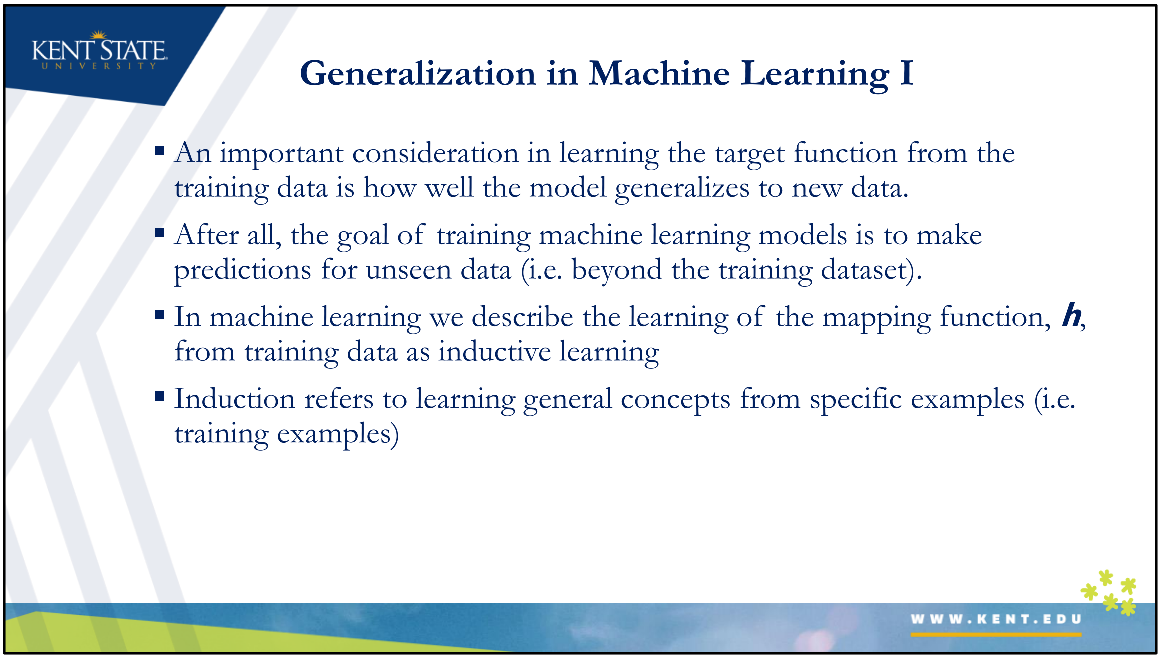
KENT STATE UNIVERSITY

Underfitting and Overfitting

WWW.KENT.EDU

One of the primary uses of ML models is for prediction. The ability of the model to predict the output of cases it has not yet seen. As we saw earlier, ML models are trained with existing or available data. The question is, “how do we ensure that such models are also able to predict well?” Underfitting and Overfitting are two aspects of model fitting that allows us to fine tune the model for better prediction.

