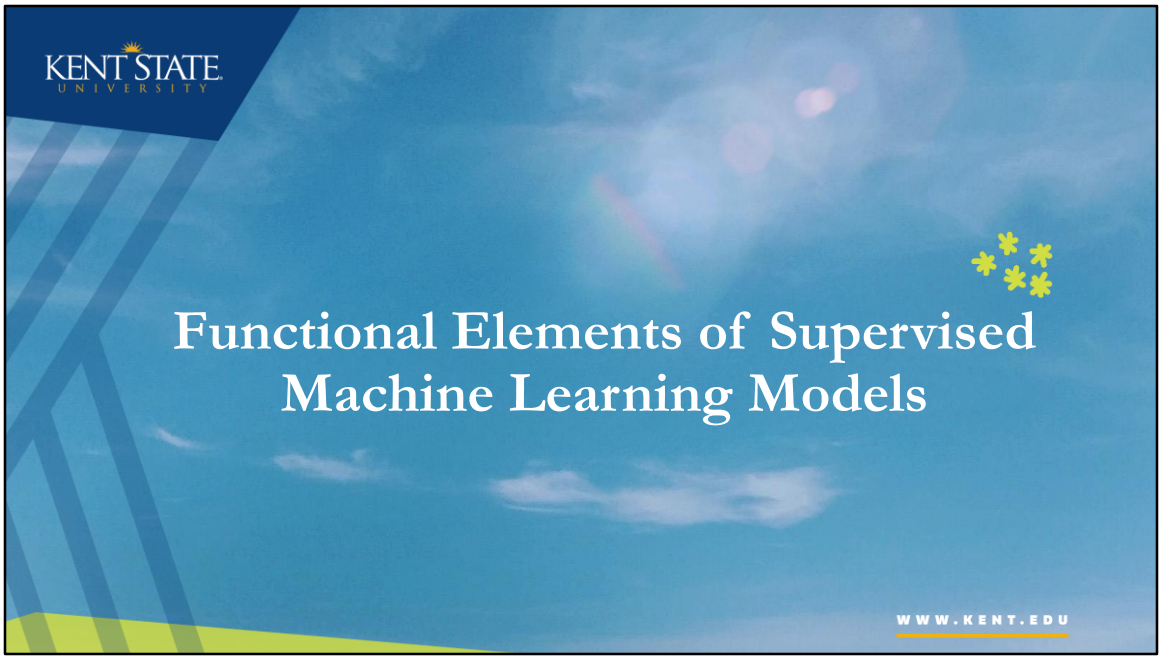
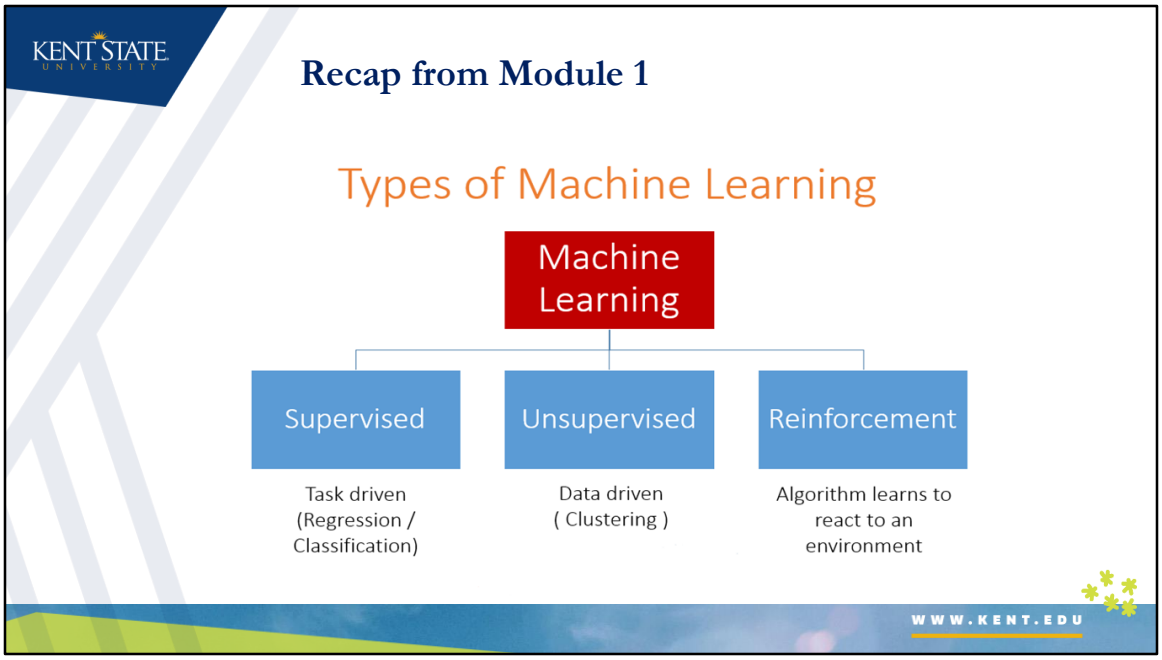


