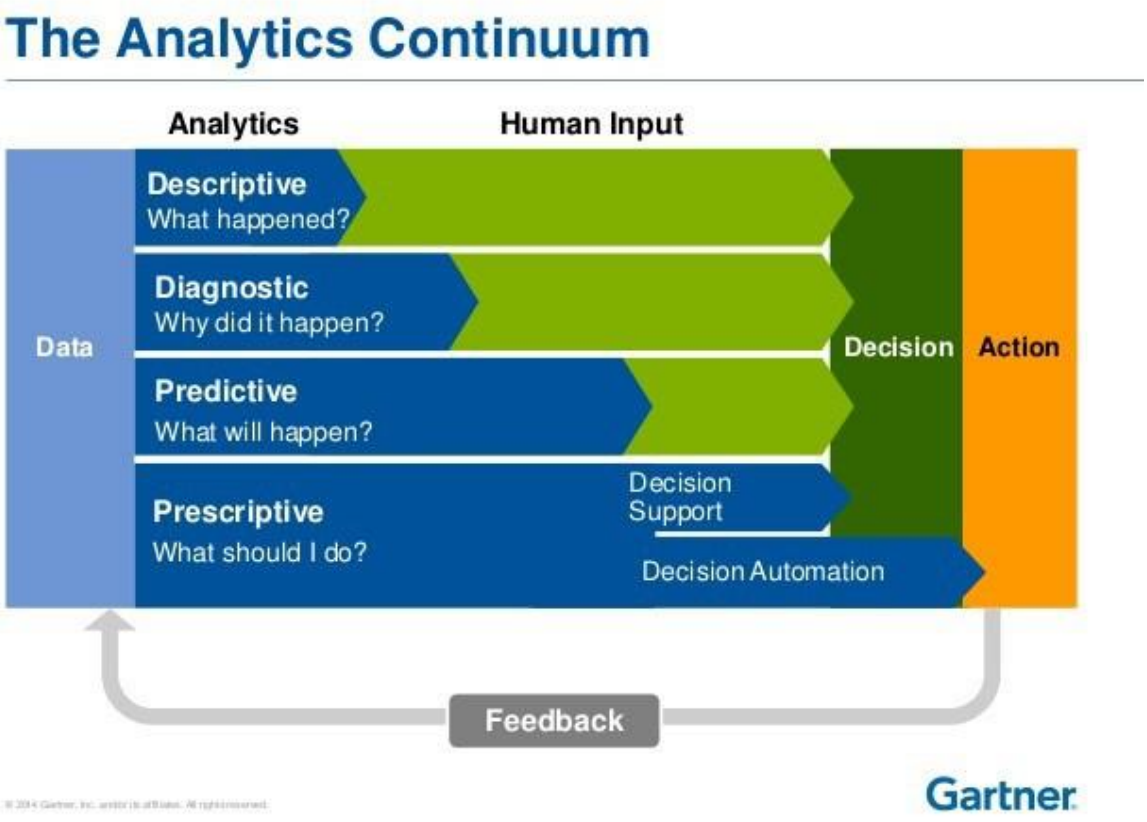
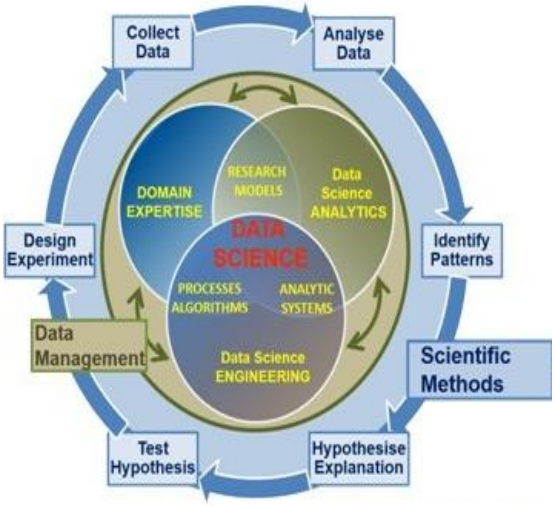
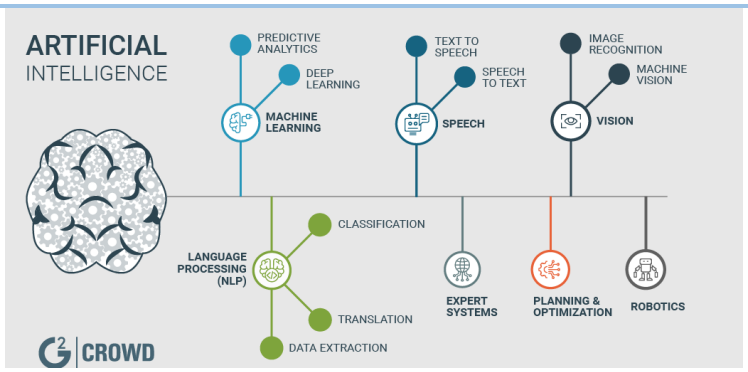
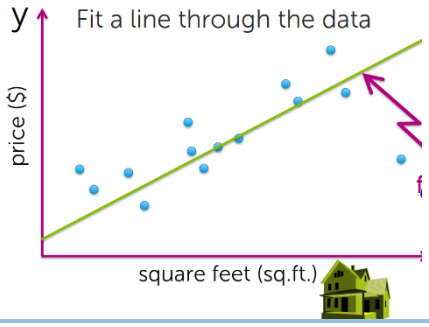
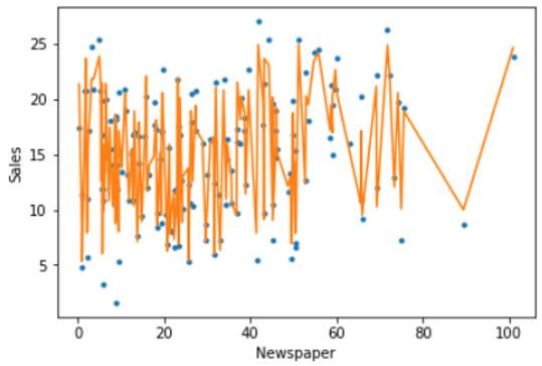
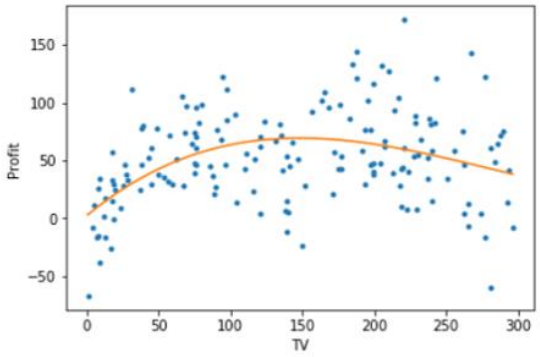
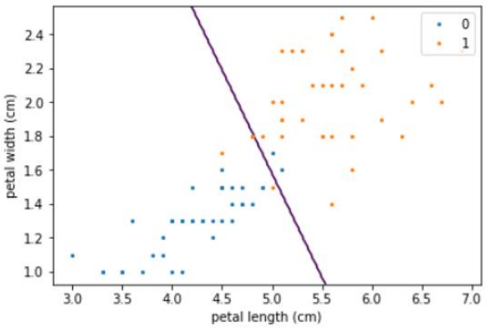
