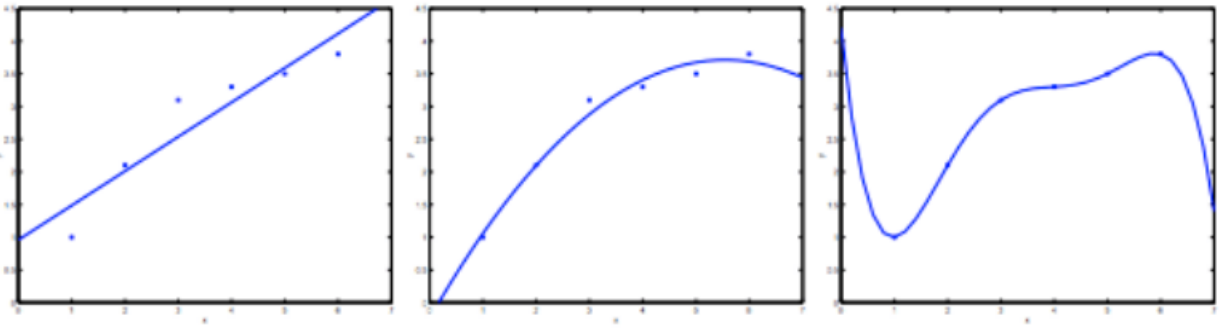
# Overfitting & Underfitting

## Overfitting Problem :

در بعضی اوقات ممکن است مدل یا خطی که از داده ها عبور می کند به طور کامل بر داده ها منطبق نباشد. می توانیم با استفاده از اضافه کردن ویژگی این مشکل را حل کنیم اما لزوما این روش خوب نخواهد بود.

برای مثال در شکل زیر هر چه ویژگی ها بیشتر شده است ، مدل بر داده ها انطباق بیشتری پیدا کرده است :

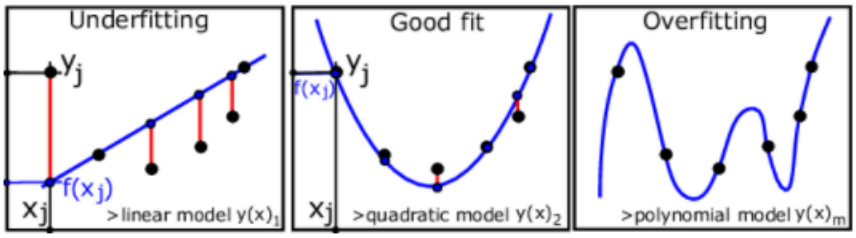


### ❖ Underfitting :

زمانی که مدل یا خط حاصل از تابع فرضیه بر داده ها انطباق کم یا ضعیفی داشته باشد به آن Underfitting می گویند. این اتفاق به دلیل خیلی ساده بودن تابع فرضیه و یا کم بودن ویژگی ها رخ می دهد.

### ❖ Overfitting :

زمانی که مدل یا خط حاصل از تابع فرضیه بر داده ها انطباق زیاد و خوبی داشته باشد به آن Overfitting می گویند. شاید تصور کنید که Overfitting اتفاق خوبی است اما به خاطر وجود انحناهای زیاد به دلیل ویژگی های زیاد ، منحنی یا مدل پیش بینی خوبی نخواهد کرد.




دو راه حل برای حل مشکل Overfitting وجود دارد :

۱- کاهش تعداد ویژگی که معمولا به دو روش انجام می شود. یا به طور دستی ویژگی های اضافی را حذف کنیم و یا از الگوریتم های انتخاب مدل استفاده کنیم.

۲- تنظیم پارامتر ( Regularization ) یعنی آن که همه ویژگی ها را نگه داریم اما پارامترهای مدل ( پارامترهای تتا ) مناسب را پیدا کنیم.

تنظیم پارامتر زمانی به خوبی عمل خواهد کرد که ویژگی های مفیدی داشته باشیم.

## ❖ تابع هزینه در Overfitting :

اگر تابع فرضیه مشکل Overfitting داشته باشد ، می توانیم با افزایش هزینه بعضی از بخش های تابع فرضیه ، وزن بخش های مد نظر را کاهش دهیم.

برای مثال اگر بخواهیم تاثیر $x^3$ و $x^4$ را از بین ببریم ، داریم :

$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$min_\theta \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + 1000 \cdot \theta_3^2 + 1000 \cdot \theta_4^2$$

تابع هزینه اصلاح شده به صورت بالا در می آید. دو بخش اضافه شده در آخر این تابع هزینه برای افزایش هزینه $x^3$ و $x^4$ است. حال باید تابع هزینه را به سمت صفر نزدیک کنیم که برای این کار $\theta_3, \theta_4$ را به صفر نزدیک می کنیم.

به طور کلی تابع هزینه به صورت زیر در می آید :

$$min_\theta \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2$$

$\lambda$ ، Regularization Parameter است که تعیین می کند هزینه پارامترهای مدل چه مقدار باید زیاد شود.

با استفاده از تابع هزینه بالا می توانیم مشکل Overfitting را حل کنیم.

**نکته !** اگر $\lambda$ خیلی بزرگ باشد باعث می شود تابع فرضیه بیش از اندازه ساده شود و Underfitting رخ می دهد.

## Regularized Linear Regression :

یکی از روش های محدود کردن وزن ها در رگرسیون خطی برای تنظیم پارامتر Ridge Regression است.

در این روش تابع هزینه به صورت زیر خواهد بود :

$$min_\theta \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2$$

حال باید با استفاده از گرادیان کاهشی پارامترهای مناسب را پیدا کرد.

حال اگر بخواهیم معادله نرمال را تنظیم کنیم که بدون نیاز به حلقه و تکرار باشد ، باید به شکل زیر معادله نرمال را بنویسیم :

$$\theta = \left(X^T X + \lambda \cdot L\right)^{-1} X^T y$$

$$\text{where } L = \begin{bmatrix} 0 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}$$

## Regularized Logistic Regression :

می توانیم تنظیم پارامتر را برای رگرسیون لجستیک نیز استفاده کنیم.

تابع هزینه در رگرسیون لجستیک به شکل زیر است :

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))]$$

تابع هزینه تنظیم شده به شکل زیر در می آید :

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))] + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2$$