Honor Pledge

“On my honor, I have neither given nor received any unauthorized aid on this assignment.”

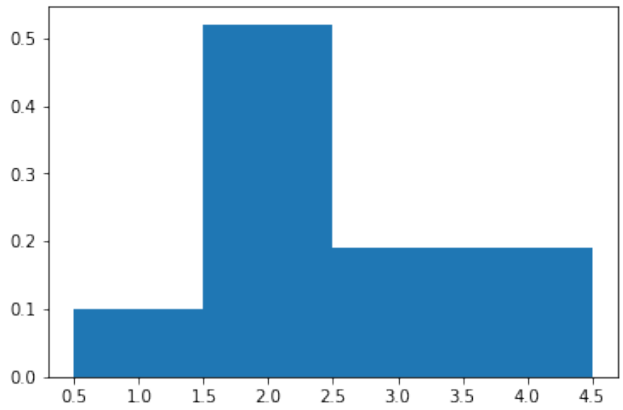
Peng Yang

Yue He

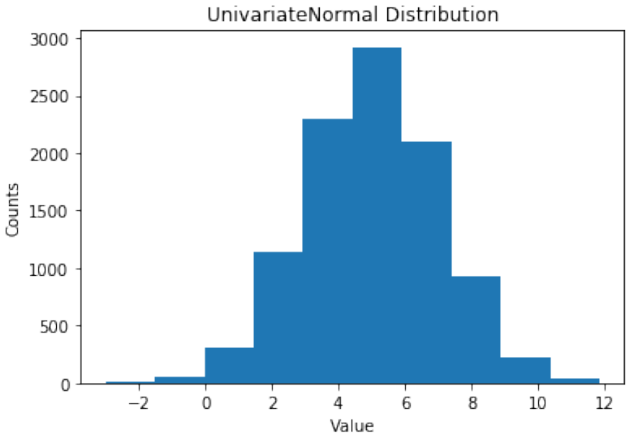
Jan. 19th 2019

Problem 0: Background refresher

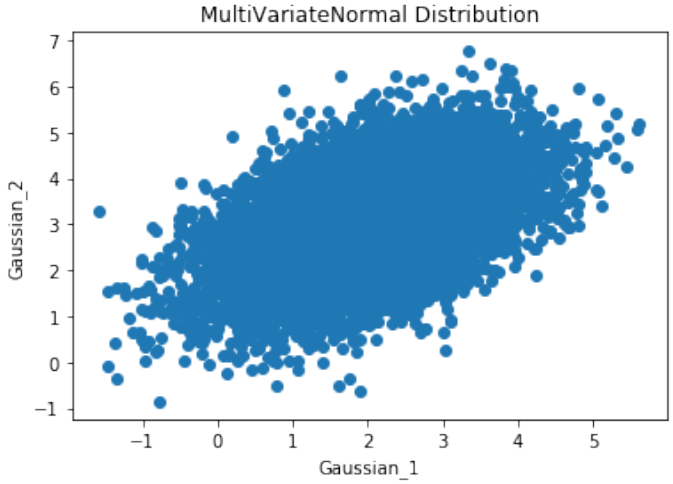
0.1 categorical distribution



0.2 univariate normal distribution



0.3 multivariate normal distribution



0.4 mixture distribution

The probability is around 0.18 (each time the answer is different)

Problem 0.

2. $X \sim \text{possion}(a)$. $Y \sim \text{possion}(b)$.

$\Rightarrow M_X(t) = \exp(a(e^{it}-1))$

$M_Y(t) = \exp(b(e^{it}-1))$

$M_{X+Y}(t) = M_X(t) M_Y(t) = \exp((a+b)(e^{it}-1))$

$\Rightarrow X+Y \sim \text{possion}(a+b)$.

3. $P(X_0 = x_0) = \alpha_0 e^{-\frac{(x_0-\mu_0)^2}{2\sigma_0^2}}$

$P(X_1 = x_1 \mid X_0 = x_0) = \alpha e^{-\frac{(x_1-x_0)^2}{2\sigma^2}}$

$\Rightarrow P(X_1 = x_1, X_0 = x_0) = P(X_1 = x_1 \mid X_0 = x_0) \cdot P(X_0 = x_0)$

$= \alpha\alpha_0 \exp\left\{-\frac{\sigma^2(x_0-\mu_0)^2 + \sigma_0^2(x_1-x_0)^2}{2\sigma_0^2\sigma^2}\right\}$.

$= \alpha\alpha_0 \exp\left\{-\frac{x_0^2 - 2\frac{\sigma^2\mu_0+\sigma_0^2x_1}{\sigma^2+\sigma_0^2}x_0 + \frac{\sigma^2\mu_0^2+\sigma_0^2x_1^2}{\sigma^2+\sigma_0^2}}{2\sigma_0^2\sigma^2/(\sigma_0^2+\sigma^2)}\right\}$.

$= \alpha\alpha_0 \exp\left\{-\frac{\left(x_0 - \frac{\sigma^2\mu_0+\sigma_0^2x_1}{\sigma^2+\sigma_0^2}\right)^2 + \frac{\sigma^2\mu_0+\sigma_0^2x_1}{\sigma^2+\sigma_0^2} - \left(\frac{\sigma^2\mu_0+\sigma_0^2x_1}{\sigma^2+\sigma_0^2}\right)^2}{2\sigma_0^2\sigma^2/(\sigma_0^2+\sigma^2)}\right\}$.

