CIS 419/519: Homework 2

{Yupeng Li}

Although the solutions are entirely my own, I consulted with the following people and sources while working on this homework: {http://www.cs.utep.edu/vladik/cs5315.13/cs5315_13kader.pdf}

1 Gradient Descent

Let k be a counter for the iterations of gradient descent, and let α_k be the learning rate for the k^{th} step of gradient descent.

a.) In one sentence, what are the implications of using a constant value for $\alpha_R k$ in gradient descent?
The α_k used is already the optimum value which will help converge θ_j to the minimum at quick enough speed
b.) In another sentence, what are the implications for setting α_k as a function of k?
We have not yet decided the optimum α_k for convergence, thus we are checking if α_k is too big or too small

2 Linear Regression

Since as we know the defined X has a Gaussian with mean 0 and variance, we can confidently assume

$$\frac{\partial}{\partial \theta}\mathcal{L}(\theta) = 0$$

Thus we can safely use the closed form solution:

$$\theta = (X^T X)^{-1} X^T y$$

Using this θ we can get from the following equation:

$$h_\theta(x_i) = x_i^T \theta$$

So that:

$$h_\theta(x) = \sum_{i=1}^{n} x_i^T \theta$$

$$h_\theta(x) = \sum_{i=1}^{n} x_i^T (X^T X)^{-1} X^T y_i$$

$$h_\theta(x) = x^T (X^T X)^{-1} X^T y$$

Where as we can see, X$^T X)^{-1} X^T isalinearfunctionof(X;x)$
So that we can express that :

$$h_\theta(x) = x^T (X^T X)^{-1} X^T y$$

in the other way:

$$f(x) = \sum_{i=1}^{n} l_i(x; X) y_i$$

, where

$$l_i(x; X) y_i = x_i^T (X^T X)^{-1} X^T y_i$$

Obviously,

$$(X^T X)^{-1} X^T$$

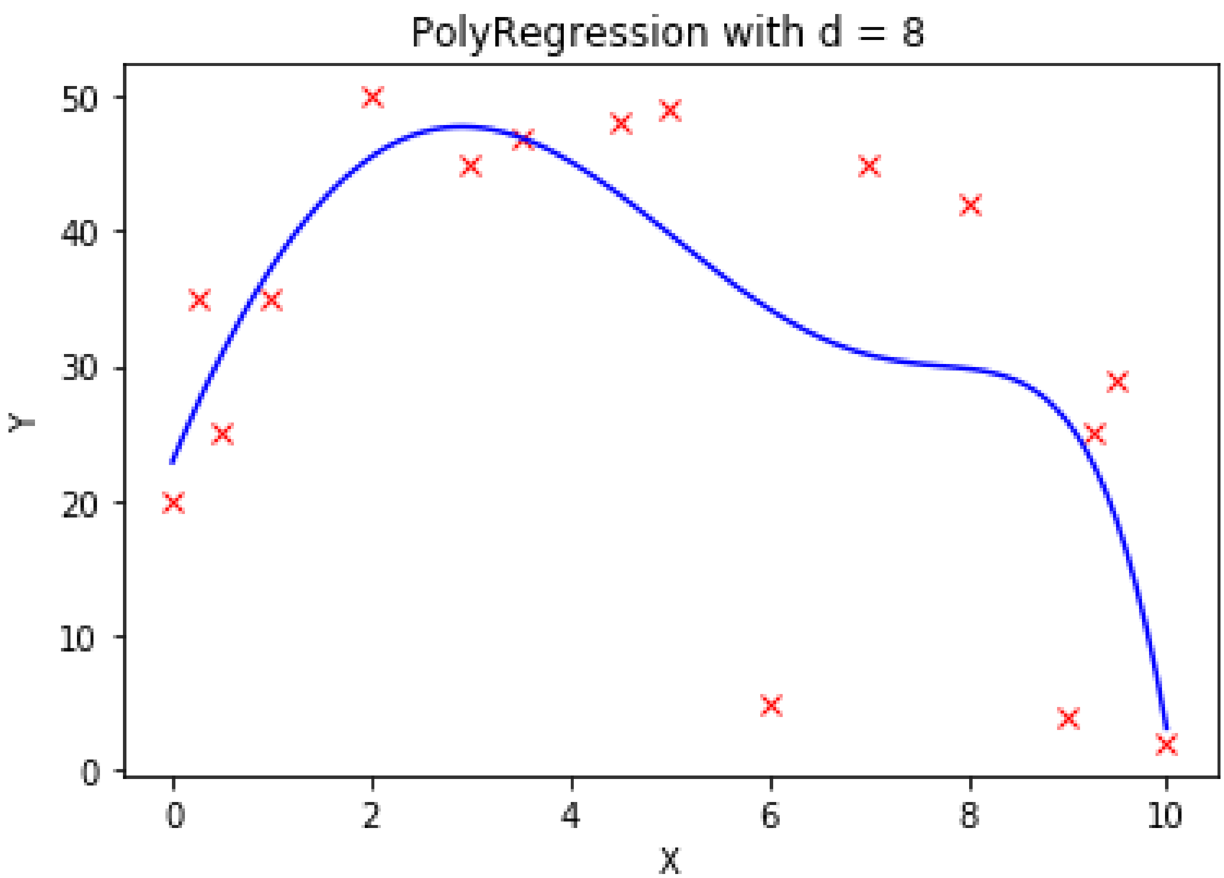
is a a linear function that does not depend on y_i