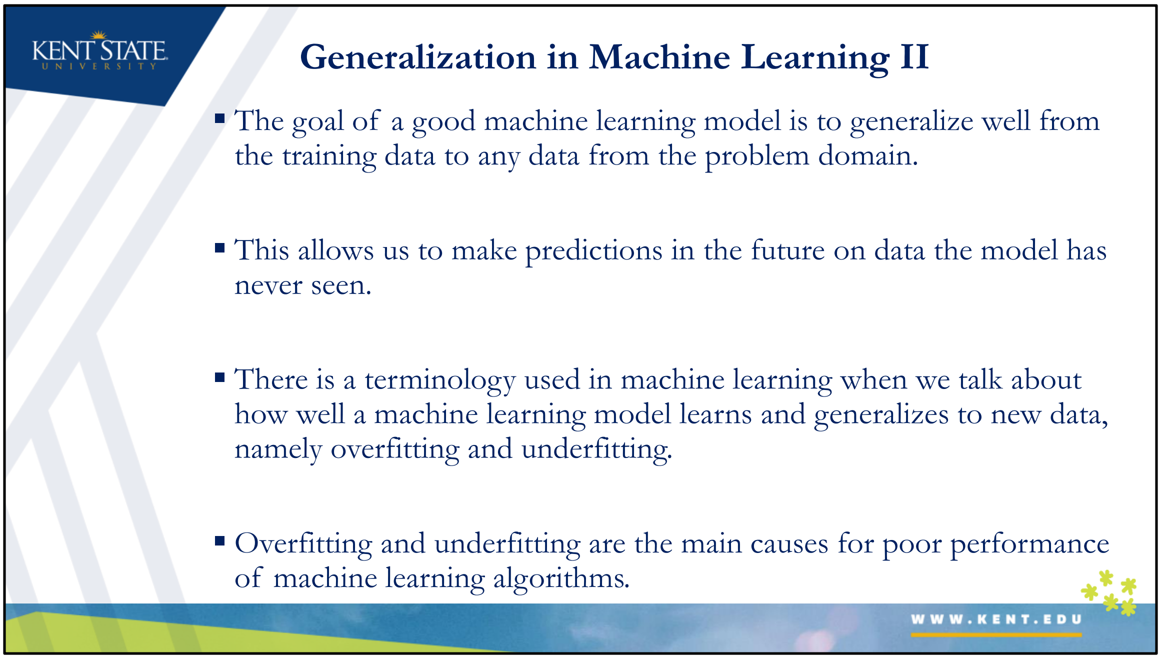


Generalization in Machine Learning I

- An important consideration in learning the target function from the training data is how well the model generalizes to new data.
- After all, the goal of training machine learning models is to make predictions for unseen data (i.e. beyond the training dataset).
- In machine learning we describe the learning of the mapping function, ***h***, from training data as inductive learning
- Induction refers to learning general concepts from specific examples (i.e. training examples)

WWW.KENT.EDU

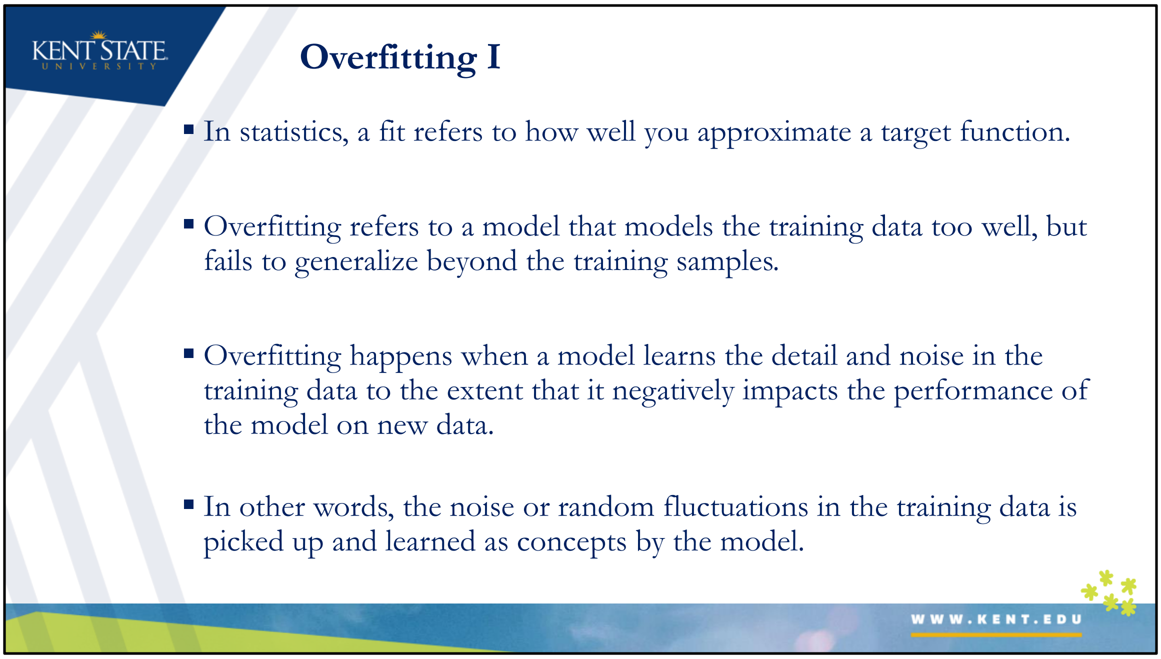
Given sample data with pairs of (X, Y), the primary objective during training a ML model is to find the mapping function h that maps X to Y. The optimization algorithm, to reduce the cost function (objective), will necessarily try to optimize the function h so that the errors are minimal. While this determines a good mapping function h for the available data, this may not serve well to predict future or new values. Why?



Generalization in Machine Learning II

- The goal of a good machine learning model is to generalize well from the training data to any data from the problem domain.
- This allows us to make predictions in the future on data the model has never seen.
- There is a terminology used in machine learning when we talk about how well a machine learning model learns and generalizes to new data, namely overfitting and underfitting.
- Overfitting and underfitting are the main causes for poor performance of machine learning algorithms.

Underfitting and Overfitting are symptoms of models that are unable to either learn the existing data well, or predict new or future data well. As you will see, we can use these approaches to refine our model to improve both training results, existing data, and prediction results, future data.