$\Rightarrow P(X_1 = x_1) = \int_{\mathbb{R}} P(X_1 = x_1, X_0 = x_0)\, dx_0$.

$= \alpha\alpha_0\sqrt{2\pi\frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2}} \exp\left\{-\frac{\frac{\sigma^2\mu_0+\sigma_0^2x_1}{\sigma^2+\sigma_0^2} - \left(\frac{\sigma^2\mu_0+\sigma_0^2x_1}{\sigma^2+\sigma_0^2}\right)^2}{2\sigma_0^2\sigma^2/(\sigma^2+\sigma_0^2)}\right\}$.

$\textcircled{1}$

$\textcircled{1} = \frac{(\sigma^2\mu_0^2+\sigma_0^2x_1^2)(\sigma^2+\sigma_0^2) - (\sigma^2\mu_0+\sigma_0^2x_1)^2}{\sigma_0^2\sigma^2(\sigma^2+\sigma_0^2)}$

$= \frac{\sigma_0^2\sigma^2(x_1-\mu_0)^2}{\sigma_0^2\sigma^2(\sigma^2+\sigma_0^2)} = \frac{(x_1-\mu_0)^2}{\sigma^2+\sigma_0^2}$

$\Rightarrow P(X_1 = x_1) = \alpha\alpha_0\sqrt{2\pi\frac{\sigma_0^2\sigma^2}{\sigma^2+\sigma_0^2}} \exp\left\{-\frac{(x_1-\mu_0)^2}{2(\sigma^2+\sigma_0^2)}\right\}$.

$\Rightarrow$. $\alpha_1 = \alpha\alpha_0\sqrt{2\pi\frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2}}$ $\mu_1=\mu_0$ $\sigma_1=\sqrt{\sigma^2+\sigma_0^2}$.

3. $A=\begin{pmatrix}0 & 1\\ -2 & -3\end{pmatrix}$ $|A-\lambda E| = \begin{vmatrix}-\lambda & 1\\ -2 & -3-\lambda\end{vmatrix} = (3+\lambda)\lambda+2=0$.

$\Rightarrow$. $\lambda_1=-1$ $\lambda_2=-2$.

when $\lambda_1=-1$

$A+E=\begin{pmatrix}1 & 1\\ -2 & -2\end{pmatrix}\sim\begin{pmatrix}1 & 1\\ 0 & 0\end{pmatrix}$ $\Rightarrow$. $P_1=\begin{pmatrix}1\\ -1\end{pmatrix}$

when $\lambda_2=-2$.

$A+2E=\begin{pmatrix}2 & 1\\ -2 & -1\end{pmatrix}\sim\begin{pmatrix}2 & 1\\ 0 & 0\end{pmatrix}$ $\Rightarrow$. $P_2=\begin{pmatrix}1\\ -2\end{pmatrix}$

4. - $(A+B)^2 \neq A^2+2AB+B^2$

$A=\begin{pmatrix}1 & 1\\ 0 & 0\end{pmatrix}$ $B=\begin{pmatrix}0 & 1\\ 0 & 1\end{pmatrix}$

$(A+B)^2=\begin{pmatrix}1 & 4\\ 0 & 1\end{pmatrix}$

$A^2+2AB+B^2=\begin{pmatrix}1 & 6\\ 0 & 1\end{pmatrix}$

- $AB=0$, $A\neq 0$ $B\neq 0$.

$A=\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}$ $B=\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix}$ $\Rightarrow$. $AB=0$.

5. $A^T=[I-2\mu\mu^T]^T=I-2(\mu\mu^T)^T=I-2\mu\mu^T$

$$A^TA = (I - 2uu^T)(I - 2uu^T)$$

$$= I - 2uu^T - 2uu^T + 4u(u^Tu)u^T = I.$$

6. (1). $f(x) = e^x \Rightarrow f'(x) = f''(x) = e^x > 0, \ \forall x \in \mathbb{R}$.

$\Rightarrow$. $f(x)$ is a convex for $x \in \mathbb{R}$.

(2). $f(x_1, x_2) = \max(x_1, x_2) \quad \lambda \in (0,1), \ x, y \in \mathbb{R}$.

$$f(\lambda x + (1-\lambda)y) = \max_i(\lambda x_i + (1-\lambda)y_i)$$

$$\leq \lambda \max_i x_i + (1-\lambda) \max_i y_i = \lambda f(x) + (1-\lambda) f(y)$$

(3). f, g are convex on S, then $\max(f, g)$ is convex

let $h(x) = \max(f(x), g(x)) \quad \lambda \in (0,1)$.

$$h(\lambda x_1 + (1-\lambda)x_2) = \max\{f(\lambda x_1 + (1-\lambda)x_2), g(\lambda x_1 + (1-\lambda)x_2)\}.$$

$$\leq \max\{\lambda f(x_1) + (1-\lambda) f(x_2), \lambda g(x_1) + (1-\lambda) g(x_2)\}.$$

$$= \max\{\lambda f(x_1), \lambda g(x_1)\} + \max\{(1-\lambda) f(x_2), (1-\lambda) g(x_2)\}.$$

$$= \lambda h(x_1) + (1-\lambda) h(x_2).$$

$\Rightarrow$. $\max(f, g)$ is convex on S.

(4). let $h(x) = f(x) \cdot g(x)$.

$$h'(x) = f'(x) g(x) + f(x) g'(x).$$

$$h''(x) = f''(x) g(x) + 2 f'(x) g'(x) + f(x) g''(x).$$

Since f, g are convex and non-negative.