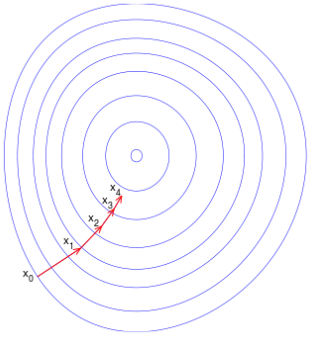
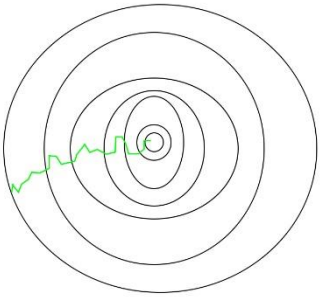
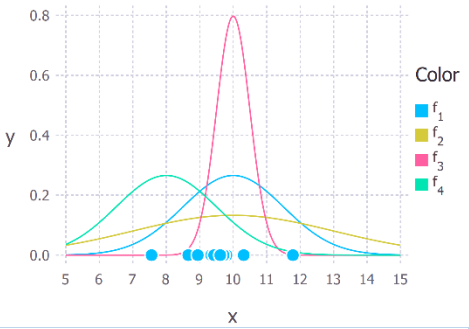
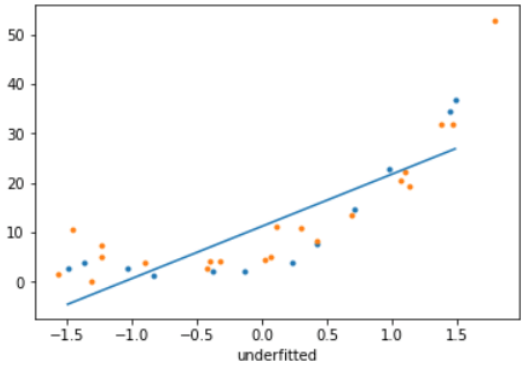
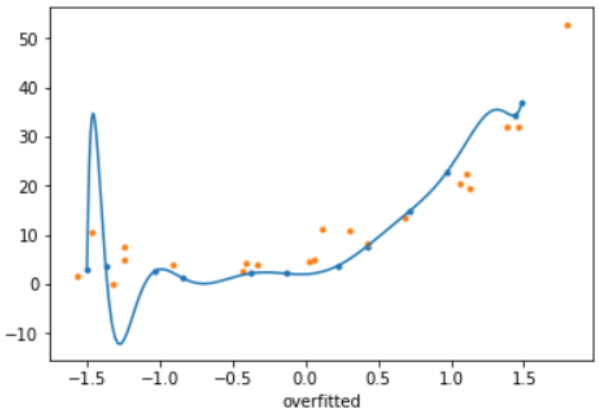
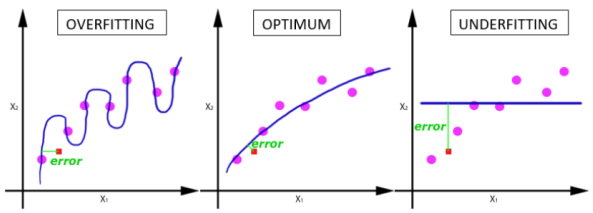
Data science: Course Machine Learning

