What is Machine Learning?

What is Machine Learning?

Two definitions of Machine Learning are offered. Arthur Samuel described it as: "the field of study that gives computers the ability to learn without being explicitly programmed." This is an older, informal definition.

Tom Mitchell provides a more modern definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Example: playing checkers.

E = the experience of playing many games of checkers

T = the task of playing checkers.

P = the probability that the program will win the next game.

In general, any machine learning problem can be assigned to one of two broad classifications:

Supervised learning and Unsupervised learning.

Supervised Learning

Supervised Learning

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.

Supervised learning problems are categorized into "regression" and "classification" problems. In a regression problem, we are trying to predict results within a continuous output, meaning that we are trying to map input variables to some continuous function. In a classification problem, we are instead trying to predict results in a discrete output. In other words, we are trying to map input variables into discrete categories.

**Example 1:**

Given data about the size of houses on the real estate market, try to predict their price. Price as a function of size is a continuous output, so this is a regression problem.

We could turn this example into a classification problem by instead making our output about whether the house "sells for more or less than the asking price." Here we are classifying the houses based on price into two discrete categories.

**Example 2**:

(a) Regression - Given a picture of a person, we have to predict their age on the basis of the given picture

(b) Classification - Given a patient with a tumor, we have to predict whether the tumor is malignant or benign.

Unsupervised Learning

Unsupervised Learning

Unsupervised learning allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables.

We can derive this structure by clustering the data based on relationships among the variables in the data.

With unsupervised learning there is no feedback based on the prediction results.

**Example:**

Clustering: Take a collection of 1,000,000 different genes, and find a way to automatically group these genes into groups that are somehow similar or related by different variables, such as lifespan, location, roles, and so on.

Non-clustering: The "Cocktail Party Algorithm", allows you to find structure in a chaotic environment. (i.e. identifying individual voices and music from a mesh of sounds at a [cocktail party](https://en.wikipedia.org/wiki/Cocktail_party_effect)).

Model Representation

To establish notation for future use, we’ll use x^{(i)}*x*(*i*) to denote the “input” variables (living area in this example), also called input features, and y^{(i)}*y*(*i*) to denote the “output” or target variable that we are trying to predict (price). A pair (x^{(i)} , y^{(i)} )(*x*(*i*),*y*(*i*)) is called a training example, and the dataset that we’ll be using to learn—a list of m training examples {(x^{(i)} , y^{(i)} ); i = 1, . . . , m}(*x*(*i*),*y*(*i*));*i*=1,...,*m*—is called a training set. Note that the superscript “(i)” in the notation is simply an index into the training set, and has nothing to do with exponentiation. We will also use X to denote the space of input values, and Y to denote the space of output values. In this example, X = Y = ℝ.

To describe the supervised learning problem slightly more formally, our goal is, given a training set, to learn a function h : X → Y so that h(x) is a “good” predictor for the corresponding value of y. For historical reasons, this function h is called a hypothesis. Seen pictorially, the process is therefore like this:

Diagram

Description automatically generated

When the target variable that we’re trying to predict is continuous, such as in our housing example, we call the learning problem a regression problem. When y can take on only a small number of discrete values (such as if, given the living area, we wanted to predict if a dwelling is a house or an apartment, say), we call it a classification problem.

Cost Function

We can measure the accuracy of our hypothesis function by using a **cost function**. This takes an average difference (actually a fancier version of an average) of all the results of the hypothesis with inputs from x's and the actual output y's.

J(θ0​,θ1​)=2m1​i=1∑m​(y^​i​−yi​)2=2m1​i=1∑m​(hθ​(xi​)−yi​)2

To break it apart, it is \frac{1}{2}21​ \bar{x}*x*ˉ where \bar{x}*x*ˉ is the mean of the squares of h\_\theta (x\_{i}) - y\_{i}*hθ*​(*xi*​)−*yi*​ , or the difference between the predicted value and the actual value.

This function is otherwise called the "Squared error function", or "Mean squared error". The mean is halved \left(\frac{1}{2}\right)(21​) as a convenience for the computation of the gradient descent, as the derivative term of the square function will cancel out the \frac{1}{2}21​ term. The following image summarizes what the cost function does:

Text, letter

Description automatically generated

Cost Function - Intuition I

If we try to think of it in visual terms, our training data set is scattered on the x-y plane. We are trying to make a straight line (defined by h\_\theta(x)*hθ*​(*x*)) which passes through these scattered data points.