$\Rightarrow$. $h''(x) \geq 0$. $\Rightarrow$. $h(x)$ is a convex.

7. $H(p) = -\sum_{i=1}^{k} p_i \log(p_i)$ $\quad g = \sum_{i=1}^{k} p_i = 1$.

let $\frac{\partial}{\partial p_i}(H(p) + \lambda(g-1)) = 0$.

$\Rightarrow \frac{\partial}{\partial p_i}(-\sum_{i=1}^{k} p_i \log(p_i) + \lambda(\sum_{i=1}^{k} p_i - 1)) = 0$.

$\Rightarrow -(\log p_i + 1) + \lambda = 0$.

the we know that $p_1 = p_2 = \cdots = p_k$.

Since $\sum_{i=1}^{k} p_i = 1$. $\Rightarrow p_i = \frac{1}{k}$

$X \sim MN(n, p)$. $f_X = \frac{n!}{x_1! x_2! \cdots x_n!}(\frac{1}{k})^n$.

$\sum x_i = n$.

Problem 2.

1. $J(\theta) = \frac{1}{2}\sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)} - y^{(i)})^2$

$\Rightarrow J(\theta) = (X\theta - y)^T W (X\theta - y)$

Since W is a diagonal matrix.

$\Rightarrow$

$$(X_{m\times d}\,\theta_{d\times 1} - y_{m\times 1})^T_{1\times m} \begin{pmatrix} \frac{w^{(1)}}{2} & & & \\ & \frac{w^{(2)}}{2} & & \\ & & \ddots & \\ & & & \frac{w^{(m)}}{2} \end{pmatrix}_{m\times m} (X_{m\times d}\,\theta_{d\times 1} - y_{m\times 1})_{m\times 1}$$

$= \frac{1}{2}\sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)} - y^{(i)})^2$.

2. $J(\theta) = (X\theta - y)^T W (X\theta - y)$.

$= (\theta^T X^T - y^T) W (X\theta - y)$

$= \theta^T X^T W X\theta - \theta^T X^T W y - y^T W X\theta + y^T W y$.

Since $\theta^T X^T W y$ and $y^T W X\theta$ are 1×1 scalar.

$\Rightarrow \theta^T X^T W y = y^T W X\theta$

$\Rightarrow J(\theta) = \theta^T X^T W X \theta - 2 y^T W X \theta + y^T W y.$

$\frac{\partial J(\theta)}{\partial \theta} = 2 X^T W X \theta - 2 X^T W y$. let $\frac{J(\theta)}{\partial \theta} = 0$.

$\Rightarrow \hat{\theta} = (X^T W X)^{-1} X^T W y.$

3. $J(\theta) = \frac{1}{2} \sum_{i=1}^{m} w^{(i)} (\theta^T x^{(i)} - y^{(i)})^2$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \begin{cases} \sum_{i=1}^{m} w^{(i)} (h_\theta(x^{(i)}) - y^{(i)}) & j = 0. \\ \sum_{i=1}^{m} w^{(i)} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} & j \geq 1. \end{cases}$$

Algorithm:

① Input X.

② Calculat W.

③ Initize θ

④ for loop:

$\theta_j \leftarrow \theta_j - \alpha \sum_{i=1}^{m} w^{(i)} (h_\theta(x^{(i)}) - y^{(i)})$. $j = 0$.

$\theta_j \leftarrow \theta_j - \alpha \sum_{i=1}^{m} w^{(i)} (h_\theta(x^{(i)}) - y^{(i)})$ $j \geq 1$.

⑤ predict.

$y_{pred} = \hat{\theta}^T X.$

Locally weighted linear regression is a non-parametric method, since the estimation rely on data only without assumption of distribution.

Problem 2.

Since we assume

$$y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}. \qquad \varepsilon^{(i)} \overset{iid}{\sim} N(0, \sigma^2)$$

$$\Rightarrow \quad y^{(i)} \sim N(\theta^T x^{(i)}, \sigma^2).$$

$$E y^{(i)} = \theta^T x^{(i)}. \qquad \mathrm{Var}\, y^{(i)} = \sigma^2.$$

Besides, we know $\hat{\theta} = (X^T X)^{-1} X^T Y$.

a. $E\hat{\theta} = (X^T X)^{-1} X^T EY = (X^T X)^{-1}(X^T X)\theta = \theta.$

Then we know $\hat{\theta}$ is an unbaised estimator of θ.

b. $\mathrm{Var}\,\hat{\theta} = (X^T X)^{-1} X^T \mathrm{Var}\, Y \cdot X[(X^T X^{-1})]^T$

$= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = (X^T X)^{-1} \cdot \sigma^2.$