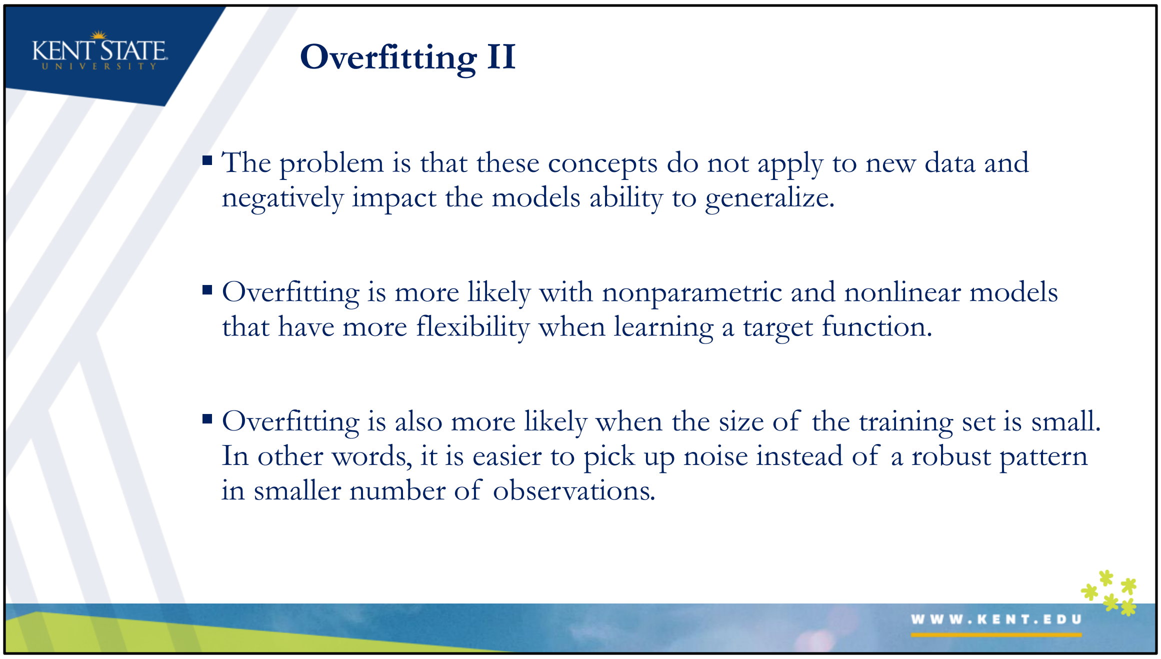


Overfitting I

- In statistics, a fit refers to how well you approximate a target function.
- Overfitting refers to a model that models the training data too well, but fails to generalize beyond the training samples.
- Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data.
- In other words, the noise or random fluctuations in the training data is picked up and learned as concepts by the model.

WWW.KENT.EDU

Recall our question on why optimizing errors in training will result in a good mapping function h for the training data, but may result in a poor predictor. There are many reasons for this, but one primary cause is *overfitting*, i.e., for the model to not only fit the signal (data), but also any noise. The larger the model, the more likely it will overfit.