Our objective is to get the best possible line. The best possible line will be such so that the average squared vertical distances of the scattered points from the line will be the least. Ideally, the line should pass through all the points of our training data set. In such a case, the value of J(\theta\_0, \theta\_1)*J*(*θ*0​,*θ*1​) will be 0. The following example shows the ideal situation where we have a cost function of 0.

Chart, scatter chart

Description automatically generated

When \theta\_1 = 1*θ*1​=1, we get a slope of 1 which goes through every single data point in our model. Conversely, when \theta\_1 = 0.5*θ*1​=0.5, we see the vertical distance from our fit to the data points increase.

Chart, scatter chart

Description automatically generated

This increases our cost function to 0.58. Plotting several other points yields to the following graph:

Diagram

Description automatically generated

Thus as a goal, we should try to minimize the cost function. In this case, \theta\_1 = 1*θ*1​=1 is our global minimum.

Cost Function - Intuition II

A contour plot is a graph that contains many contour lines. A contour line of a two variable function has a constant value at all points of the same line. An example of such a graph is the one to the right below.

Chart, diagram

Description automatically generated

Taking any color and going along the 'circle', one would expect to get the same value of the cost function. For example, the three green points found on the green line above have the same value for J(\theta\_0,\theta\_1)*J*(*θ*0​,*θ*1​) and as a result, they are found along the same line. The circled x displays the value of the cost function for the graph on the left when \theta\_0*θ*0​ = 800 and \theta\_1*θ*1​= -0.15. Taking another h(x) and plotting its contour plot, one gets the following graphs:

Chart

Description automatically generated

When \theta\_0*θ*0​ = 360 and \theta\_1*θ*1​ = 0, the value of J(\theta\_0,\theta\_1)*J*(*θ*0​,*θ*1​) in the contour plot gets closer to the center thus reducing the cost function error. Now giving our hypothesis function a slightly positive slope results in a better fit of the data.

Chart, line chart

Description automatically generated

The graph above minimizes the cost function as much as possible and consequently, the result of \theta\_1*θ*1​ and \theta\_0*θ*0​ tend to be around 0.12 and 250 respectively. Plotting those values on our graph to the right seems to put our point in the center of the inner most 'circle'.

# Gradient Descent

So we have our hypothesis function and we have a way of measuring how well it fits into the data. Now we need to estimate the parameters in the hypothesis function. That's where gradient descent comes in.

Imagine that we graph our hypothesis function based on its fields \theta\_0*θ*0​ and \theta\_1*θ*1​ (actually we are graphing the cost function as a function of the parameter estimates). We are not graphing x and y itself, but the parameter range of our hypothesis function and the cost resulting from selecting a particular set of parameters.

We put \theta\_0*θ*0​ on the x axis and \theta\_1*θ*1​ on the y axis, with the cost function on the vertical z axis. The points on our graph will be the result of the cost function using our hypothesis with those specific theta parameters. The graph below depicts such a setup.

Chart, surface chart

Description automatically generated

We will know that we have succeeded when our cost function is at the very bottom of the pits in our graph, i.e. when its value is the minimum. The red arrows show the minimum points in the graph.

The way we do this is by taking the derivative (the tangential line to a function) of our cost function. The slope of the tangent is the derivative at that point and it will give us a direction to move towards. We make steps down the cost function in the direction with the steepest descent. The size of each step is determined by the parameter α, which is called the learning rate.

For example, the distance between each 'star' in the graph above represents a step determined by our parameter α. A smaller α would result in a smaller step and a larger α results in a larger step. The direction in which the step is taken is determined by the partial derivative of J(\theta\_0,\theta\_1)*J*(*θ*0​,*θ*1​). Depending on where one starts on the graph, one could end up at different points. The image above shows us two different starting points that end up in two different places.

The gradient descent algorithm is:

repeat until convergence:

\theta\_j := \theta\_j - \alpha \frac{\partial}{\partial \theta\_j} J(\theta\_0, \theta\_1)*θj*​:=*θj*​−*α*∂*θj*​∂​*J*(*θ*0​,*θ*1​)

where

j=0,1 represents the feature index number.

At each iteration j, one should simultaneously update the parameters \theta\_1, \theta\_2,...,\theta\_n*θ*1​,*θ*2​,...,*θn*​. Updating a specific parameter prior to calculating another one on the j^{(th)}*j*(*th*) iteration would yield to a wrong implementation.