Problem 3.1: Implementing linear regression

Problem 3.1.A: Linear regression with one variable

Plotting the data

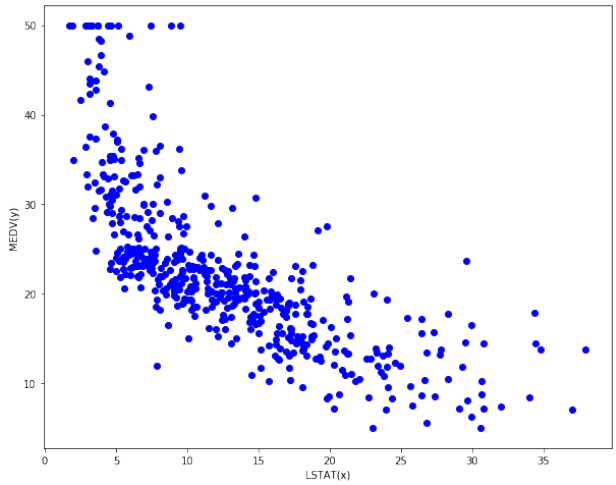


Figure 1: Scatter plot of training data

Problem 3.1.A1: Computing the cost function $J(\theta)$

Problem 3.1.A2: Implementing gradient descent

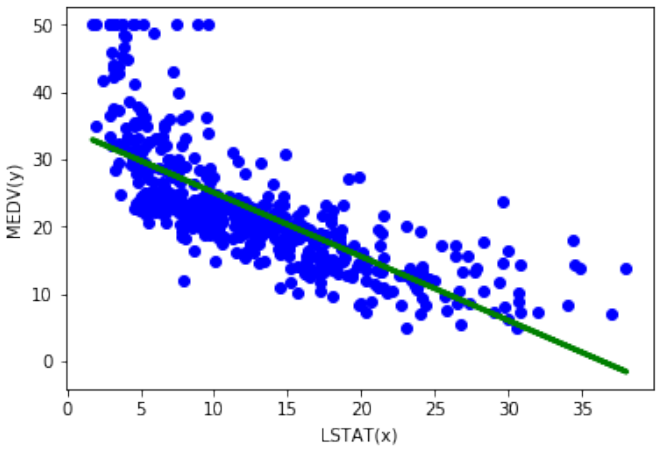


Figure 2: Fitting a linear model to the data in Figure 1

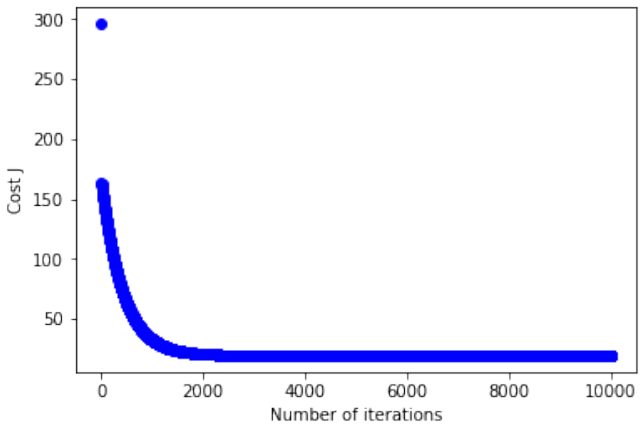


Figure 3: Convergence of gradient descent to _t the linear model in Figure 2

Problem 3.1.A3: Predicting on unseen data

```
For lower status percentage = 5, we predict a median home value of 29.80
For lower status percentage = 50, we predict a median home value of -12.95
```

Visualizing J(θ)

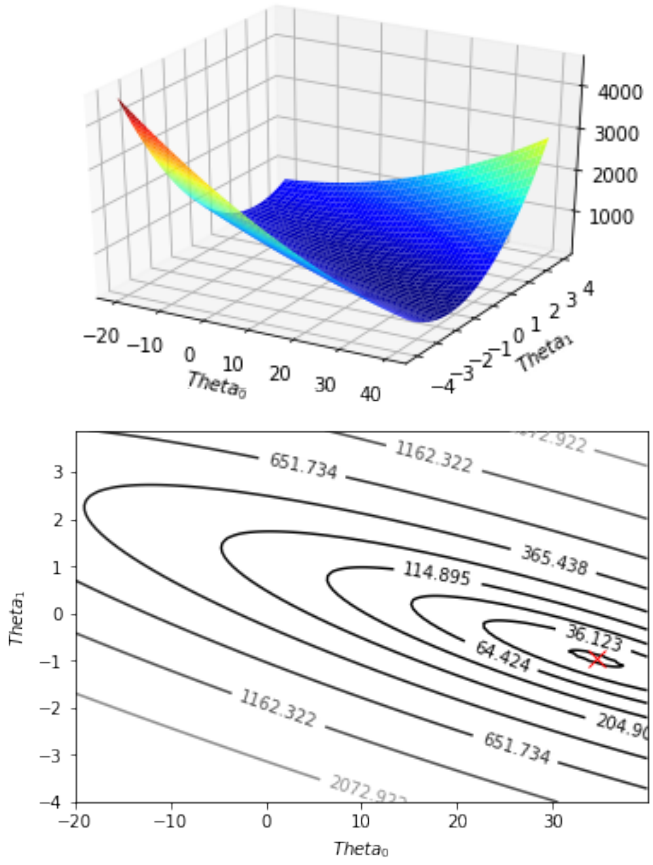


Figure 4: Surface and contour plot of cost function J (linear regression with single variable)

