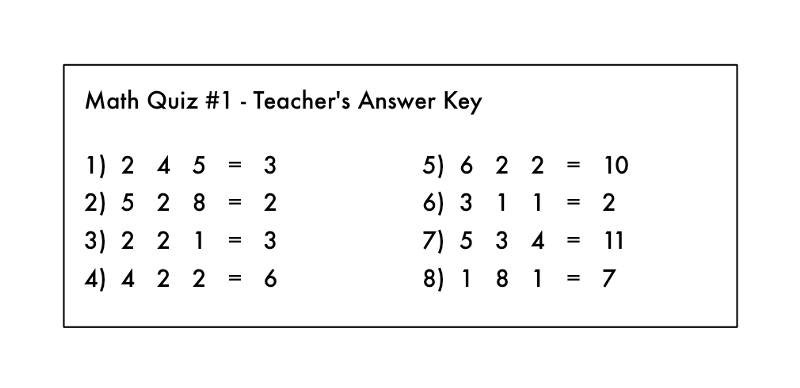
1. **Vấn đề đặt ra**

* Giả sử bạn là một người làm về bất động sản. Công ty của bạn phát triển nhanh, bạn tuyển hàng loạt thực tập. Nhưng có một vấn đề - bạn có thể dễ dàng nhìn lướt qua và đánh giá chính xác giá trị của một ngôi nhà, nhưng với các thực tập viên không có kinh nghiệm, họ không biết mỗi căn trị giá bao nhiêu.
* Để giúp đỡ các bạn thực tập này, bạn quyết định viết một ứng dụng nhỏ có thể ước lượng được giá của một ngôi nhà dựa vào diện tích, khu vực lân cận, ...
* Và bạn ghi lại thông tin của mọi căn nhà được bán trong thành phố, trong vòng 3 tháng. Với mỗi ngôi nhà, bạn ghi lại mọi thông tin: số phòng ngủ, diện tích (feet vuông), neighborhood, ... Nhưng quan trọng nhất là giá (price) cuối cùng của căn nhà được bán:



* Chúng ta có được "training data" , Với dữ liệu "training data" như trên, chúng ta muốn viết một ứng dụng có thể ước tính được giá của một căn nhà tương tự khác:



* Chúng ta muốn sử dụng training data để dự đoán giá của những ngôi nhà khác.
* Đây được gọi là supervised learning. Bạn biết được giá mỗi căn nhà được bán đi, nói cách khác, bạn biết được câu trả lời của bài toán, và có thể từ đây suy ra được logic của vấn đề.
* Để xây dựng ứng dụng này, bạn cho training data về mỗi ngôi nhà này vào một thuật toán machine learning. Thuật toán sẽ cố gắng tìm ra loại tính toán nào để các con số có thể work.
* Nó giống như việc tìm các toán tử trong bài tập toán hồi lớp 1 chúng ta vẫn hay được học:
* Từ bảng trên, bạn có thể tìm ra được các phép toán nào để có được kết quả bên phải? Bạn biết bạn có nghĩa vụ phải "làm gì đó" với những con số ở bên trái để có được câu trả lời ở bên phải.

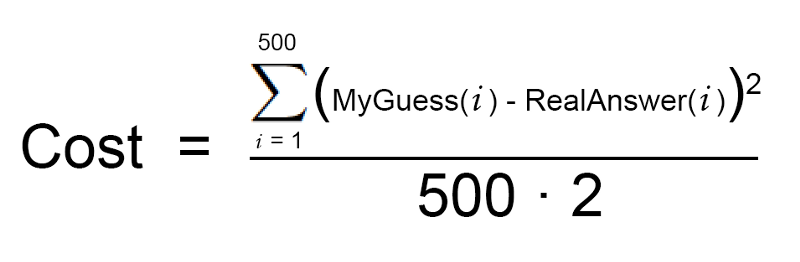
1. **Thuật toán áp dụng vào đề tài : Supervised Learning (Học có giám sát)**

* Supervised learning là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (input, outcome) đã biết từ trước. Cặp dữ liệu này còn được gọi là (data, label), tức (dữ liệu, nhãn). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning.
* cách toán học, Supervised learning là khi chúng ra có một tập hợp biến đầu vào X={x1,x2,…,xN}X={x1,x2,…,xN} và một tập hợp nhãn tương ứng Y={y1,y2,…,yN}Y={y1,y2,…,yN}, trong đó xi,yixi,yi là các vector. Các cặp dữ liệu biết trước (xi,yi)∈X×Y(xi,yi)∈X×Y được gọi là tập training data (dữ liệu huấn luyện). Từ tập traing data này, chúng ta cần tạo ra một hàm số ánh xạ mỗi phần tử từ tập XX sang một phần tử (xấp xỉ) tương ứng của tập YY:
* yi≈f(xi),  ∀i=1,2,…,Nyi≈f(xi),  ∀i=1,2,…,NMục đích là xấp xỉ hàm số ff thật tốt để khi có một dữ liệu xx mới, chúng ta có thể tính được nhãn tương ứng của nó y=f(x)y=f(x).
* Thuật toán gồm 2 loại : Classification (Phân loại) , Regression (Hồi quy)
* **Regression (Hồi quy)**
* Linear Regression
* Logistic Regression
* Stepwise Regression
* **Classification (Phân loại)**
* Linear Classifier
* Support Vector Machine (SVM)
* Kernel SVM
* Sparse Representation-based classification (SRC)