In this module, we will look at the basic components of supervised learning models.



As we saw earlier, there were three, really four, types of ML algorithms. Here, we study supervised learning models. These are models that require a target, usually annotated by human. The main types of models that we will study here are Regression, and Classification models.

Supervised Machine Learning Models

- In supervised learning, we are given a data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output.
- Let ***x*** to represent "input" variables (e.g. living area of a house), also called input features, and ***y*** to denote the "output" or target variable that we are trying to predict.
- A supervised machine learning modelling can be seen as the task of finding a mapping function, ***h***, that can accurately map the values of ***x*** to ***y***.

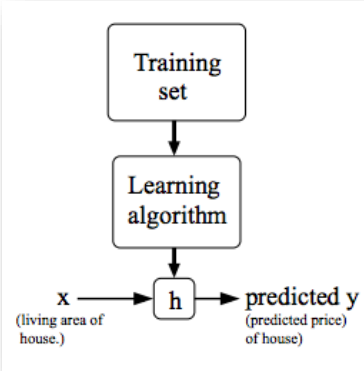
WWW.KENT.EDU

We introduced the concept of X and Y earlier. To illustrate, Y here represents the target, or in other terminology, the dependent or response variable. X here represents the input, or independent or predictor variables. The objective is to find a function h that maps X to Y. In mathematical notation Y = h(X).

KENT STATE UNIVERSITY

Supervised Machine Learning Models continued

- Let x^i to represent i-th element of x and y^i to represent i-th element of y.
- A pair (x^i, y^i) is called a training example, and the list of the dataset that contains a list of m training examples (i.e. (x^i, y^i); i=1,2,3,…m) is called a training set.
- The role of the learning algorithm is to use the training set to approximate a function h, that can accurately map the values of x to y.



WWW.KENT.EDU

To determine the function form h, we supply samples of (X,Y) values to the learning algorithm. Based on these samples, the learning algorithm determines an approximation to the hidden, and unknown, function h. This process of supplying samples to the learning algorithm so that it can map X to Y is called *learning* in ML terminology, and the samples themselves are called a training set.

Remember, though, our primary purpose in doing this is to really use this model to predict. That is, to guess the value of Y for values of X that the learning algorithm may or may not have seen as part of the training set. Clearly, the performance of the ML algorithm on a training set is likely to be better than its performance on an unseen dataset. Can you think of why?

Example: House Price Prediction

House Area (Square foot) | Price ($)