A picture containing text

Description automatically generated

# Gradient Descent Intuition

In this video we explored the scenario where we used one parameter \theta\_1*θ*1​ and plotted its cost function to implement a gradient descent. Our formula for a single parameter was :

Repeat until convergence:

|  |
| --- |
| \theta\_1:=\theta\_1-\alpha \frac{d}{d\theta\_1} J(\theta\_1)*θ*1​:=*θ*1​−*αdθ*1​*d*​*J*(*θ*1​) |

Regardless of the slope's sign for \frac{d}{d\theta\_1} J(\theta\_1)*dθ*1​*d*​*J*(*θ*1​), \theta\_1*θ*1​ eventually converges to its minimum value. The following graph shows that when the slope is negative, the value of \theta\_1*θ*1​ increases and when it is positive, the value of \theta\_1*θ*1​ decreases.

Diagram

Description automatically generated with medium confidence

On a side note, we should adjust our parameter \alpha*α* to ensure that the gradient descent algorithm converges in a reasonable time. Failure to converge or too much time to obtain the minimum value imply that our step size is wrong.

Diagram

Description automatically generated with medium confidence

### **How does gradient descent converge with a fixed step size \alpha*α*?**

The intuition behind the convergence is that \frac{d}{d\theta\_1} J(\theta\_1)*dθ*1​*d*​*J*(*θ*1​) approaches 0 as we approach the bottom of our convex function. At the minimum, the derivative will always be 0 and thus we get:

|  |
| --- |
| \theta\_1:=\theta\_1-\alpha \* 0*θ*1​:=*θ*1​−*α*∗0 |

Graphical user interface

Description automatically generated with medium confidence

# Gradient Descent For Linear Regression

**Note:** [At 6:15 "h(x) = -900 - 0.1x" should be "h(x) = 900 - 0.1x"]

When specifically applied to the case of linear regression, a new form of the gradient descent equation can be derived. We can substitute our actual cost function and our actual hypothesis function and modify the equation to :

|  |
| --- |
| repeat until convergence: {*θ*0:=*θ*1:=}*θ*0−*α*1*m*∑*i*=1*m*(*hθ*(*xi*)−*yi*)*θ*1−*α*1*m*∑*i*=1*m*((*hθ*(*xi*)−*yi*)*xi*) |

where m is the size of the training set, \theta\_0*θ*0​ a constant that will be changing simultaneously with \theta\_1*θ*1​ and x\_{i}, y\_{i}*xi*​,*yi*​are values of the given training set (data).

Note that we have separated out the two cases for \theta\_j*θj*​ into separate equations for \theta\_0*θ*0​ and \theta\_1*θ*1​; and that for \theta\_1*θ*1​ we are multiplying x\_{i}*xi*​ at the end due to the derivative. The following is a derivation of \frac {\partial}{\partial \theta\_j}J(\theta)∂*θj*​∂​*J*(*θ*) for a single example :

Text, letter

Description automatically generated

The point of all this is that if we start with a guess for our hypothesis and then repeatedly apply these gradient descent equations, our hypothesis will become more and more accurate.

So, this is simply gradient descent on the original cost function J. This method looks at every example in the entire training set on every step, and is called **batch gradient descent**. Note that, while gradient descent can be susceptible to local minima in general, the optimization problem we have posed here for linear regression has only one global, and no other local, optima; thus gradient descent always converges (assuming the learning rate α is not too large) to the global minimum. Indeed, J is a convex quadratic function. Here is an example of gradient descent as it is run to minimize a quadratic function.

Chart

Description automatically generated

The ellipses shown above are the contours of a quadratic function. Also shown is the trajectory taken by gradient descent, which was initialized at (48,30). The x’s in the figure (joined by straight lines) mark the successive values of θ that gradient descent went through as it converged to its minimum.

# Multiple Features

**Note:** [7:25 - \theta^T*θT* is a 1 by (n+1) matrix and not an (n+1) by 1 matrix]

Linear regression with multiple variables is also known as "multivariate linear regression".

We now introduce notation for equations where we can have any number of input variables.

|  |
| --- |
| *x*(*i*)*jx*(*i*)*mn*=value of feature *j* in the *ith* training example=the input (features) of the *ith* training example=the number of training examples=the number of features |

The multivariable form of the hypothesis function accommodating these multiple features is as follows:

h\_\theta (x) = \theta\_0 + \theta\_1 x\_1 + \theta\_2 x\_2 + \theta\_3 x\_3 + \cdots + \theta\_n x\_n*hθ*​(*x*)=*θ*0​+*θ*1​*x*1​+*θ*2​*x*2​+*θ*3​*x*3​+⋯+*θn*​*xn*​

In order to develop intuition about this function, we can think about \theta\_0*θ*0​ as the basic price of a house, \theta\_1*θ*1​ as the price per square meter, \theta\_2*θ*2​ as the price per floor, etc. x\_1*x*1​ will be the number of square meters in the house, x\_2*x*2​ the number of floors, etc.