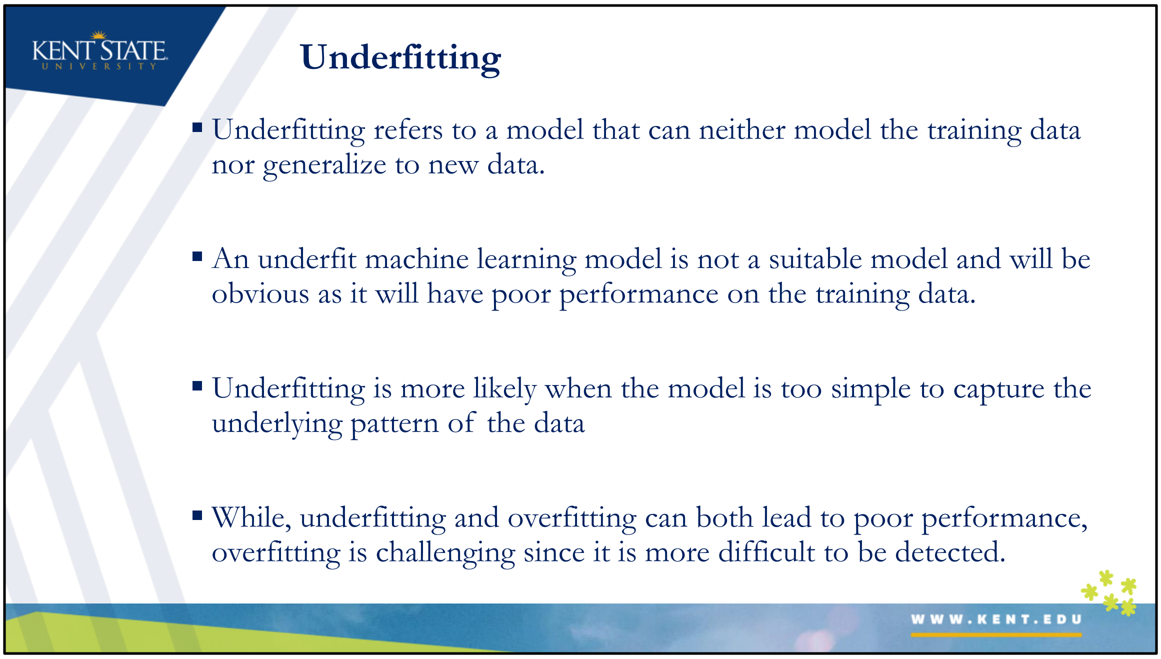


Overfitting II

- The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.
- Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function.
- Overfitting is also more likely when the size of the training set is small. In other words, it is easier to pick up noise instead of a robust pattern in smaller number of observations.

Thus, when a model learns not only the relationship between X and Y, but also any noise that might be in the data, it does not generalize well. From a statistical perspective, this can be viewed as a “sampling” problem. If the sample was large and representative enough, then of course, we would want the model to train for the data well. But, this is rarely the case. Nevertheless, the larger the sample, less the chance of overfitting.

To summarize, overfitting is more likely when the model is large, or has a large number of parameters, which is typical in nonlinear and nonparametric models. The effect of overfitting is reduced when the sample size is large and representative of the population.



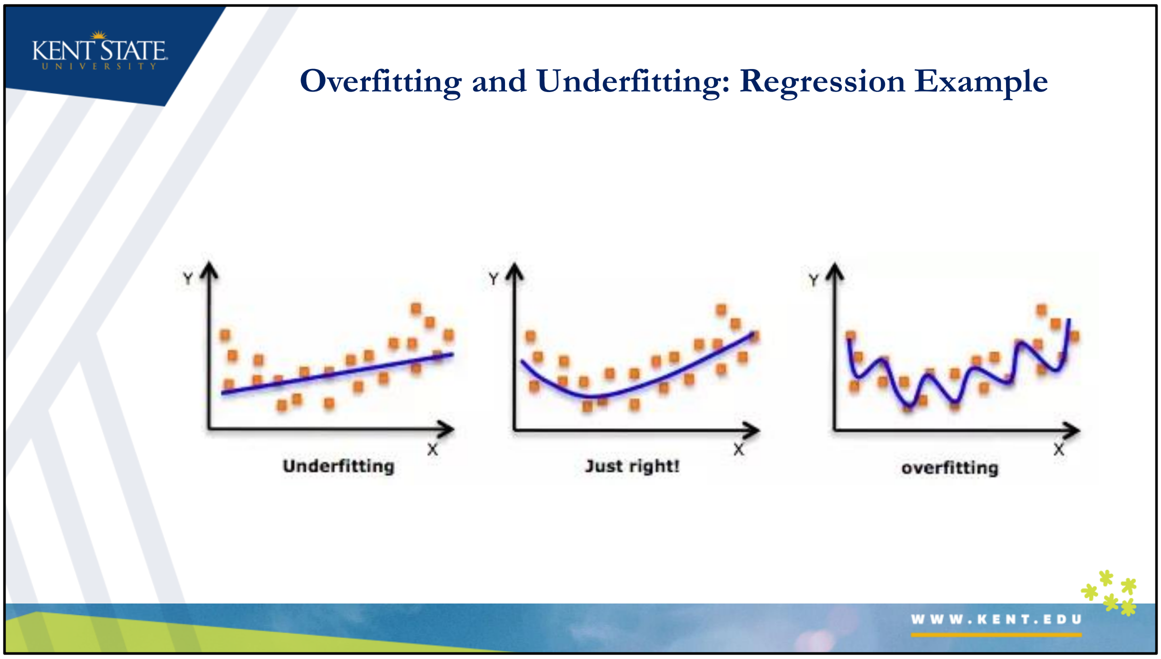
KENT STATE UNIVERSITY

Underfitting

- Underfitting refers to a model that can neither model the training data nor generalize to new data.
- An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.
- Underfitting is more likely when the model is too simple to capture the underlying pattern of the data
- While, underfitting and overfitting can both lead to poor performance, overfitting is challenging since it is more difficult to be detected.

