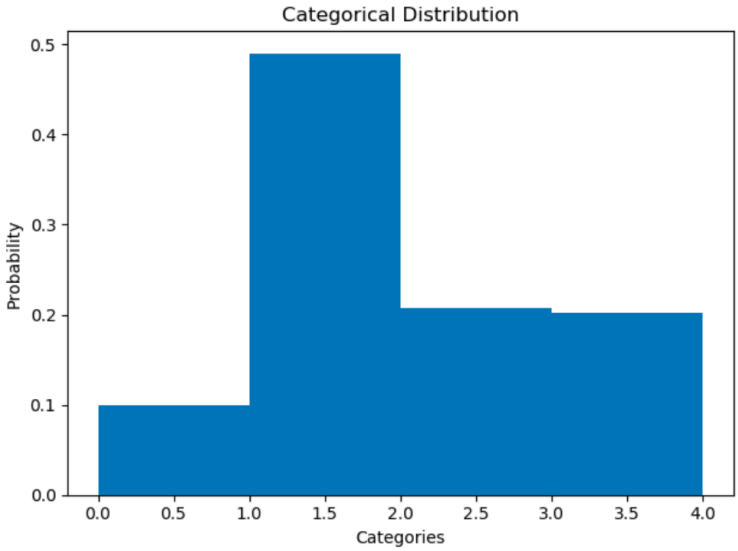
HW1

Zheyu Ding (zd12)
M.Viginesh Vaibhav (vm24)

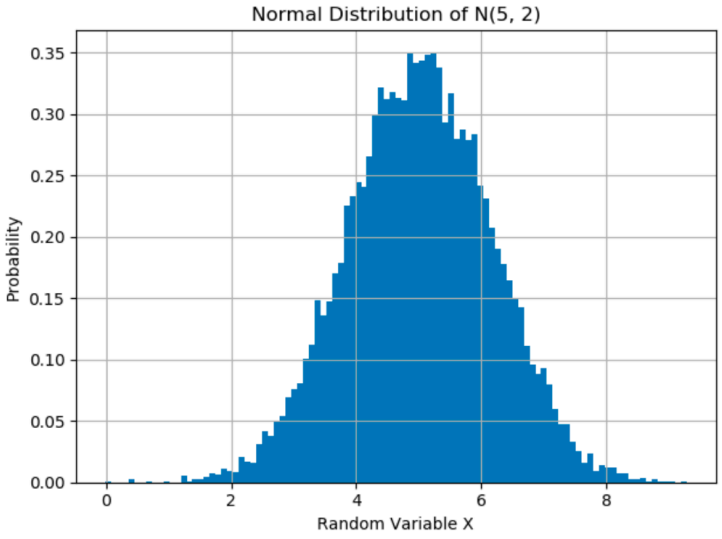
Problem 0: Background refresher

1. Plot 4 kinds of distribution

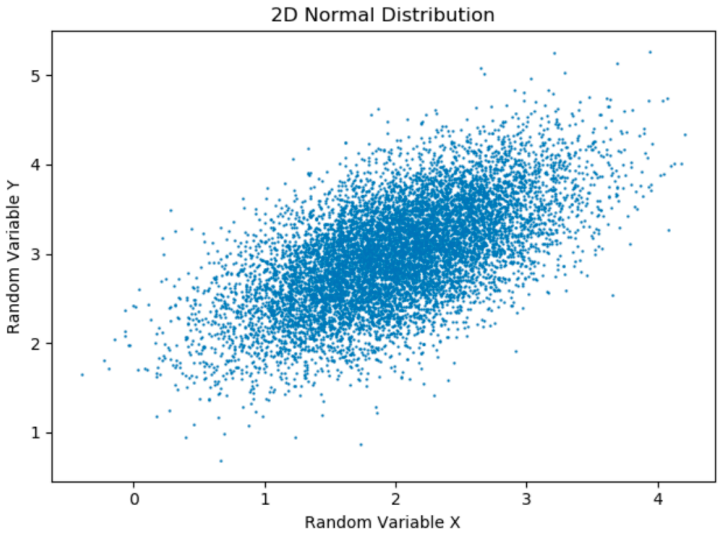
a) Plot the histogram of samples generated by a categorical distribution with probabilities [0.1,0.5,0.2,0.2]



b) Plot the univariate normal distribution with mean of 5 and standard deviation of 2.



c) Produce a scatter plot of the samples for a 2-D Gaussian with mean at [2,3] and a covariance matrix [[1,0.5], [0.2,1]].



d) Test your mixture sampling code by writing a function that implements an equal-weighted mixture of four Gaussians in 2 dimensions, centered at (±1, ±1) and having covariance I. Estimate the probability that a sample from this distribution lies within the unit circle centered at (0.1, 0.2) and include that number in your writeup.

The probability is 0.1689.

2. Prove that the sum of two independent Poisson random variables is also a Poisson random variable.

Solution

There are two random variables: X~P(λ), Y~P(μ)

So, $P(X=x)=\frac{\lambda^x}{x!}\,e^{-\lambda}$, $P(Y=y)=\frac{\lambda^y}{y!}e^{-\lambda}$, U=X+Y

$$P(U=u)=P(X+Y=u)=\sum_{x=0}^{u}P(X=x)\cdot P(Y=u-x)$$

$$=\sum_{x=0}^{u}\frac{\lambda^x}{x!}e^{-\lambda}\cdot\frac{\mu^{u-x}}{(u-x)!}e^{-\mu}=\sum_{x=0}^{u}\frac{\mu^{u-x}\cdot\lambda^x}{x!\,(u-x)!}e^{-(\mu+\lambda)}$$

$$=\sum_{x=0}^{u}\frac{C_u^x\mu^{u-x}\cdot\lambda^x}{u!}e^{-(\mu+\lambda)}=\frac{e^{-(\mu+\lambda)}}{u!}\sum_{x=0}^{u}C_u^x\mu^{u-x}\cdot\lambda^x$$

$$=\frac{e^{-(\mu+\lambda)}}{u!}\cdot(\mu+\lambda)^u$$

That is to say $U\sim P(\mu+\lambda)$.