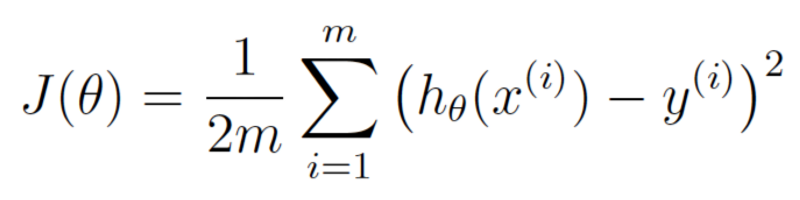
1. Mô tả chi tiết thuật toán

Để nhanh chóng tìm ra các weights này mà không cần phải thử quá nhiều lần. Đây là một cách:

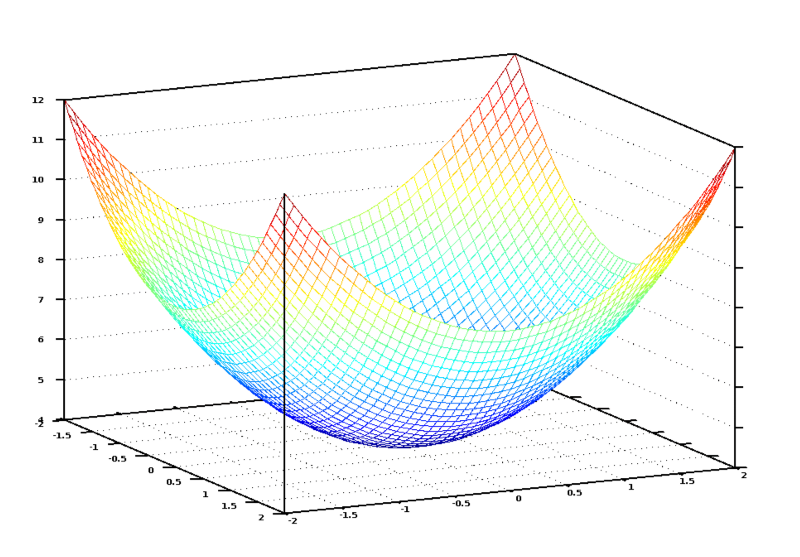
* Đầu tiên, viết một biểu thức đơn giản để biểu diễn cost :



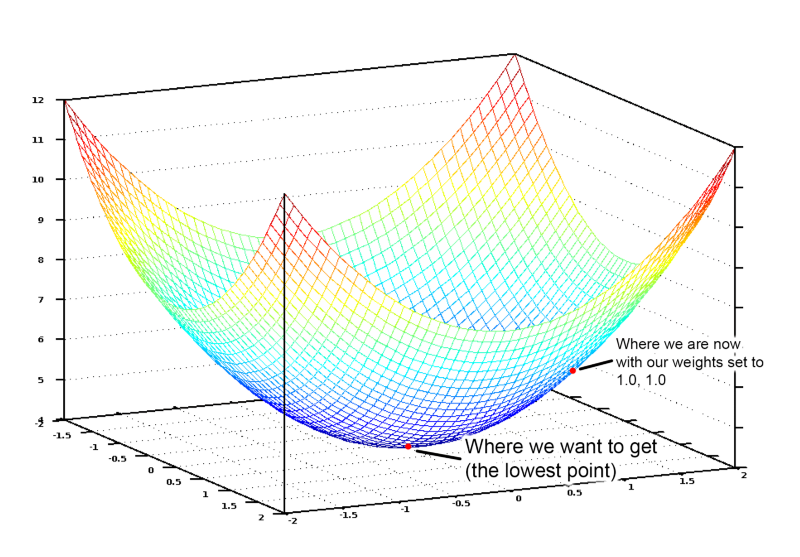
* Bây giờ chúng ta viết lại dưới dạng một biểu thức với ký hiệu toán machine learning



* Công thức trên biểu diễn độ sai của function ước lượng price của chúng ta với tập weights θ.
* Nếu chúng ta vẽ hết tất cả giá trị của biểu thức J(θ) với các weights có thể có ứng với number\_of\_bedrooms và sqft, biểu đồ sẽ có thể có dạng như sau:



* Trên hình trên, điểm thấp nhất màu xanh ứng với nơi có cost thấp nhất - vì thế chương trình sẽ có độ lỗi thấp nhất. Điểm càng cao sẽ có độ lệch/lỗi càng cao. Vì thế nếu chúng ta có thể tìm được weights đưa chúng ta đến điểm thấp nhất trên đồ thị, chúng ta sẽ tìm ra được câu trả lời



* Vì thế, chúng ta cần điều chỉnh weights, giống như việc "đi xuống thung lũng" trong đồ thị để tìm được điểm thấp nhất. Nếu như chúng ta điều chỉnh từng chút một và luôn đi xuống, ta sẽ tìm được điểm cực tiểu mà không cần phải thử quá nhiều weights.