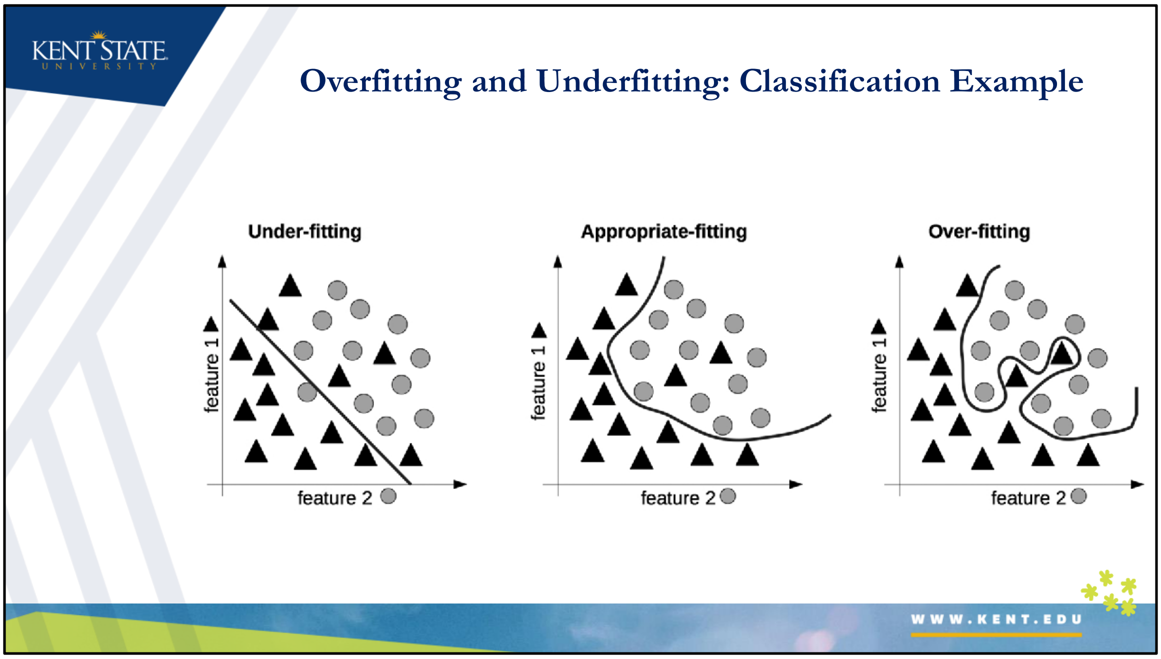
WWW.KENT.EDU

Underfitting is when a model typically does not fit the available data well. This happens when a model is not suitable enough to capture the relationship between X and Y. In statistical terms, an underfitted model leads to bias. That is, on average, the relationship captured by the model doesn't represent the true relationship between X and Y. It is relatively easy to correct for underfitting. Using a larger, size or complexity, model will usually lead to better training results. But, note that nothing can correct for this if the sample size is not large enough. Fortunately, in most practical applications of ML, getting adequate data is not an issue.



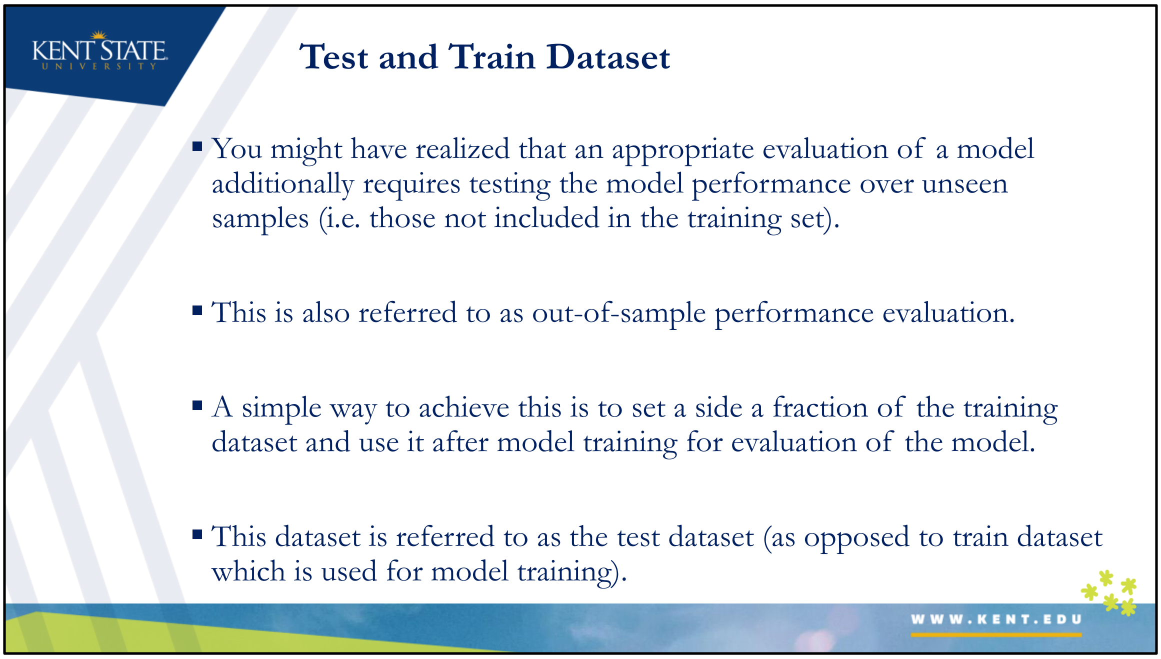
Overfitting and Underfitting: Regression Example