3. Let X_0 and X_1 be continuous random variables. Show that if p ($X_0 = x_0$) P ($X_1 = x_1 | X_0 = x_0$) there exists α1, μ1 and σ1 such that. Write down expressions for these quantities in terms of α0, α, μ0, σ0 and σ.

$$p(X_0 = x_0) = \alpha_0 e^{-\frac{(x_0-\mu_0)^2}{2\sigma_0^2}}$$
$$p(X_1 = x_1|X_0 = x_0) = \alpha e^{-\frac{(x_1-x_0)^2}{2\sigma^2}}$$

there exists α1, μ1 and σ1 such that

$$p(X_1 = x_1) = \alpha_1 e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}}$$

Write down expressions for these quantities in terms of α0, α, μ0, σ0 and σ.

Solution

$$P(X_1 = x_1) = \int P(X_1 = x1|X_0 = x_0)P(X_0 = x_0)dx_0$$

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

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

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

Because $\int e^{-ax^2+bx+c}\,dx = \sqrt{\frac{\pi}{a}}e^{\frac{b^2}{4a}+c}$

$$original = \alpha_0\alpha\sqrt{\frac{2\sigma^2\sigma_0^2\pi}{(\sigma^2+\sigma_0^2)}}\exp\left(\frac{(2\mu_0\sigma^2+2x_1\sigma_0^2)^2}{8(\sigma^2+\sigma_0^2)/(\sigma^2\sigma_0^2)} - \mu_0^2\sigma^2 - x_1\sigma_0^2\right)$$

$$= \alpha_0\alpha\sqrt{\frac{2\sigma^2\sigma_0^2\pi}{(\sigma^2+\sigma_0^2)}}\exp\left(\frac{(2\mu_0\sigma^2+2x_1\sigma_0^2)^2}{8(\sigma^2+\sigma_0^2)/(\sigma^2\sigma_0^2)} - \mu_0^2\sigma^2 - x_1\sigma_0^2\right)$$

$$= \alpha_0\alpha\sqrt{\frac{2\pi\sigma^2\sigma_0^2}{(\sigma^2+\sigma_0^2)}}\exp\left(\frac{-(x_1-\mu_0)^2}{2(\sigma^2+\sigma_0^2)}\right) = \alpha_1 e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}}$$

So,

$$\begin{cases} \sigma_1 = \sqrt{\sigma^2 + \sigma_0^2} \\ \mu_1 = \mu_0 \\ \alpha_1 = \alpha_0 \alpha \sqrt{\dfrac{2\pi\sigma^2\sigma_0^2}{(\sigma^2 + \sigma_0^2)}} \end{cases}$$

4. Find the eigenvalues and eigenvectors of the following 2 × 2 matrix A.
Solution

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

$$= \lambda^2 + 3\lambda + 2 = (\lambda + 1)(\lambda + 2)$$

So, the eigenvalues of A are $\lambda_1 = -1,\ \lambda_2 = -2$.

When $\lambda_1 = -1,\ the$ eigenvector of A can be $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$.

When $\lambda_2 = -2,\ the$ eigenvector of A can be $\begin{pmatrix} -1 \\ 2 \end{pmatrix}$.

5. Provide one example for each of the following cases, where A and B are 2 × 2 matrices.
Solution

1) $(A+B)^2 \neq A^2 + B^2$+2AB

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

$$(A+B)^2 = \left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right)^2 = 0 \neq A^2 + B^2 + 2AB = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

2) AB=0, A≠0, B≠0

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

$$\text{AB} = 0$$

6. Let u denote a real vector normalized to unit length. That is, $uu^T = 1$. Show that

$A = I - 2uu^T$ is orthogonal, i.e., $AA^T = 1$.

Solution

$$u^T u = E \leftrightarrow uu^T = E$$

$$AA^T = (I - 2uu^T)^T(I - 2uu^T) = (I - 2uu^T)(I - 2uu^T)$$

$$= I - 4uu^T + 4uu^T = I = 1$$

So, A is orthogonal.

7. A function f is convex on a given set S if and only if for $\lambda \in [0,1]$ and for all

x, y ∈ S, the following holds.

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

Moreover, a univariate function f(x) is convex on a set S if and only if its second derivative ″

f (x) is non-negative everywhere in the set. prove the following assertions.

Solution

1) $f(x) = e^x$ is convex for. X ∈ R.

$$f'(x) = e^x$$

$$f''(x) = e^x > 0 \quad X \in R$$

So, $f(x)$ is convex.