Problem 3.1.B: Linear regression with multiple variables

Problem 3.1.B1: Feature normalization

Problem 3.1.B2: Loss function and gradient descent

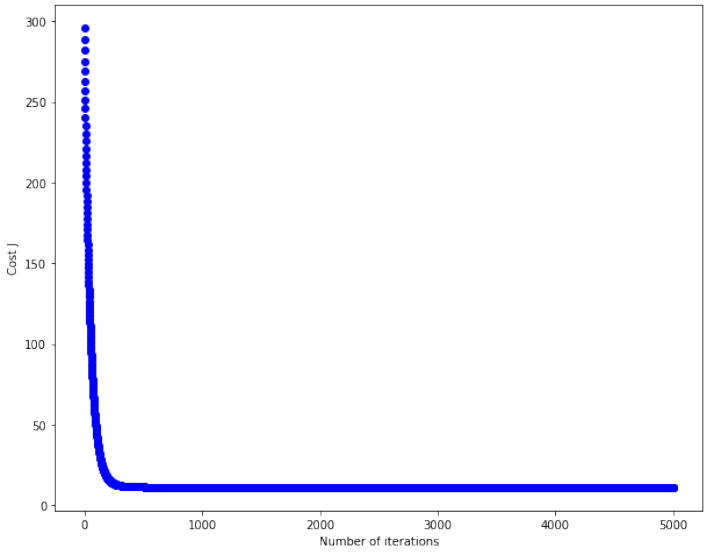


Figure 5-1: Convergence of gradient descent for linear regression with multiple variables (Boston housing data set)

Problem 3.1.B3: Making predictions on unseen data

```
For average home in Boston suburbs, we predict a median home value of 2253
28.06
```

Problem 3.1.B4: Normal equations

```
For average home in Boston suburbs, we predict a median home value of $225
328.06$.
```

Problem 3.1.B5: Exploring convergence of gradient descent

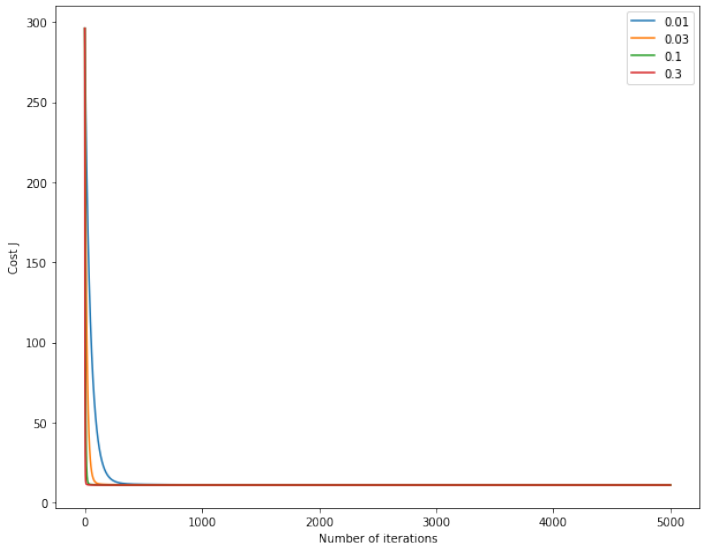


Figure 5-2: Convergence of gradient descent for linear regression with multiple variables (Boston housing data set) with different lambda

When learning rates of 0.1 and 0.3, the cost J drops too quickly and the converge process is not that obvious. So we prefer learning rate to be 0.01 and 0.03. The best we choose is learning rate of 0.01 with iteration of 400.