These examples illustrate underfitting and overfitting. In underfitting, the true relationship between X and Y is a non-linear function, but only a linear model was used. This does not truly capture the relationship. On the other hand, in overfitting, the model is complex enough that it captures natural variation among the points, rather than the relationship. This is akin to capturing the value for each individual person’s age in predicting height, when what we are interested in is predicting the average height for a person of a certain age. Obviously, people of a certain age may have different heights. We are not interested in predicting each person’s height as much as generalizing the results to a population, where we could choose a person at random and predict their average height.



Overfitting and Underfitting: Classification Example

These examples illustrate the concepts, but now for classification. Again note that a nonlinear classifier is need, but the level of nonlinearity determines whether the classifier is appropriate, or overfitted. Overfitting is usually the harder of the two to resolve. We will look at how we can identify overfitting, and the steps we can take to determine an appropriate model.



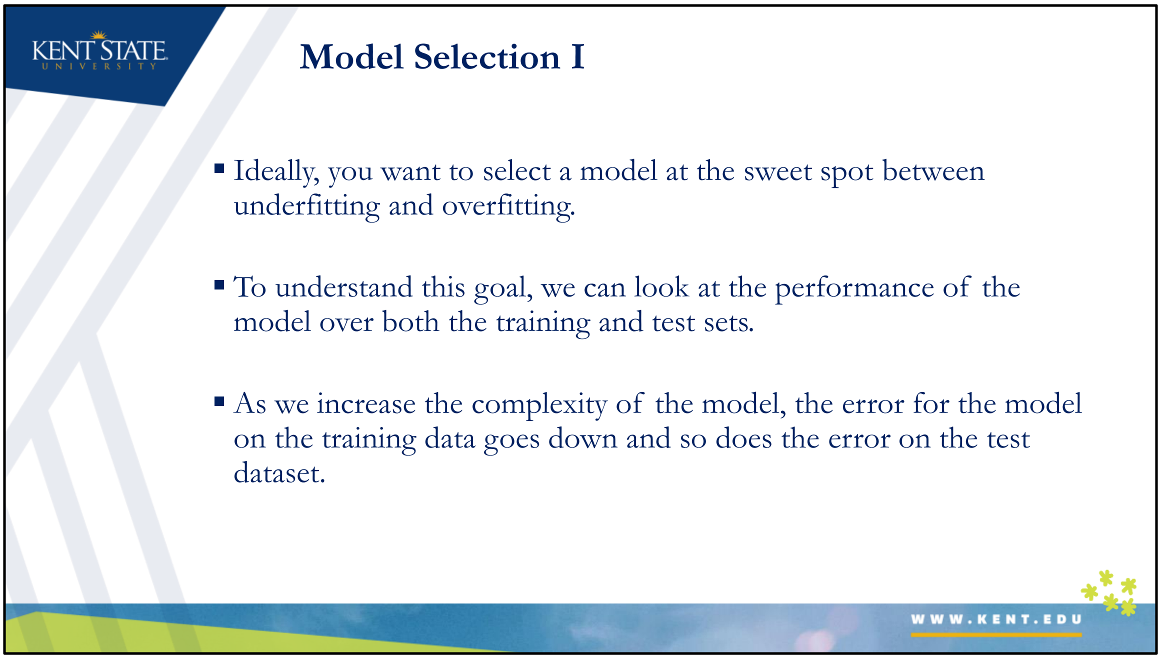
KENT STATE UNIVERSITY

Test and Train Dataset

- You might have realized that an appropriate evaluation of a model additionally requires testing the model performance over unseen samples (i.e. those not included in the training set).
- This is also referred to as out-of-sample performance evaluation.
- A simple way to achieve this is to set a side a fraction of the training dataset and use it after model training for evaluation of the model.
- This dataset is referred to as the test dataset (as opposed to train dataset which is used for model training).

WWW.KENT.EDU

One approach to identify models that predict well is to determine the model based on its performance on an unseen (not trained) sample. An easy way to accomplish that is to divide the current sample into multiple parts; a training dataset, and a test dataset. We use the training dataset to build the model, and the test dataset to evaluate the models on unseen samples. There are several ways of dividing datasets into training and test, this is the simplest one.