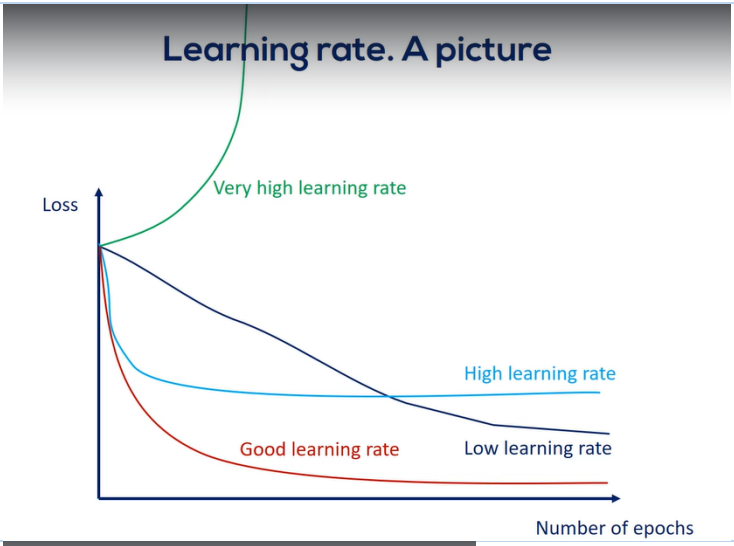
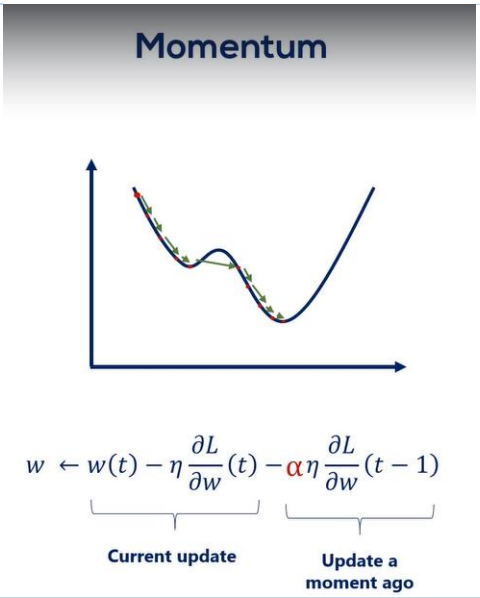