2) $f(x) = \max(x_1, x_2)$ is convex on R^2.

$$Suppose\ we\ have\ f(x) = \max(x_1, x_2), f(y) = \max(y_1, y_2)$$

$$left = f(\lambda x + (1-\lambda)y) = \max(\lambda x_1 + (1-\lambda)y), \lambda x_2 + (1-\lambda)y)$$

$$right = \lambda f(x) + (1-\lambda)f(y) = \lambda \max(x_1, x_2) + (1-\lambda) \cdot \max(y_1, y_2)$$

$$= \lambda x_{max} + (1-\lambda)y_{max}$$

$$left \leq right.$$

$So, f(x) = \max(x_1, x_2)$ is convex on R^2

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

$$Suppose\ h(x) = \max(f, g),$$

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \leq \lambda h(x) + (1-\lambda)h(y)$$
$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y) \leq \lambda h(x) + (1-\lambda)h(y)$$

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

$$\leq \lambda h(x_1) + (1-\lambda)h(x_2)$$

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

4) If univariate functions f and g are convex and non-negative on S, and have their minimum within S at the same point, then fg is convex on S.

$$Suppose\ h = fg,$$

$$h' = f'g + fg'$$

$$h'' = f''g + g''f + 2f'g',$$

$$because\ f\ and\ g\ are\ convex\ and\ non-negative\ on\ S, then$$

$$f''g + g''f > 0$$

$$because\ f\ and\ g\ have\ their\ minimum\ within\ S\ at\ the\ same\ point, then$$

$$2f'g' > 0$$

$$h'' > 0$$

Then we can imply fg is convex on S.

8. The entropy of a categorical distribution on K values is defined as

$$H(P) = -\sum_{i=1}^{K} p_i \log(p_i)$$

Using the method of Lagrange multipliers, find the categorical distribution that has the highest entropy.

Solution

$$g(P) = 1 \rightarrow \sum_{i=1}^{K} p_i = 1$$

$$\Phi(p_i, \lambda) = -\sum_{i=1}^{K} p_i \log(p_i) + \lambda\left(\sum_{i=1}^{K} p_i - 1\right)$$

$$\frac{\partial \Phi(p_i, \lambda)}{\partial p_i} = -(1 + logp_i) + \lambda = 0$$

$$\lambda = (1 + logp_i)$$

$$This\ prove\ that\ p_i\ is\ varible\ of \lambda, and \sum_{i=1}^{K} p_i = 1,$$

$$p_i = \frac{1}{n}$$

Therefore, the uniform distribution will get the highest entropy.

Problem 1: Locally weighted linear regression

1) The proof is attached in the pictures below. 'W' is a diagonal matrix of dimensions m x m, where each element of the diagonal is half of its corresponding w value, i.e., $W^{(i)} = w^{(i)}/2$.

Let, $X = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_d^1 \\ x_1^2 & x_2^2 & \cdots & x_d^2 \\ \vdots & & & \\ x_1^m & x_2^m & \cdots & x_d^m \end{bmatrix}$

$y = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^m \end{bmatrix}$ $\quad \theta = [\theta_1 \cdots \theta_d]$

$w = \begin{bmatrix} w^1 & 0 & 0 & \cdots & 0 \\ 0 & w^2 & 0 & \cdots & 0 \\ & & \ddots & & \\ 0 & 0 & \cdots & & w^m \end{bmatrix}$ ('w' is a diagonal matrix)

$$(X\theta - y) = \begin{bmatrix} (x_1^1\theta_1 + x_2^1\theta_2 + \cdots + x_d^1\theta_d) - y^1 \\ (x_1^2\theta_1 + x_2^2\theta_2 + \cdots + x_d^2\theta_d) - y^2 \\ \vdots \\ (x_1^m\theta_1 + x_2^m\theta_2 + \cdots + x_d^m\theta_d) - y^m \end{bmatrix}$$

$$(X\theta - y)^T w = [w^1((x_1^1\theta_1 + \cdots + x_d^1\theta_d) - y^1),\ w^2((x_1^2\theta_1 + \cdots + x_d^2\theta_d) - y^2),\ \ldots\ldots\ldots,\ w^m((x_1^m\theta_1 + \cdots + x_d^m\theta_d) - y^d)]$$

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

$$= w^1\left((x_1^1\theta_1 + \cdots + x_d^1\theta_d) - y^1\right)^2 + w^2\left((x_1^2\theta_1 + x_2^2\theta_2 + \cdots + x_d^2\theta_d) - y^2\right)^2 + \cdots + w^m\left((x_1^m\theta_1 + x_2^m\theta_2 + \cdots + x_d^m\theta_d) - y^m\right)^2$$

Here, if each element $w^1, w^2, \ldots w^M$ is half of its corresponding $w^{(i)}$ value, then we can re-write the above equation as,

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

Here, W is a diagonal Matrix of dimensions $m \times m$, where each diagonal element $W^{(i)}$ is half of its corresponding element $w^{(i)}$, i.e.,

$$W^{(i)} = w^{(i)}/2$$

2) The proof is attached in the picture below. The derived equation for θ is,

$$\theta = [X^T(W + W^T)]^{-1}X^T(W^T + W)y$$

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

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

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

Here, $y^T W X\theta = (W^T y)^T X\theta = [(X\theta)^T (W^T y)]^T$

$= [(X\theta)^T (W^T y)]$, Since it is a scalar

Similarly, $\theta^T X^T W X\theta = (X\theta)^T W X\theta$

$\therefore J(\theta) = [(X\theta)^T W X\theta - (X\theta)^T W^T y - (X\theta)^T W y + y^T W y]$

$\frac{\partial J}{\partial \theta} = 0$

$\Rightarrow (X^T W X)\theta + (X^T W^T X)\theta - X^T W^T y - X^T W y = 0$

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

$\Rightarrow \boxed{\theta = [X^T (W + W^T) X]^{-1} X^T (W^T + W) y}$

3) The algorithm for using batch gradient descent for locally weighted linear regression is as follows:
   a. Take the input vector x for the target to be predicted y
   b. Compute $w^{(i)}$ for every $x^{(i)}$ in the training data using x as the input vector, using the formula,

$$w^{(i)} = \exp\left(-\frac{\left(x - x^{(i)}\right)^T \left(x - x^{(i)}\right)}{2\tau^2}\right)$$

   Here, τ is the bandwidth defining the sphere of influence around x.
   c. Initialize theta to a matrix of dimensions d x 1, with all the elements as zero.
   d. Compute the cost as,

$$J(\theta) = \frac{1}{2}\sum w^{(i)}\left(\theta^T x^{(i)} - y^{(i)}\right)^2$$

   e. Compute the gradient as,

$$\nabla J(\theta) = \frac{1}{m} X^T (X\theta - y)$$

   f. Compute θ in the next step as,

$$\theta = \theta - \alpha \nabla J(\theta)$$

   Here, α is the learning rate.
   g. Repeat steps d to f over all of the training data, until J(θ) does not change, i.e., the value of θ converges.