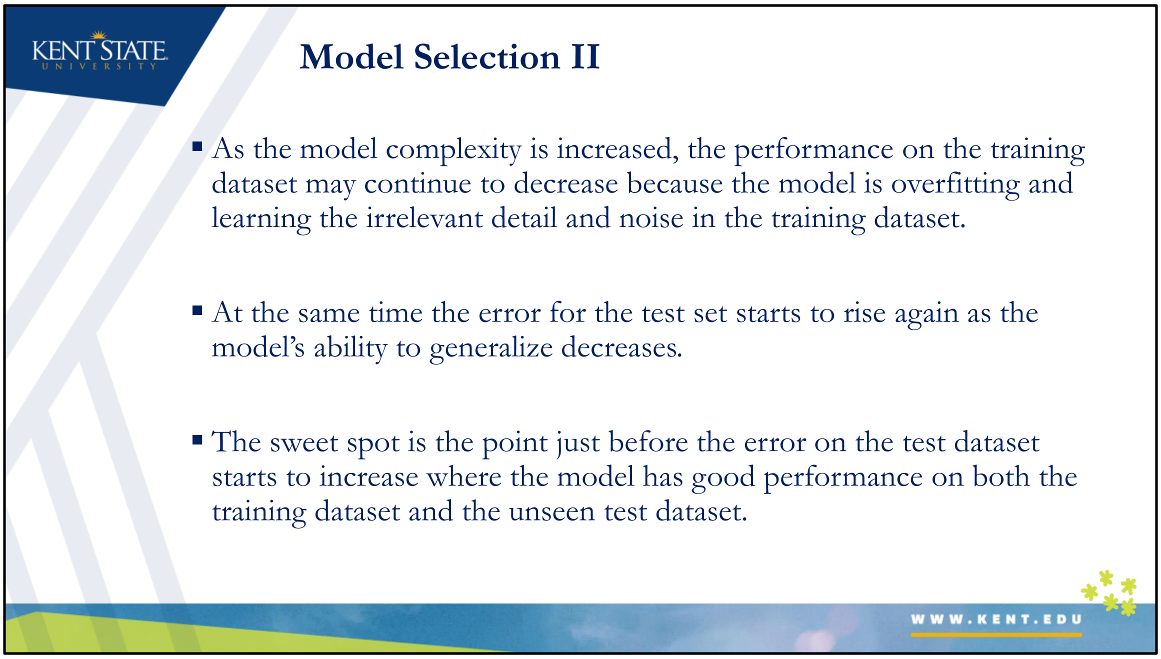


Model Selection I

- Ideally, you want to select a model at the sweet spot between underfitting and overfitting.
- To understand this goal, we can look at the performance of the model over both the training and test sets.
- As we increase the complexity of the model, the error for the model on the training data goes down and so does the error on the test dataset.

WWW.KENT.EDU

So, the objective is to construct a model that not only is able to train well, i.e., explain the relationship between X and Y, but also able to predict future or unseen values correctly. A small model will typically underfit, and lead to bias, while larger models will overfit, and lead to overfitting. Thus, there is a tradeoff. To understand this tradeoff, we can look at the performance of the model on the training set, which will give us an indication of underfitting, and the performance on the test set, which will provide an indication of overfitting.