Using the definition of matrix multiplication, our multivariable hypothesis function can be concisely represented as:

|  |
| --- |
| *hθ*(*x*)=[*θ*0*θ*1...*θn*]⎡⎣⎢⎢⎢*x*0*x*1⋮*xn*⎤⎦⎥⎥⎥=*θTx* |

This is a vectorization of our hypothesis function for one training example; see the lessons on vectorization to learn more.

Remark: Note that for convenience reasons in this course we assume x\_{0}^{(i)} =1 \text{ for } (i\in { 1,\dots, m } )*x*0(*i*)​=1 for (*i*∈1,…,*m*). This allows us to do matrix operations with theta and x. Hence making the two vectors '\theta*θ*' and x^{(i)}*x*(*i*) match each other element-wise (that is, have the same number of elements: n+1).]

# Gradient Descent For Multiple Variables

## ****Gradient Descent for Multiple Variables****

The gradient descent equation itself is generally the same form; we just have to repeat it for our 'n' features:

|  |
| --- |
| }repeat until convergence:{*θ*0:=*θ*0−*α*1*m*∑*i*=1*m*(*hθ*(*x*(*i*))−*y*(*i*))⋅*x*(*i*)0*θ*1:=*θ*1−*α*1*m*∑*i*=1*m*(*hθ*(*x*(*i*))−*y*(*i*))⋅*x*(*i*)1*θ*2:=*θ*2−*α*1*m*∑*i*=1*m*(*hθ*(*x*(*i*))−*y*(*i*))⋅*x*(*i*)2⋯ |

In other words:

|  |
| --- |
| }repeat until convergence:{*θj*:=*θj*−*α*1*m*∑*i*=1*m*(*hθ*(*x*(*i*))−*y*(*i*))⋅*x*(*i*)*j*for j := 0...n |

The following image compares gradient descent with one variable to gradient descent with multiple variables:

Diagram, text

Description automatically generated

# Gradient Descent in Practice I - Feature Scaling

**Note:** [6:20 - The average size of a house is 1000 but 100 is accidentally written instead]

We can speed up gradient descent by having each of our input values in roughly the same range. This is because θ will descend quickly on small ranges and slowly on large ranges, and so will oscillate inefficiently down to the optimum when the variables are very uneven.

The way to prevent this is to modify the ranges of our input variables so that they are all roughly the same. Ideally:

−1 ≤ x\_{(i)}*x*(*i*)​ ≤ 1

or

−0.5 ≤ x\_{(i)}*x*(*i*)​ ≤ 0.5

These aren't exact requirements; we are only trying to speed things up. The goal is to get all input variables into roughly one of these ranges, give or take a few.

Two techniques to help with this are **feature scaling** and **mean normalization**. Feature scaling involves dividing the input values by the range (i.e. the maximum value minus the minimum value) of the input variable, resulting in a new range of just 1. Mean normalization involves subtracting the average value for an input variable from the values for that input variable resulting in a new average value for the input variable of just zero. To implement both of these techniques, adjust your input values as shown in this formula:

x\_i := \dfrac{x\_i - \mu\_i}{s\_i}*xi*​:=*si*​*xi*​−*μi*​​

Where μ\_i*μi*​ is the **average** of all the values for feature (i) and s\_i*si*​ is the range of values (max - min), or s\_i*si*​ is the standard deviation.

Note that dividing by the range, or dividing by the standard deviation, give different results. The quizzes in this course use range - the programming exercises use standard deviation.

For example, if x\_i*xi*​ represents housing prices with a range of 100 to 2000 and a mean value of 1000, then, x\_i := \dfrac{price-1000}{1900}*xi*​:=1900*price*−1000​.

# Gradient Descent in Practice II - Learning Rate

**Note:** [5:20 - the x -axis label in the right graph should be \theta*θ* rather than No. of iterations ]

**Debugging gradient descent.** Make a plot with number of iterations on the x-axis. Now plot the cost function, J(θ) over the number of iterations of gradient descent. If J(θ) ever increases, then you probably need to decrease α.

**Automatic convergence test.** Declare convergence if J(θ) decreases by less than E in one iteration, where E is some small value such as 10^{−3}10−3. However in practice it's difficult to choose this threshold value.

