# Week 5

## Regression

- Supervised learning

### Training data

$$\{X, y\}$$

- $X$ = feature/ data matrix of shape $(d, n)$
- y = label vector of shape $(n, 1)$

# Goal:

Given a data point, predict the outcome

Learn a function $h_w : \mathbb{R}^d \rightarrow \mathbb{R}$

- $h$ is called a **MODEL**.

It takes a data points and maps it to a real number.

- $w =$ weight vector (or parameter vector) of shape $(d, 1)$

Can we find such function??

How to find $w$??

Is there any measure to define the performance of the model?

# Error

Error = Predicted value $-$Actual value

Total loss/ Total error = $\sum_{i=1}^{n} (\mathrm{Error}_i)^2$

Is there any way to define error?

## Goal:

**Minimize the error**

How minimum it can be???

Why not the below model??

$$h_w(x_i) = y_i$$

# Linear regression

Assumption: Labels are linearly related to features with some noise.

- We only look for **linear function** that is

$$\begin{aligned} h_w(x) &= w_1 x^{(1)} + w_2 x^{(2)} + \ldots + w_n x^{(n)} \\ &= w^T x \end{aligned}$$

## We want to minimize the squared error

That is our goal is to

$$\min_{w} \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

## Vector notations

The above optimization problem can be written as

$$\min_{w} ||X^T w - y||^2$$

OR

$$\min_{w} (X^T w - y)^T (X^T w - y)$$

Does every model need pass through origin??

**Allow model to be any line (hyperplane) in the space.**

That is the model is

$$\begin{aligned} h_w(x) &= w_0 x^{(0)} + w_1 x^{(1)} + w_2 x^{(2)} + \ldots + w_n x^{(n)} \\ &= w^T x \end{aligned}$$

Add a dummy feature $x^0$ and set it to 1. that is

$$\begin{aligned} h_w(x) &= w_0 x + w_1 x^{(1)} + w_2 x^{(2)} + \ldots + w_n x^{(n)} \\ &= w^T x \end{aligned}$$

## Solution of optimization problem

- Normal equation
- Gradient descent

## Normal equation

$$\min_{w} \sum ||X^T w - y||^2$$

- Convex function
- Without any constraint

Take derivative and set it to 0

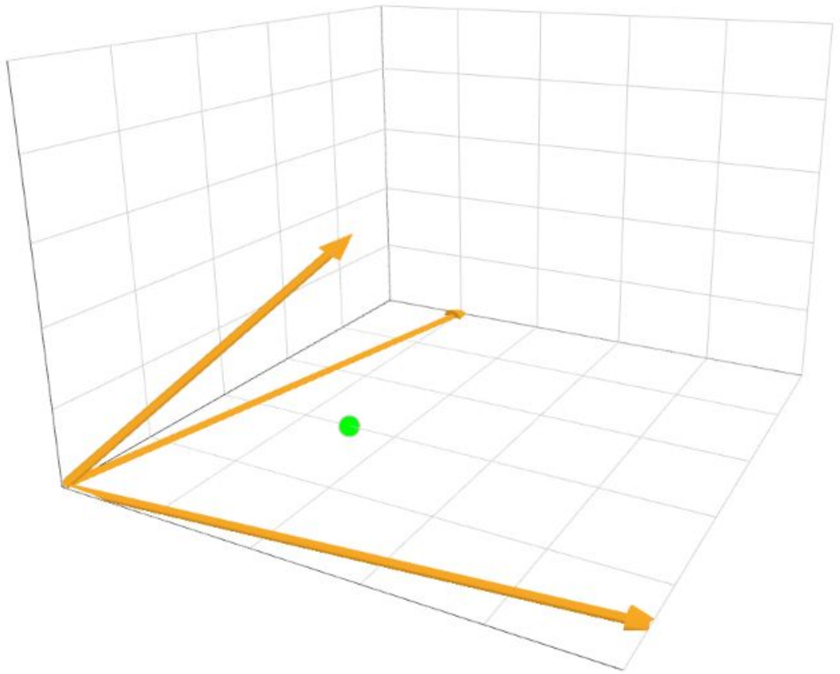
Let $L(w) = (X^T w - y)^T (X^T w - y)$

$L'(w) = 2(XX^T)w - 2Xy$

Setting it to zero, we have

$w^* = (XX^T)^\dagger Xy$

## Geometric interpretation



# Gradient descent

Why GD??

What is GD??

$\nabla f$ = direction of maximum slope at that point

- We want to move in the direction where the function value decreases
    - Move in the direction of gradient descent

But how much to move

- handle by a learning rate $\eta$.
    - A hyperparameter

$$w^{t+1} = w^t - \eta\nabla(f(w^t))$$
$$w^{t+1} = w^t - \eta(2(XX^T)w - 2Xy)$$

# Stochastic gradient descent

## Why?

- Don't update weights using all the samples
- Take random subsets of samples (batches) and update the weights in batches

## Probabilistic view of linear regression

- labels are linearly associated with features but with some noise.
- Assumption: noise $\epsilon \sim N(0, \sigma^2)$

That is

$$y_i | x_i = w^T x_i + \epsilon$$
$$y_i | x_i \sim N(w^T x_i, \sigma^2)$$

Can be estimate the parameters $w$???

- We have the data- The samples points
- We have distribution with unkonwn parameters $w$
    - Let's apply the MLE

$$
\begin{aligned}
L(w; X, y) &= \prod_{i=1}^{n} f_{y_i|x_i}(y_i) \\
&= \exp\left(\sum_{i=1}^{n} \frac{-(y_i - w^T x_i)^2}{2\sigma^2}\right) \\
\log L &= \sum_{i=1}^{n} \frac{-(y_i - w^T x_i)^2}{2\sigma^2}
\end{aligned}
$$

$$w^* = \argmax_w \sum_{i=1}^{n} \frac{-(y_i - w^T x_i)^2}{2\sigma^2}$$

$$= \argmin_w \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

**we have already solved the same problem.**

## Kernel regression

What if the data have some non-linear relationship??

- Can we transform the data into higher dimension such that it comes in linear relationship??

- Can we use the kernel function?

Remember the kernel function has been defined to follow:

$$k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

Can we write $w^*$ as a linear combination of data points??

$$
\begin{aligned}
w^* &= (\phi(X)\phi(X)^T)^\dagger \phi(X) y \\
\phi(X)\alpha &= (\phi(X)\phi(X)^T)^\dagger \phi(X) y
\end{aligned}
$$

$$\alpha = K^{-1}y$$

## Can we find the weight vector??

- If we have $\phi$: we can find

## Prediction for $x_t$

$$
\begin{aligned}
\hat{y} &= (w^*)^T \phi(x_t) \\
&= \left( \sum_{i=1}^{n} \alpha_i \phi(x_i) \right) \phi(x_t)^T \\
&= \sum_{i=1}^{n} \alpha_i k(x_i, x_t)
\end{aligned}
$$