Good Fit: a minimal gap between the two final loss values

Learning Curves

- One important aspect of model complexity is that the optimal model will generally depend on the size of the training data.
- A plot of the training/validation score with respect to the size of the training set is known as a learning curve
- A model will never except by chance give a better score to the validation set that the training set.



Learning Curves

- A learning curve is a plot of model learning performance over experience or time..
- Learning curves are a widely used diagnostic tool in machine learning for algorithms that learn from a training dataset incrementally.
- Generally, a learning curve is a plot that shows time or experience on the x-axis and learning or improvement on the y-axis.
- **Learning Curve**: Line plot of learning (y-axis) over experience (x-axis)
  - **Train Learning Curve**: Learning curve calculated from the training dataset that gives an idea of how well the model is learning.
  - **Validation Learning Curve**: Learning curve calculated from a hold-out validation dataset that gives an idea of how well the model is generalizing.
- There are three common dynamics that you are likely to observe in learning curves; they are:
  - Underfit: cannot learn the training dataset
  - Overfit: random fluctuations in the training dataset
  - Good Fit: a minimal gap between the two final loss values