A picture containing graphical user interface

Description automatically generated

It has been proven that if learning rate α is sufficiently small, then J(θ) will decrease on every iteration.

Diagram

Description automatically generated

To summarize:

If \alpha*α* is too small: slow convergence.

If \alpha*α* is too large: ￼may not decrease on every iteration and thus may not converge.

# Features and Polynomial Regression

We can improve our features and the form of our hypothesis function in a couple different ways.

We can **combine** multiple features into one. For example, we can combine x\_1*x*1​ and x\_2*x*2​ into a new feature x\_3*x*3​ by taking x\_1*x*1​⋅x\_2*x*2​.

### **Polynomial Regression**

Our hypothesis function need not be linear (a straight line) if that does not fit the data well.

We can **change the behavior or curve** of our hypothesis function by making it a quadratic, cubic or square root function (or any other form).

For example, if our hypothesis function is h\_\theta(x) = \theta\_0 + \theta\_1 x\_1*hθ*​(*x*)=*θ*0​+*θ*1​*x*1​ then we can create additional features based on x\_1*x*1​, to get the quadratic function h\_\theta(x) = \theta\_0 + \theta\_1 x\_1 + \theta\_2 x\_1^2*hθ*​(*x*)=*θ*0​+*θ*1​*x*1​+*θ*2​*x*12​ or the cubic function h\_\theta(x) = \theta\_0 + \theta\_1 x\_1 + \theta\_2 x\_1^2 + \theta\_3 x\_1^3*hθ*​(*x*)=*θ*0​+*θ*1​*x*1​+*θ*2​*x*12​+*θ*3​*x*13​

In the cubic version, we have created new features x\_2*x*2​ and x\_3*x*3​ where x\_2 = x\_1^2*x*2​=*x*12​ and x\_3 = x\_1^3*x*3​=*x*13​.

To make it a square root function, we could do: h\_\theta(x) = \theta\_0 + \theta\_1 x\_1 + \theta\_2 \sqrt{x\_1}*hθ*​(*x*)=*θ*0​+*θ*1​*x*1​+*θ*2​*x*1​​

One important thing to keep in mind is, if you choose your features this way then feature scaling becomes very important.

eg. if x\_1*x*1​ has range 1 - 1000 then range of x\_1^2*x*12​ becomes 1 - 1000000 and that of x\_1^3*x*13​ becomes 1 - 1000000000

# Normal Equation

**Note:** [8:00 to 8:44 - The design matrix X (in the bottom right side of the slide) given in the example should have elements x with subscript 1 and superscripts varying from 1 to m because for all m training sets there are only 2 features x\_0*x*0​ and x\_1*x*1​. 12:56 - The X matrix is m by (n+1) and NOT n by n. ]

Gradient descent gives one way of minimizing J. Let’s discuss a second way of doing so, this time performing the minimization explicitly and without resorting to an iterative algorithm. In the "Normal Equation" method, we will minimize J by explicitly taking its derivatives with respect to the θj ’s, and setting them to zero. This allows us to find the optimum theta without iteration. The normal equation formula is given below:

\theta = (X^T X)^{-1}X^T y*θ*=(*XTX*)−1*XTy*

Diagram, schematic

Description automatically generated with medium confidence

There is **no need** to do feature scaling with the normal equation.

The following is a comparison of gradient descent and the normal equation:

|  |  |
| --- | --- |
| **Gradient Descent** | **Normal Equation** |
| Need to choose alpha | No need to choose alpha |
| Needs many iterations | No need to iterate |
| O (kn^2*kn*2) | O (n^3*n*3), need to calculate inverse of X^TX*XTX* |
| Works well when n is large | Slow if n is very large |

With the normal equation, computing the inversion has complexity \mathcal{O}(n^3)O(*n*3). So if we have a very large number of features, the normal equation will be slow. In practice, when n exceeds 10,000 it might be a good time to go from a normal solution to an iterative process.

Normal Equation Noninvertibility

When implementing the normal equation in octave we want to use the 'pinv' function rather than 'inv.' The 'pinv' function will give you a value of \theta*θ* even if X^TX*XTX* is not invertible.

If X^TX*XTX* is **noninvertible,** the common causes might be having :

* Redundant features, where two features are very closely related (i.e. they are linearly dependent)
* Too many features (e.g. m ≤ n). In this case, delete some features or use "regularization" (to be explained in a later lesson).

Solutions to the above problems include deleting a feature that is linearly dependent with another or deleting one or more features when there are too many features.