   Locally weighted linear regression is non-parametric, because the values of the model's weights depends on the training dataset as well as the input of the test data.

Problem 2: Properties of the linear regression estimator

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

1) Show that $E[\theta] = \theta^*$ for the least squares estimator.
   Solution

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

$$E(\theta) = E[(X^T X^{-1}) X^T (X\theta^* + \epsilon)]$$

$$= E[\theta^* + (X^T X)^{-1} X^T \epsilon]$$

$$= \theta^* + (X^{-1} X)^{-1} X^T E(\epsilon)$$

$$because\ E(\epsilon) = 0$$

$$original = \theta^*$$

   So, $E[\theta] = \theta^*$ for the least squares estimator.

2) Show that the variance of the least squares estimator is $Var(\theta) = (X^T X)^{-1}\sigma^2$.
Solution

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

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

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

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

So, the variance of the least squares estimator is $Var(\theta) = (X^T X)^{-1}\sigma^2$.

Problem 3: Implementing linear regression and regularized linear regression

Problem 3.1: Implementing linear regression (45 points)

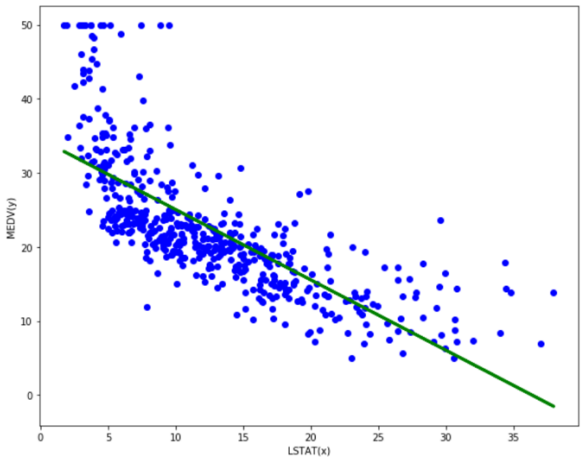
Problem 3.1.A: Linear regression with one variable (15 points)

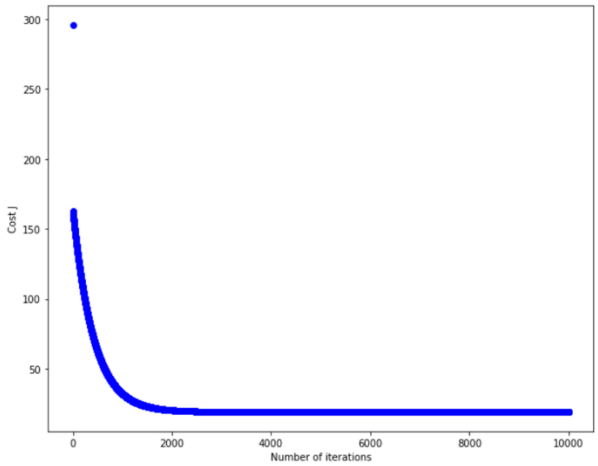
The predicted home values are as follows:

For **lower status percentage = 5**, we predict a median home value of **298034.49**

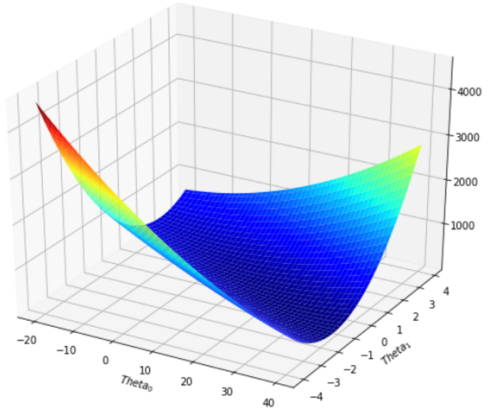
For **lower status percentage = 50**, we predict a median home value of **-129482.13**

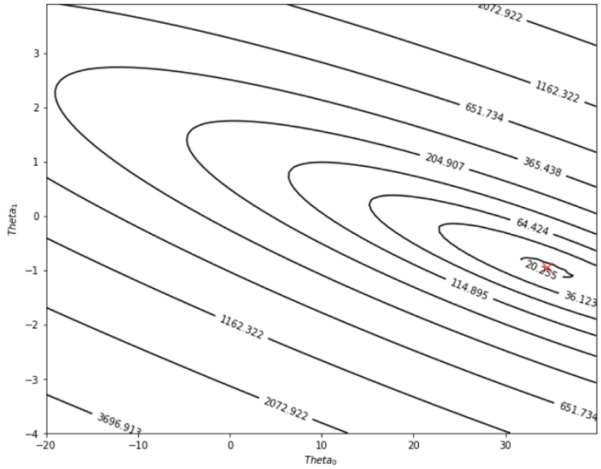
The plots produced for best fit curve and convergence of cost are as follows:





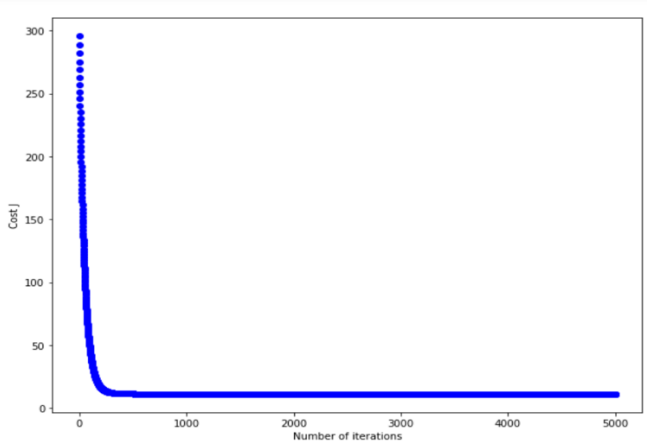
The visualization graphs of θ_0 and θ_1 are as follows:





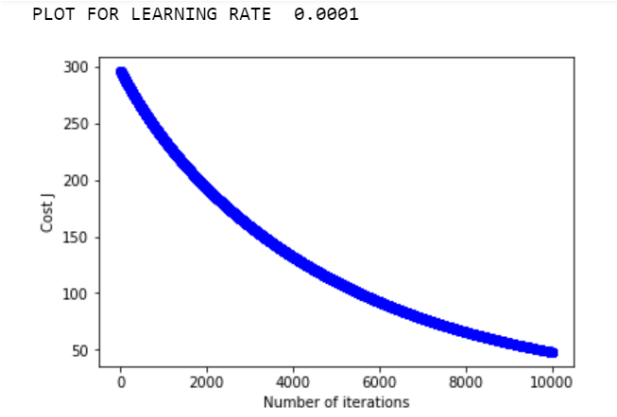
Problem 3.1.B: Linear regression with multiple variables

The plot showing the convergence of the cost with respect to the number of iterations is:

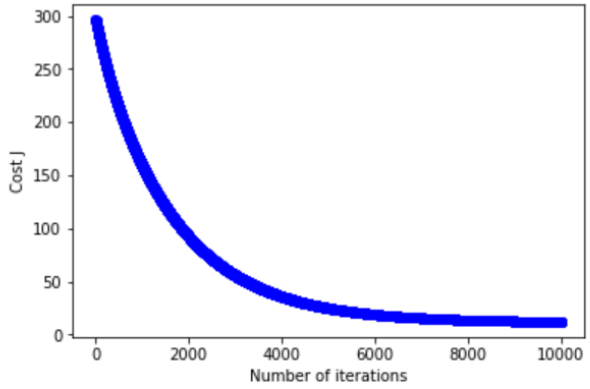


For the **average home in Boston** suburbs, we predict a median home value of **225328.06.** We obtain the same predicted result after computing theta using gradient descent and using the closed form solution.

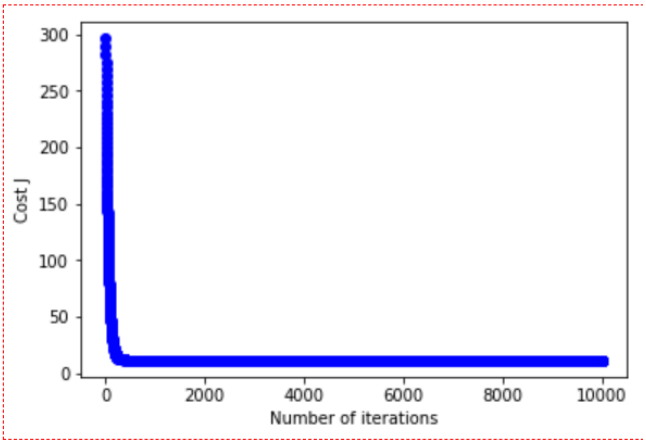
From what we have observed, the ideal learning rate for this problem is **0.01**, because it best resembles the curve we would like to obtain on the graph of convergence of cost versus the number of iterations. The initial cost isn't too high for this learning rate, and it converges quickly. A similar convergence is obtained when we use a learning rate of 0.03, but the convergence occurs quickly and the curve does not resemble the shape we'd like to see. For higher learning rates such as 0.9, 1,5, and higher, the graph diverges and the cost increases exponentially with the number of iterations. For lower learning rates of 0.003 and 0.001, the convergence happens too slowly over the number of iterations. The following figures illustrate these points:



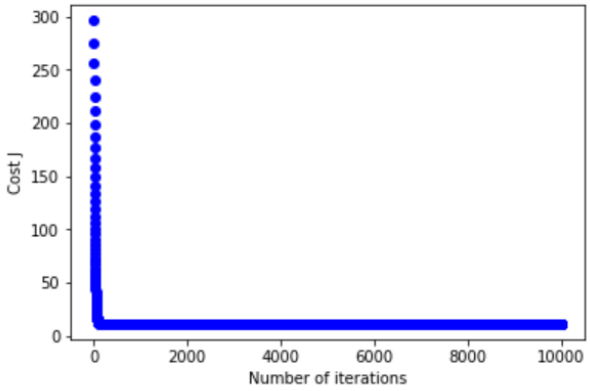
PLOT FOR LEARNING RATE 0.0003



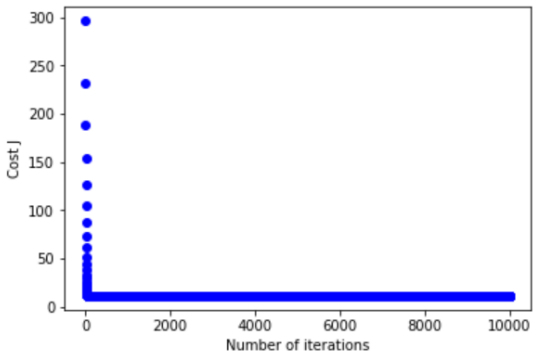
PLOT FOR LEARNING RATE 0.01



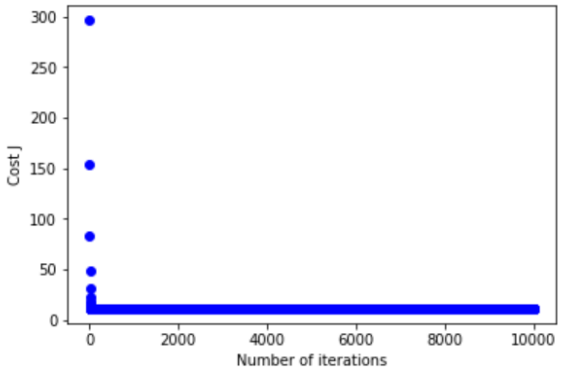
PLOT FOR LEARNING RATE 0.03



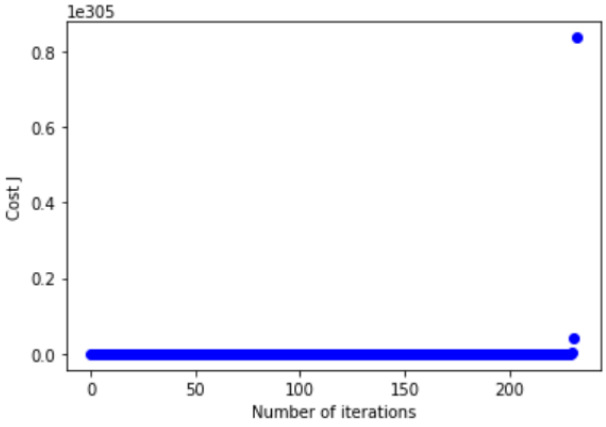
PLOT FOR LEARNING RATE 0.1

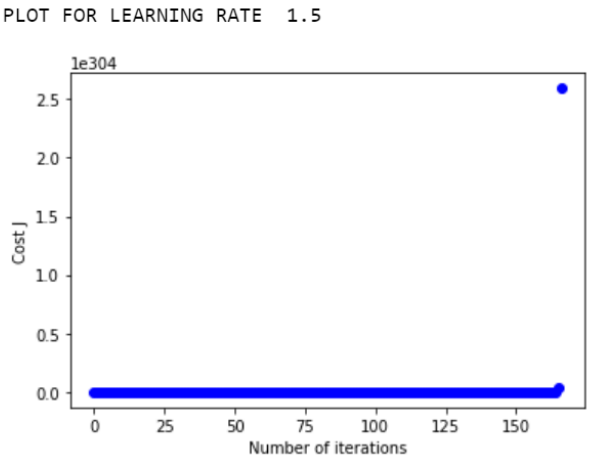


PLOT FOR LEARNING RATE 0.3



PLOT FOR LEARNING RATE 0.9





Problem 3.2: Implementing regularized linear regression (35 points)

Problem 3.2. A1: Regularized linear regression cost function (Figure 7)