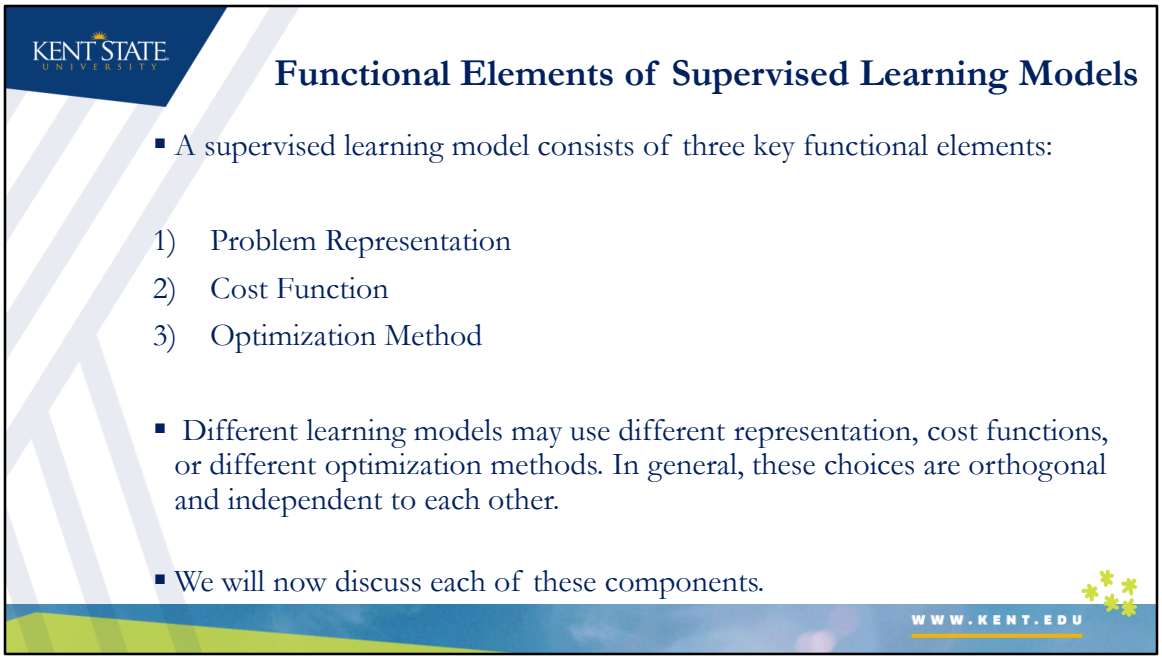
	X	Y
1	3392	339000
2	4100	899900
3	3200	448641
4	1436	239999
5	1944	377500
6	1500	299900
7	1700	265000
8	2507	449000
9	1580	439950
10	1500	699888

Training Example

Training Set

WWW.KENT.EDU

In this example, we have one independent variable X (House Area) and one dependent variable (Price). A total of 10 samples are available, i.e., pairs of (X,Y). If that is all we have, then this would comprise our training set.



There are three key elements to any supervised learning model. Problem representation, cost function, and optimization method.

Problem representation defines the mapping function h, which can take many forms. The cost function defines how we measure accuracy, and the optimization method is the process of determining h so as to improve our cost function. Let us look at each in turn.

Problem Representation

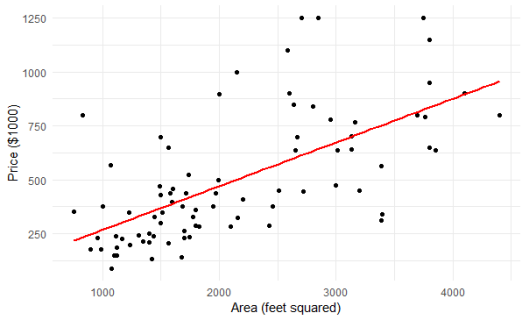
- As the name implies, problem representation defines the general structure of the mapping function, $\boldsymbol{h}$.
- For example, the mapping function can be represented through a linear function that takes $\boldsymbol{x}$ values and produces estimated $\boldsymbol{y}$ values, denoted as $\hat{\boldsymbol{y}}$, (e.g. liner regression $\hat{\boldsymbol{y}} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\,\boldsymbol{x}$)
- Another example is a decision tree representation where $\hat{\boldsymbol{y}}$ is determined by following a series of conditional statements.

There are different types of mapping functions, and h can take any form. H can be linear as in linear regression, it can be non-linear as it is in deep learning networks, or it could take the form of a set of rules, as in a decision tree. Regardless, h represents the mapping from X to Y.