Problem 3.2: Implementing regularized linear regression

Visualizing the dataset

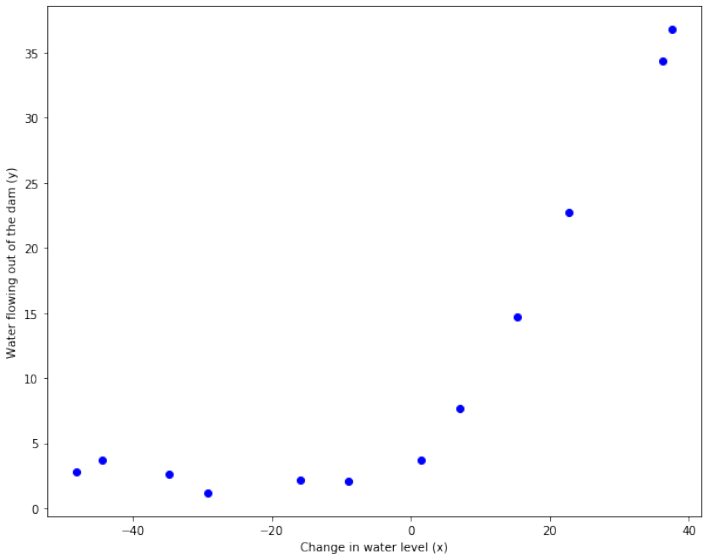


Figure 6: The training data for regularized linear regression

Problem 3.2.A1: Regularized linear regression cost function

Problem 3.2.A2: Gradient of the Regularized linear regression cost function

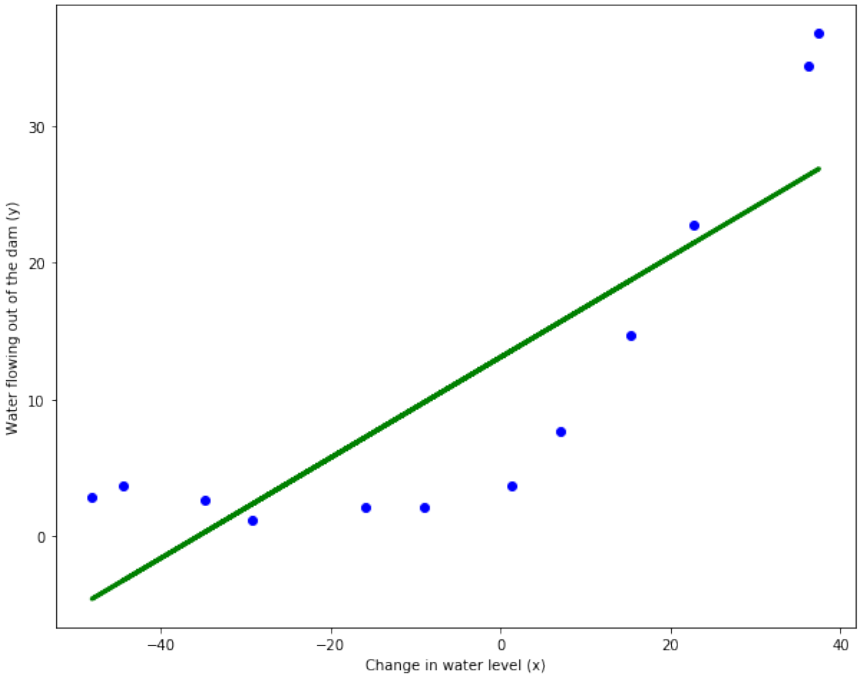


Figure 7: The best _t line for the training data

Problem 3.2.A3: Learning curves

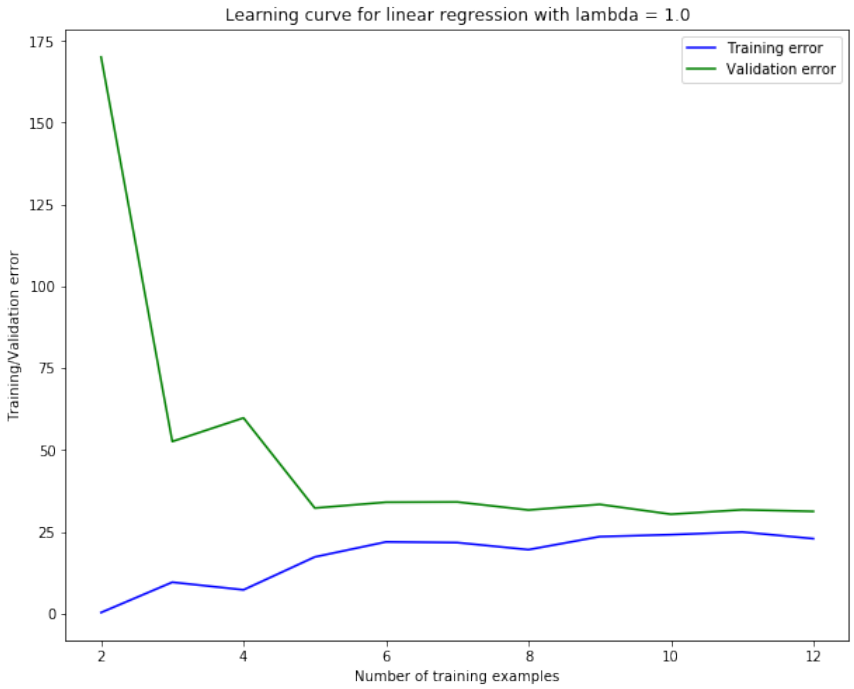


Figure 8: Learning curves

Learning polynomial regression models

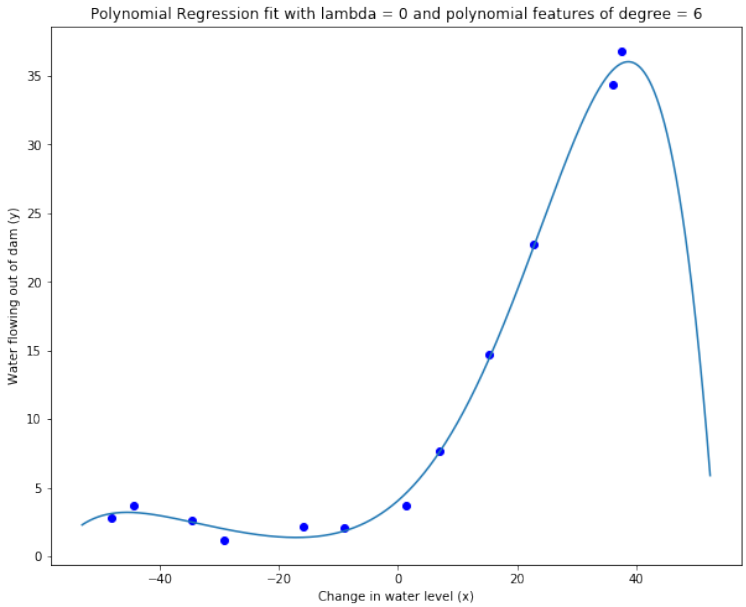


Figure 9: Polynomial _t for lambda = 0 with a p=6 order model.