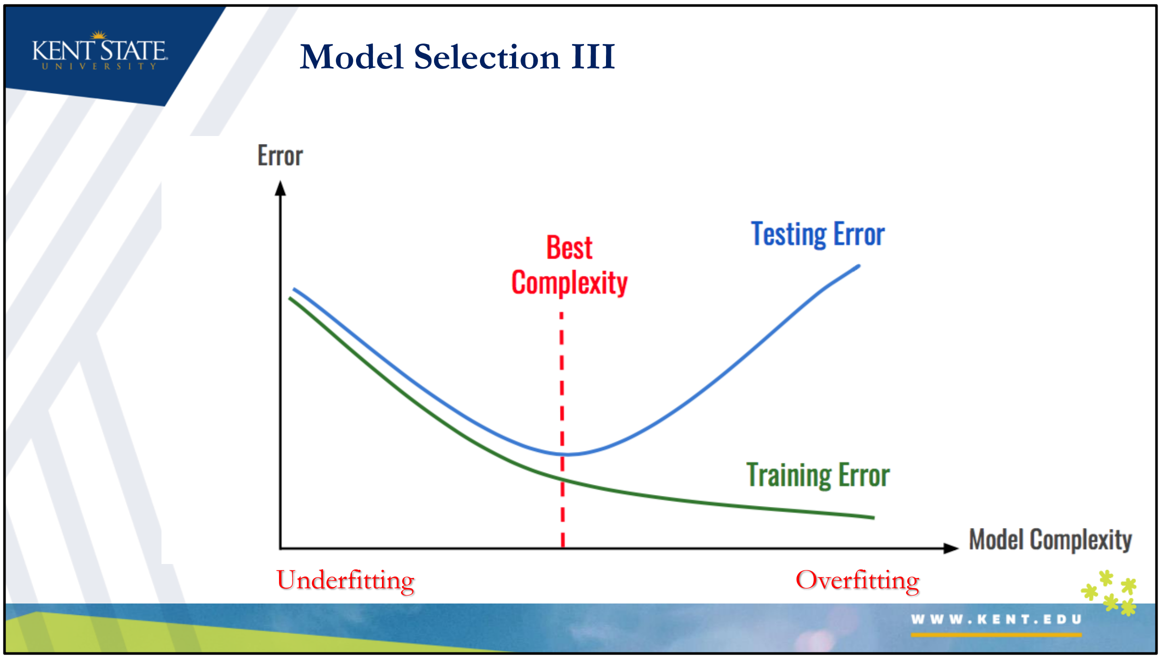
KENT STATE UNIVERSITY

Model Selection II

- As the model complexity is increased, the performance on the training dataset may continue to decrease because the model is overfitting and learning the irrelevant detail and noise in the training dataset.
- At the same time the error for the test set starts to rise again as the model's ability to generalize decreases.
- The sweet spot is the point just before the error on the test dataset starts to increase where the model has good performance on both the training dataset and the unseen test dataset.

WWW.KENT.EDU

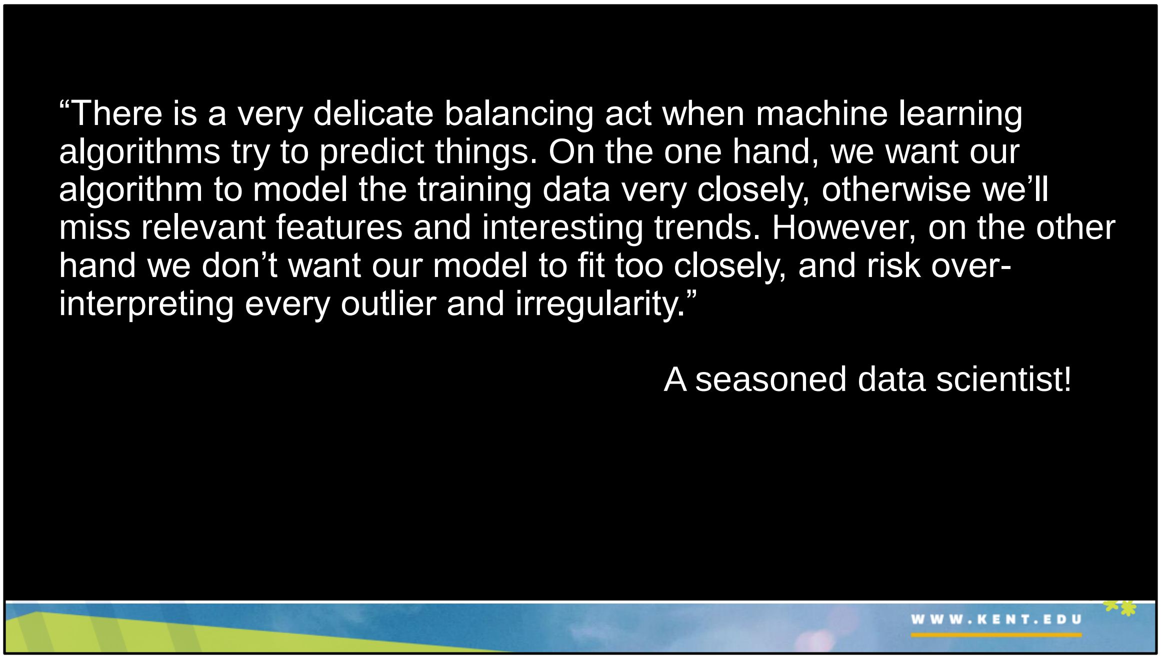
As model complexity increases, it is rare, but can happen, that the performance on the training set will deteriorate. Typically, as model complexity increases, the training set performance will become better. For the test set, initially, the performance will improve, but after some point of model complexity, the performance will worsen. The next graph illustrates this.



Model Selection III

WWW.KENT.EDU

This graphs is, of course, an oversimplification of what you might observe on real-world problems, but it does illustrate that as model complexity increases training error reduces, while testing error follows a nonlinear curve. There is a point at which we get low training and testing error. So, the model is able to explain current data well, and also is a good predictor. Note that in most cases, we tend to choose models based on the test set. Secondly, once a model has been chosen, we tend to retrain the model by combining the train and test datasets. This so as it is nearly always beneficial to have larger amount of data to construct the model. More on this in later lectures.



> “There is a very delicate balancing act when machine learning algorithms try to predict things. On the one hand, we want our algorithm to model the training data very closely, otherwise we’ll miss relevant features and interesting trends. However, on the other hand we don’t want our model to fit too closely, and risk over-interpreting every outlier and irregularity.”

A seasoned data scientist!

This concludes our lecture on underfitting and overfitting.