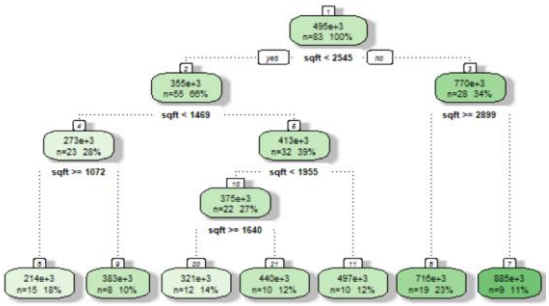
KENT STATE UNIVERSITY

Problem Representation: Example

Linear Function Representation



Decision Tree Representation



WWW.KENT.EDU

Here are two representations, one linear, and the other a decision tree. Note that regardless of representation, we can still use the same cost function, though the optimization method may differ. This is because the optimization method is usually dependent on both the cost function and the functional form of h.

KENT STATE UNIVERSITY

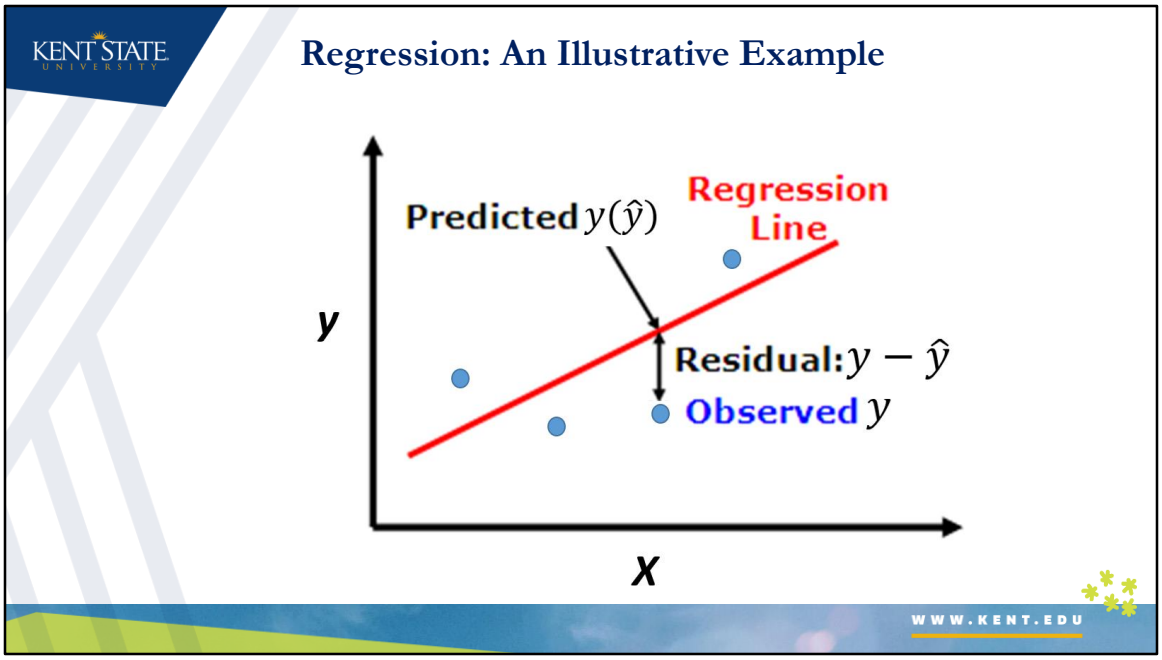
Cost Function

- We can measure the accuracy of our mapping function, ***h***, by using a cost function.
- For example, a commonly used cost function for linear regression models is sum square of errors (SSE).
- Error terms, also referred to as residuals, are calculated as the difference between the actual ($\boldsymbol{y}$) and estimated ($\hat{\boldsymbol{y}}$) values of the target variable:

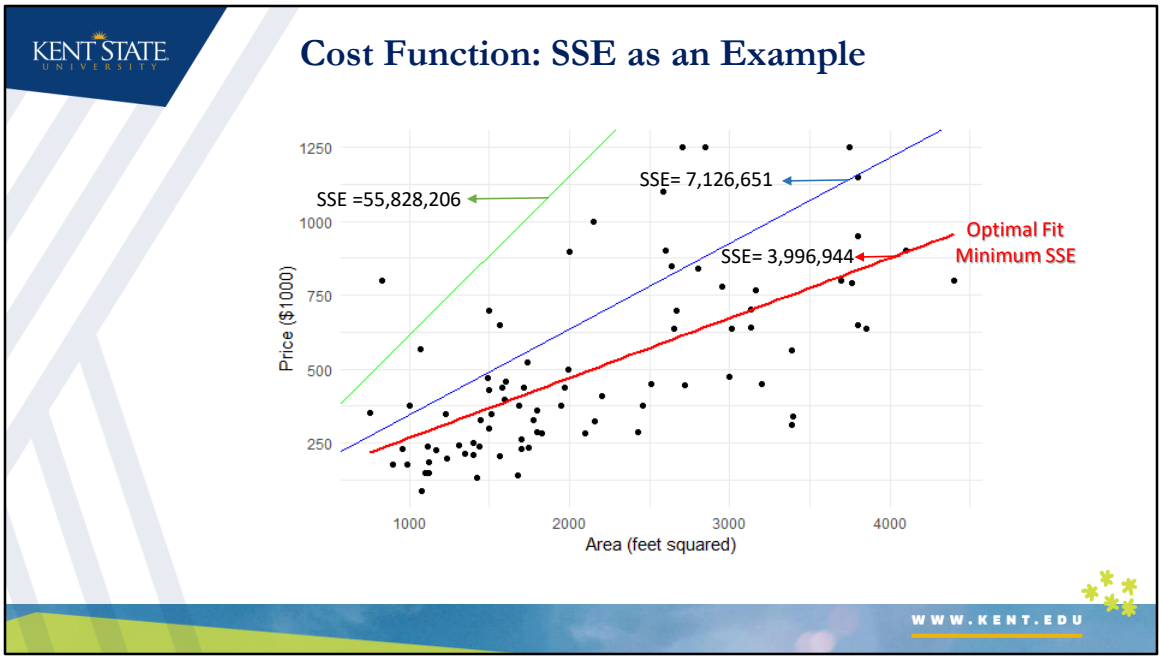
$$SSE = \sum_{i=1}^{m} (y - \hat{y})^2$$

WWW.KENT.EDU

If you are familiar with linear regression, then you will know that the cost function used there is SSE. There are statistical and computational reasons for using SSE, rather than say the sum of absolute errors, for example. Nevertheless, it is very important to define accuracy measures appropriate for the task.



Notice the errors here are defined as the difference between the actual value Y, and the predicted value Yhat. This difference is called Residual. The SSE cost function that we saw in the previous slide then therefore estimates the Sum of Squared Residuals. For a linear regression, the objective becomes to minimize the SSE. It turns out that it is relatively easy to calculate the estimates of the regression line that minimizes this function. Secondly, these parameter estimates have the statistical properties that indicate that they are the best linear unbiased estimates (BLUE). In other words, you will not be able to find better linear estimates than what we calculated using this method.



This graph illustrates the part of the properties of the least-squares (regression) line.