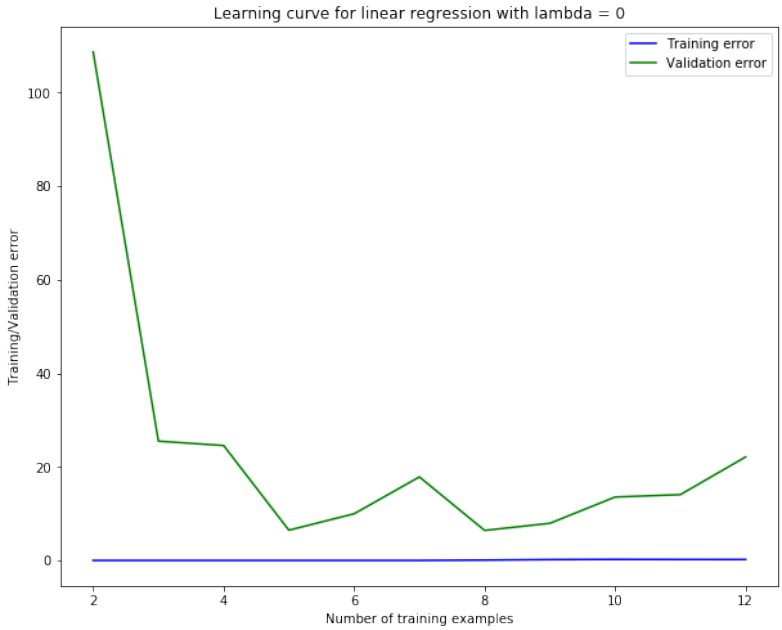
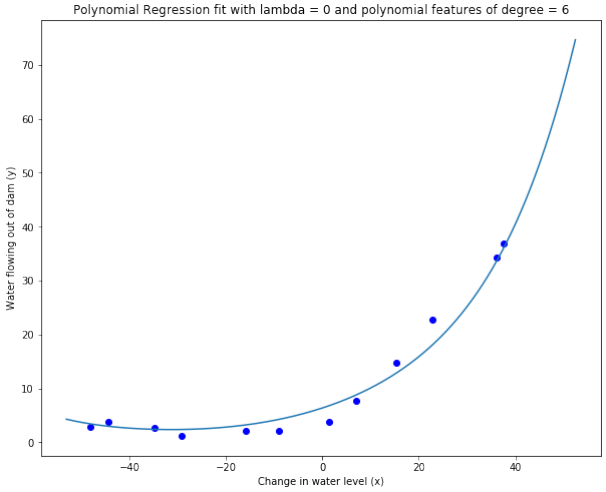
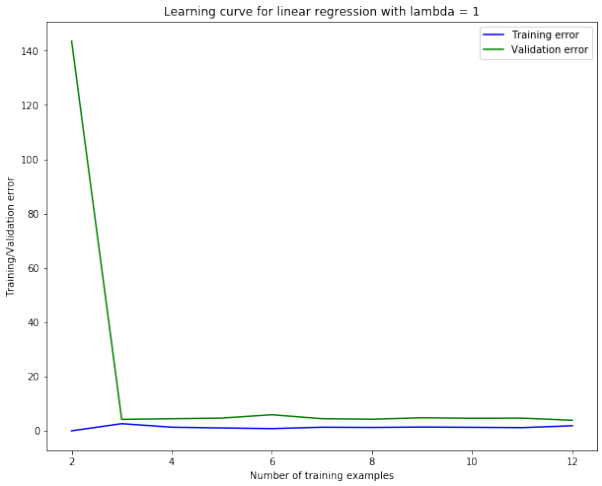
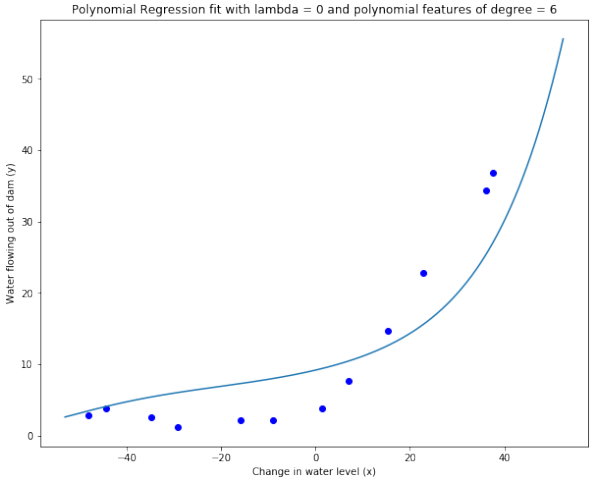


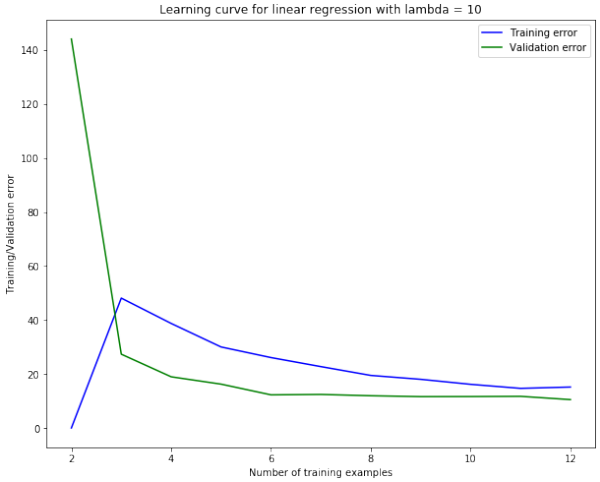
Figure 10: Learning curve for lambda = 0.