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

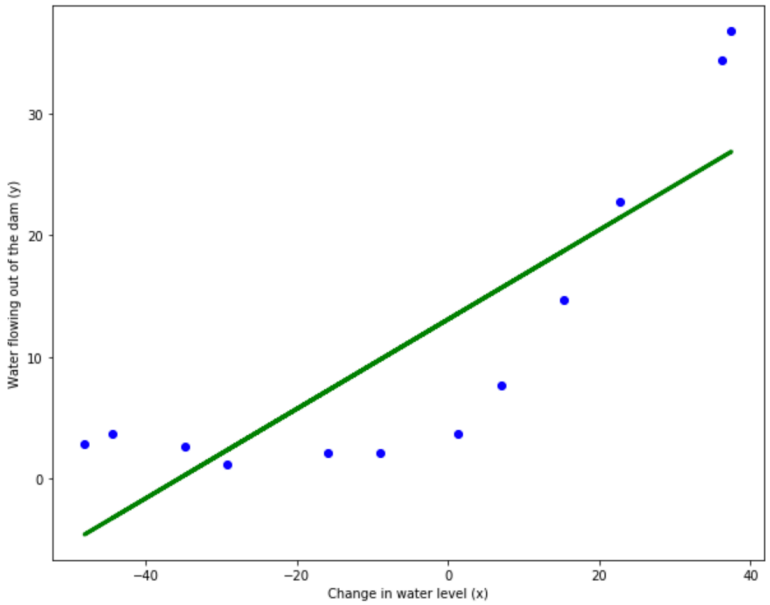


Figure 7: The best fit line for the training data

Problem 3.2.A3: Learning curves

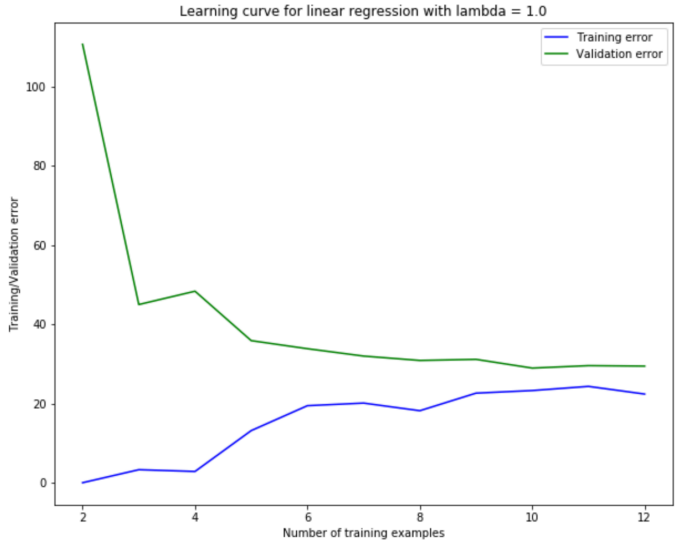


Figure 8: Learning curves $\lambda = 1$

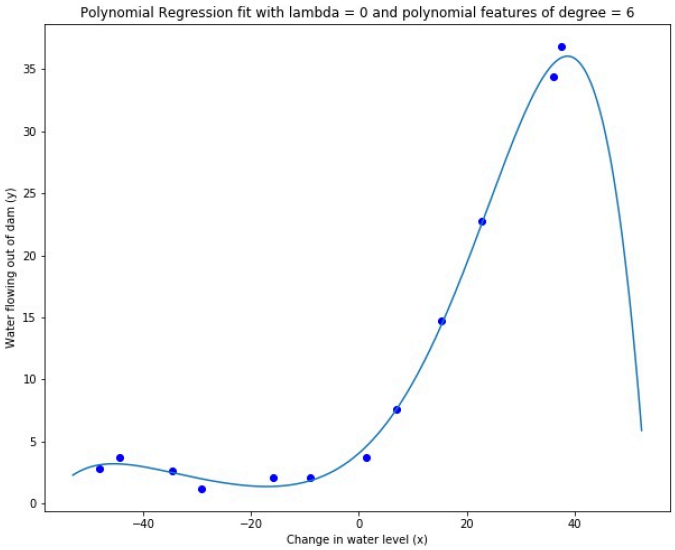


Figure 9: Polynomial fit for lambda = 0 with a p=6 order model.

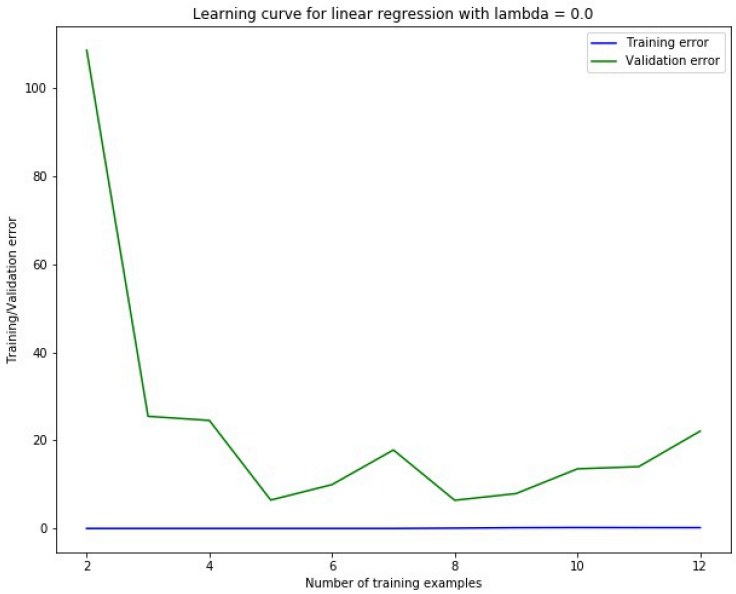


Figure 10: Learning curve for lambda = 0.

Problem 3.2.A4: Adjusting the regularization parameter

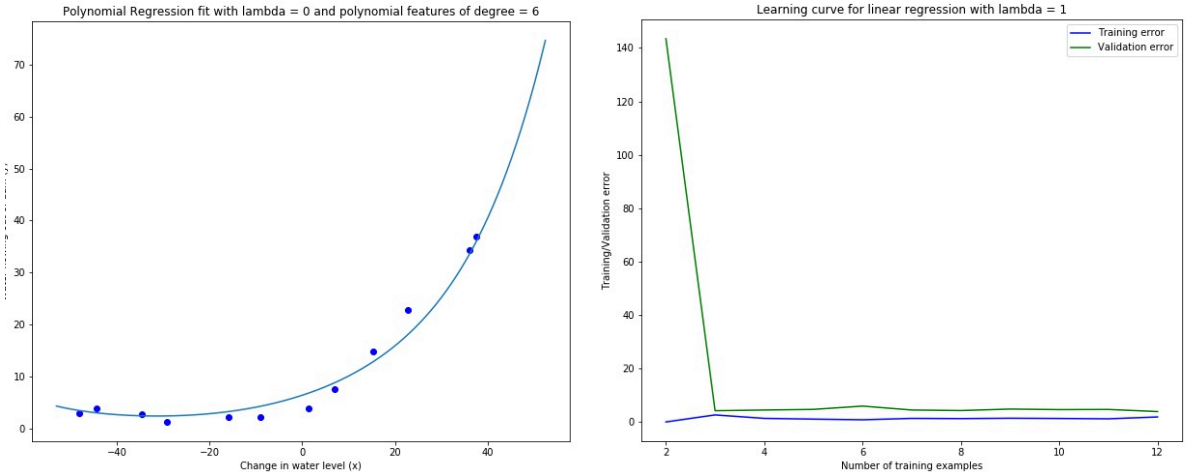


Figure 11: Polynomial fit and learning curve for lambda = 1 with a p=6 order model