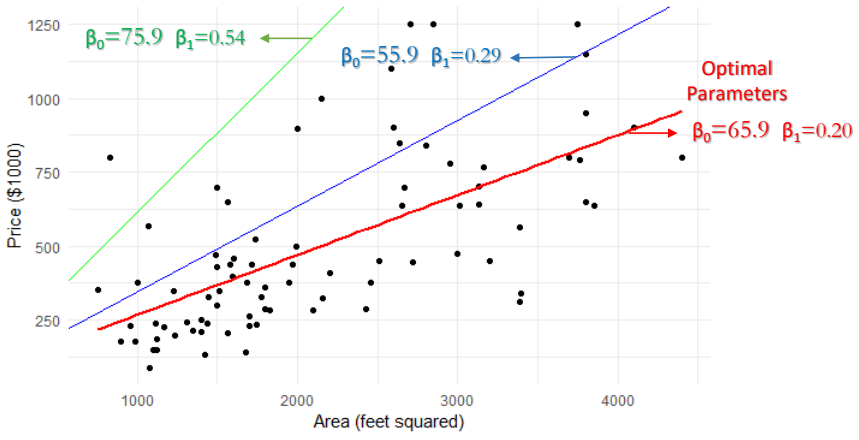
Optimization Method

- As we saw the loss function is a mathematical way of measuring how wrong your predictions are.
- During the training process, the learning algorithm tweak and change the parameters of the model to try and minimize that loss function. This is where optimizers come in.
- In other words, the optimizer shape and mold the model into its most accurate possible form by finding the optimal value of model parameters.

WWW.KENT.EDU

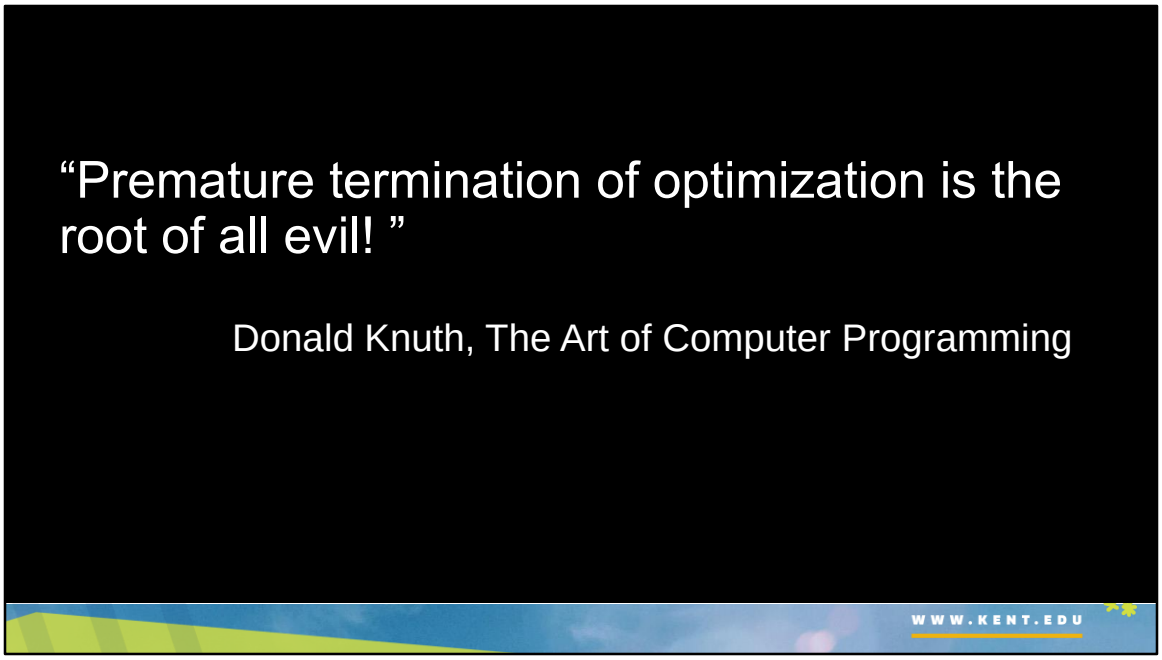
How does one calculate the parameter estimates of the regression line? That is, how do we determine these values to minimize our cost function, SSE? That process is our optimization method. As we proceed deeper into other ML models, the optimization method will change. But, for linear regression, we can actually use calculus to determine these optimal parameter estimates. Note that we could have also used other methods. For example, we could have done a search for these values, each time calculating the value of SSE. These search methods are not efficient for this model, but for other ML models, they might prove more effective.

Parameter Optimization: Example



WWW.KENT.EDU

If we were to use search, instead of calculus, to calculate the linear regression parameter estimates, what would your range of search be? How would you start?



This concludes our lecture on the functional elements for supervised learning models.