3 Polynomial Regression

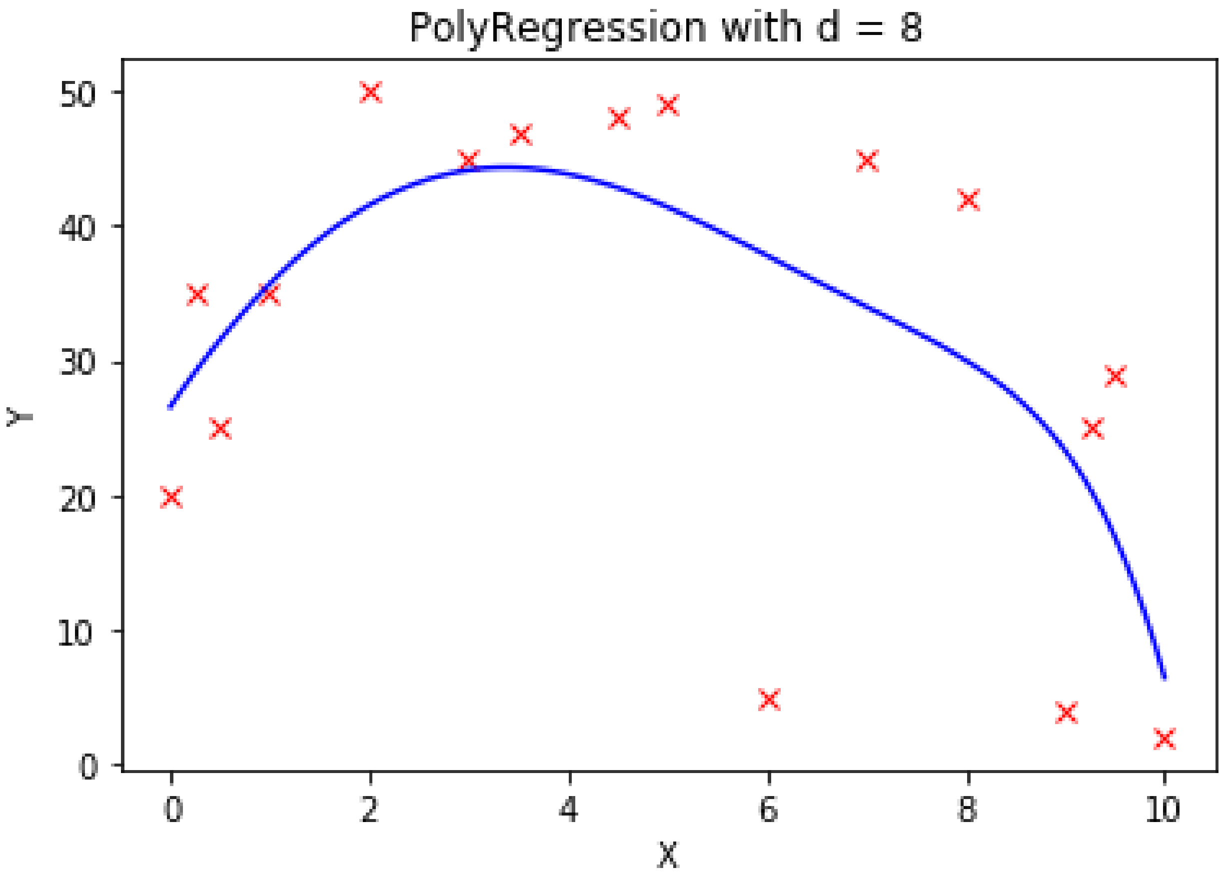
3.1 Implementating Polynomial Regression

This section is submitted in a separate .py file.

3.2 Choosing the optimal λ



Unregularized graph with $\lambda = 0$



Regularized graph with $\lambda = 0.01$

One thing interesting is that, using the α given such that $\alpha = 0.25$, the polynomial fitting does not converge when $\lambda is 0$.

Because of that, I tuned α to 0.01 so that the graph both converges at $\lambda = 0$ and λ greater than 0.

The graphs are provided as above.