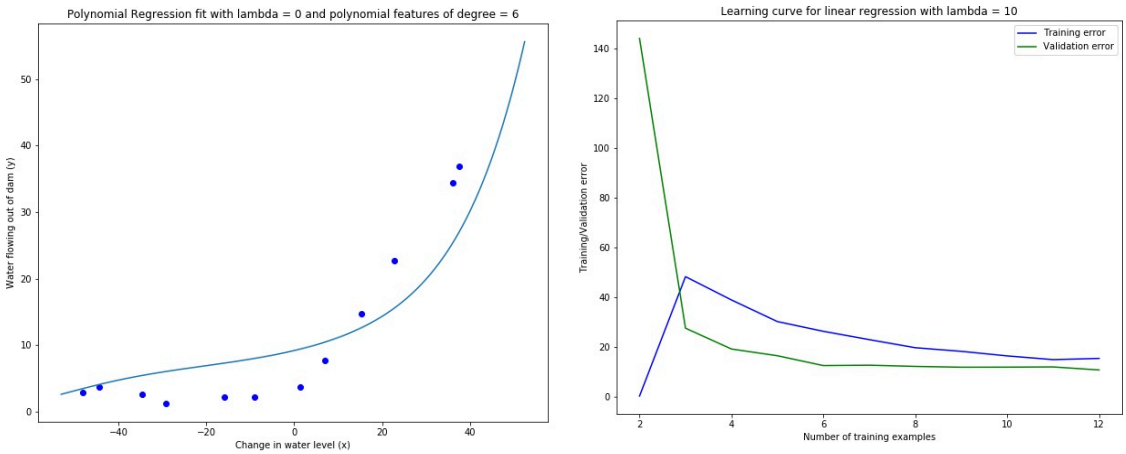


Figure 12: Polynomial fit and learning curve for lambda = 10 with a p=6 order model

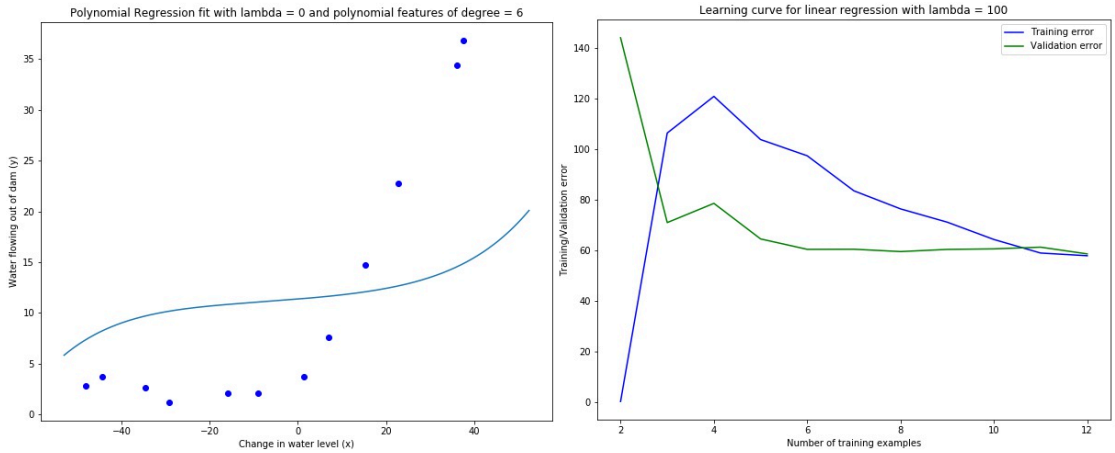


Figure 13: Polynomial fit and learning curve for lambda = 100 with a p=6 order model

According to these three plots, on the one hand, when the value of lambda increasing, it would be less likely to become overfitting. But on the other the hand, the test and train error of model would be larger, that means the bias and variance of model would be larger. Judging from these three lambdas, it would be better to choose lambda=1.

Problem 3.2.A5: Selecting λ using a validation set

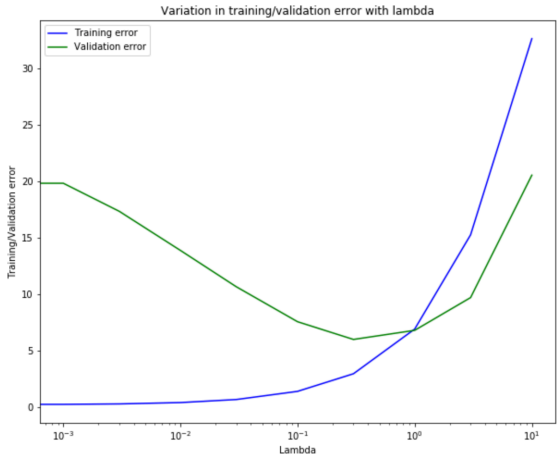


Figure 14: Variation of training/validation error with lambda

According to the plot, the plot of Validation Error is a convex function, so it exists a lowest point. If lambda is too small, it cannot restrict the problem of overfitting enough. If lambda is too large, the bias and variance of model would be larger. And the best lamb da is 0.3. That's the lowest point of Validation Error.

Problem 3.2.A6: Computing test set error

When lambda=0.3, the value of Validation Error will be lowest.
Error of training: 2.9313342465235688
Error of validation: 5.969807577475818

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

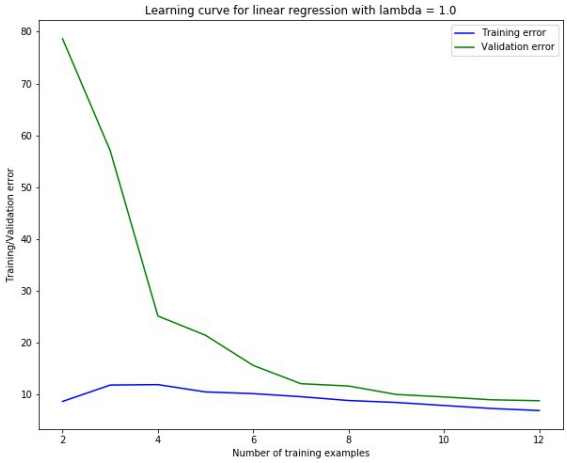


Figure 11: Averaged Learning curve for lambda = 1.