Since there are many technical words that are recurring in the lectures, I have decided to make a small glossary that explains briefly most of the machine learning terms.

Word	meaning	example
Data Science	The extraction of knowledge from large volumes of data that are structured or unstructured, which is a continuation of the field of Data Mining and Predictive Analytics.	
Artificial Intelligence	The Simulation of human intelligence by machines, such as abstraction, learning or problem solving	
Machine Learning	Scientific study of algorithms and statistical models that computers use to perform a specific task without using explicit instruction, relying on learned patterns and inference functions instead.	Computer program that learns to improve actions based on experiences. Suppose we want a model to predict the price of a house, we could use the data from houses relative to the house you want to predict and by this information the model will be able to predict the price of a house to a certain degree of accuracy.
Linear regression	Learn a linear regression function for prediction. Fitting a line through the data. Predict the best possible line that minimizes the distance between each pint to the line.	

Multivariate Regression	Use multiple explanatory variables. A normal regression that has multiple explanatory variables.	
Polynomial Regression	Polynomial Regression is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial.	
Classification	classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs to. Red/ blue category.	
Gradient descent	While training the model, the model calculates the cost function which measures the error between the predicted value (pred) and true value (y). The model targets to minimize the cost function. This is the goal of gradient descent. It iterates over the whole dataset before updating the weights, each update is very small.	

SGD	SGD is a technique that is used to find the minima of a function. Instead of updating the weights after iterating over the whole data set, it updates the data set after each epoch and the data set is divided into batches of one.	
Maximum likelihood estimation	Maximum likelihood estimation is a method that will find the values of μ (mean) and σ (variance) that result in the curve that best fits the data. It is used for getting the best fit sigmoid function.	
High Bias **Aka: underfitting**	The class of models is unable to fit the data, i.e. the systematic error of the model, how much the true value differs from the 'best possible prediction'. Possible reason: too simple model. Results underfitting.	
High Variance **Aka overfitting**	The class of models could fit the data but it doesn't because parameters are hard to optimize, i.e. the variance of the mean over different systems. Too many parameters. More like connecting the dots. Results overfitting.	
Epochs	One Epoch is when an ENTIRE dataset is passed forward and backward through the neural network only ONCE.	*One epoch leads to underfitting of the curve in the graph (below).* 

Batch	Divide dataset into Number of Batches or sets or parts.	
Batch size	Total number of training examples present in a single batch.	
Iteration	Iterations is the number of batches needed to complete one epoch.	Example: (data set of 2000) We can divide the dataset of 2000 examples into batches of 500 then it will take 4 iterations to complete 1 epoch.
Learning rate	The amount that the weights are updated during training is referred to as the step size or the "learning rate."	
Momentum	An update rule on the gradient descent that allows you to converge at the global minimum instead of a local one. Momentum is a hyperparameter just like the learning rate.	

Adam	Adaptive momentum gradient descent: It combines the momentum rule with the RMSprop into a new formula. This makes it a very good/ advanced optimizer.	**Adam** /adaptive moment estimation/ $\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t)} + \epsilon}\frac{\partial L}{\partial w_i}(t)$ $w \leftarrow w(t) - \eta\frac{\partial L}{\partial w}(t) - \alpha\eta\frac{\partial L}{\partial w}(t-1)$ $\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t)} + \epsilon} M_i(t)$ $M_i(t) = \alpha M_i(t-1) + (1-\alpha)\frac{\partial L}{\partial w_i}(t)$ $M_i(0) = 0$