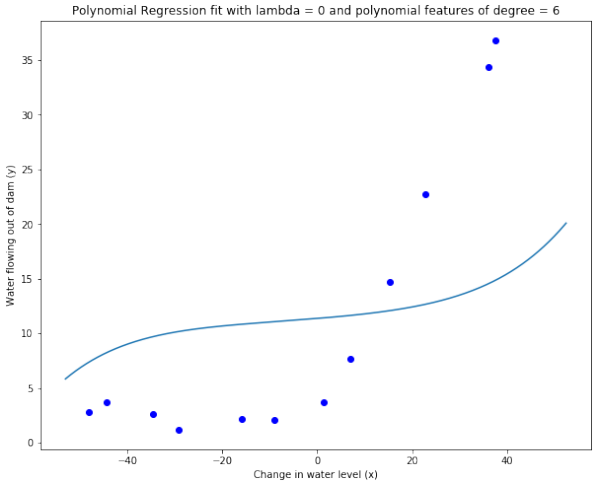
Problem 3.2.A4: Adjusting the regularization parameter

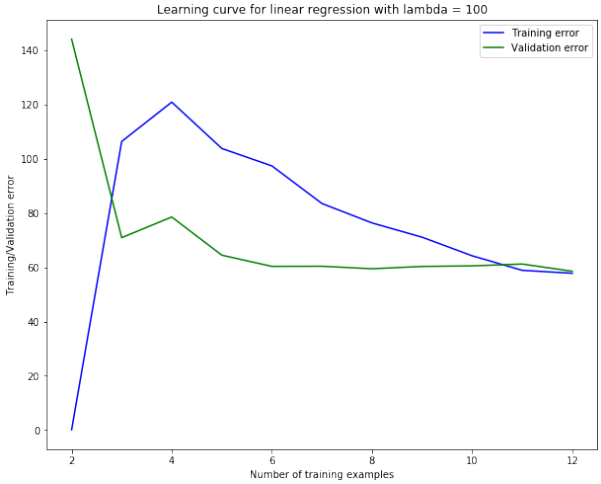












From the above pictures, we shall see that when lambda = 10 and lambda = 100, the model is under fitted; however, when lambda = 1, the model fits well.

It is reasonable, since lambda can control the coefficients of the model, which is θ, when lambda increase, θ goes smaller in order to minimize $J(\theta)$. Thus, we need to find the proper lambda for the model.

Problem 3.2.A5: Selecting lambda using a validation set

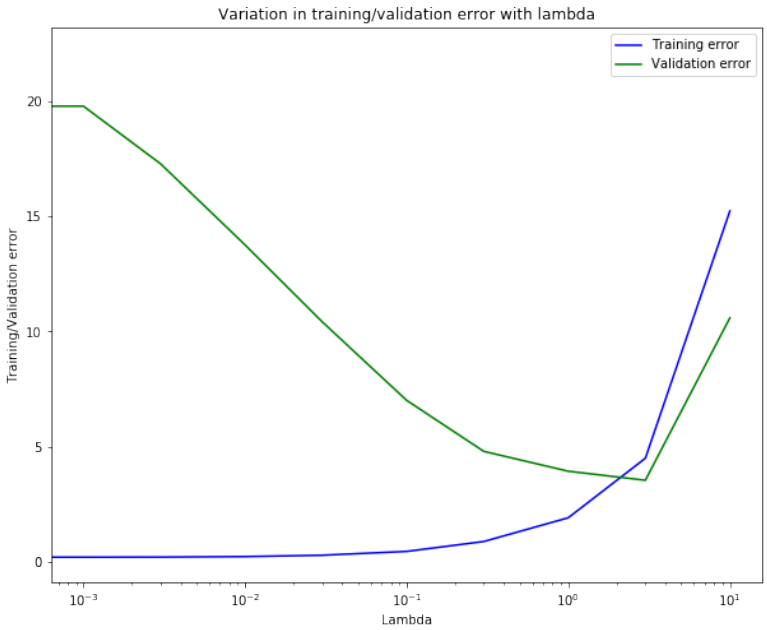


Figure 11: Averaged Learning curve for lambda = 1.

From the figure above, we shall see that when lambda = 3, validation error has the lowest value; Before 3, the validation error is high, while the training error is low; After 3, the training error is high, while the validation error is high. Thus, the best choice of lambda for this problem is 3, where validation error and training error both perform well.

Problem 3.2.A6: Computing test set error

```
Best error: 4.39762337668 when reg =  3.0
```

Problem 3.2.A7: Plotting learning curves with randomly selected examples

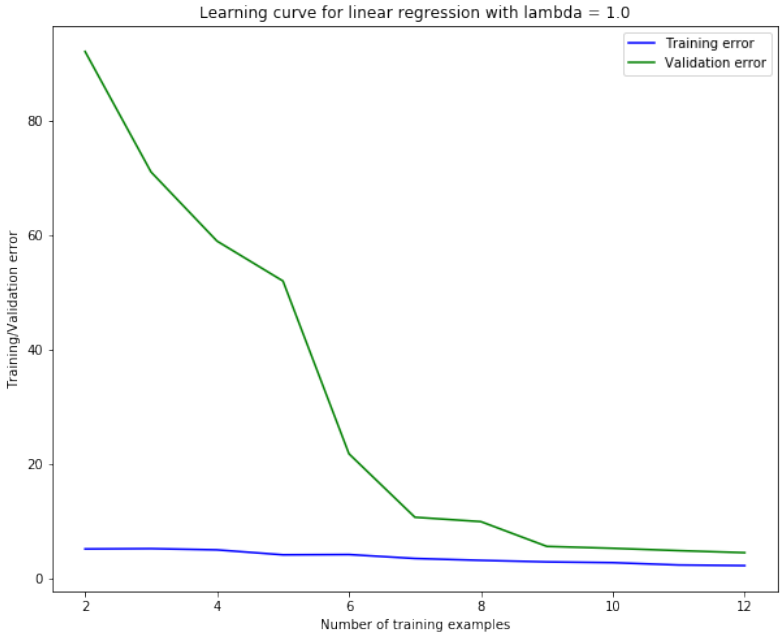


Figure 12: Learning curve for lambda = 1.