RICE UNIVERSITY

COMP 540

Machine Learning

Honor Pledge:

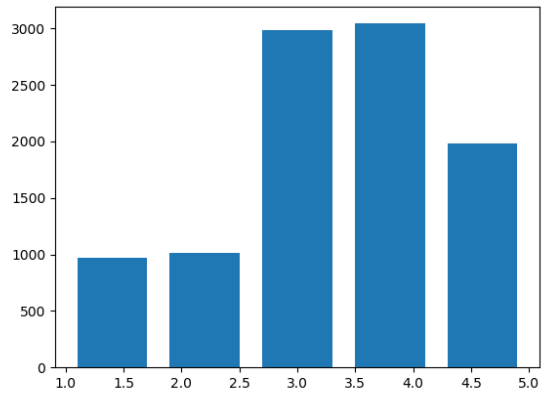
On my honor, I have neither given nor received any unauthorized aid on this assignment.

Mingrui Liang (ml86)
Yuetong Yang (yy72)

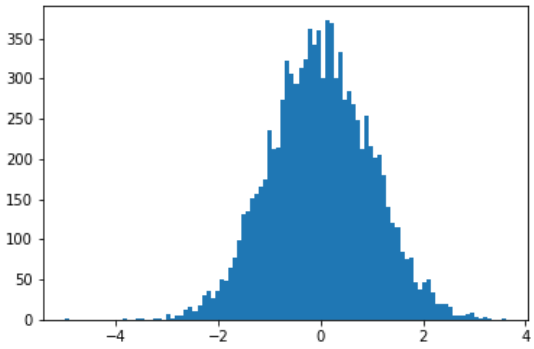
HW # 1

Problem 0: Background refresher

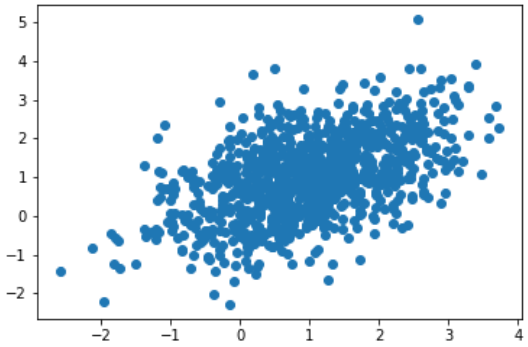
- Sampling Distributions (please see file q0p1.ipynb for more details if needed)
  - Plot the histogram of samples generated by a categorical distribution with probabilities [0.1,0.1,0.3,0.3,0.2]



    These 5 bars are for categories 1, 2, 3, 4 and 5.
  - Plot the univariate normal distribution with mean of 0 and standard deviation of 1.



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



– Test your mixture sampling code by writing a function that implements an equal-weighted mixture of four Gaussians in 2 dimensions, centered at $(\pm 1, \pm 1)$ 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 about 0.1797.

- For part 2 to 10 of problem 0, problem 1 and problem 2, please see the scanned document of our handwritten solution (starts at next page)

Problem 0

(2)

Let $X \sim \text{Poisson}(\lambda_1)$, $Y \sim \text{Poisson}(\lambda_2)$

Use moment generating function:

$\Rightarrow M_X(t) = e^{\lambda_1(e^t-1)}$, $M_Y(t) = e^{\lambda_2(e^t-1)}$

Since X and Y are independent.

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) = e^{(\lambda_1+\lambda_2)(e^t-1)}$$

$\Rightarrow X+Y \sim \text{poisson}(\lambda_1+\lambda_2)$

(3)

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

$$P(X_1 = x_1 \mid X_0 = x_0) = \alpha_1 e^{-\frac{(x_1-x_0)^2}{2\sigma^2}}$$

$$\Rightarrow P(X_1 = x_1) = \int_{-\infty}^{+\infty} P(X_0 = x_0) \cdot P(X_1 = x_1 \mid X_0 = x_0)\, dx_0$$

$$= \alpha_0\alpha_1 \int_{-\infty}^{+\infty} e^{-\frac{\sigma^2(x_0-\mu_0)^2 + \sigma_0^2(x_1-x_0)^2}{2\sigma_0^2\sigma^2}} dx_0$$

$$= \alpha_0\alpha_1 \int_{-\infty}^{+\infty} e^{-\frac{(\sigma^2+\sigma_0^2)x_0^2 - 2x_0(\sigma^2\mu_0+\sigma_0^2x_1) + \mu_0^2\sigma^2 + \sigma_0^2x_1^2}{2\sigma_0^2\sigma^2}} dx_0$$

$$= \alpha_0\alpha_1 \int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma_0^2\sigma^2}\left[\left(\sqrt{\sigma^2+\sigma_0^2}\, x_0 - \frac{-\sigma^2\mu_0+\sigma_0^2x_1}{\sqrt{\sigma^2+\sigma_0^2}}\right)^2 + \sigma^2\mu_0^2 + \sigma_0^2x_1^2 - \frac{\sigma^2\mu_0^2+\sigma_0^2x_1^2}{\sigma^2+\sigma_0^2}\right]} dx_0$$

Since $\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$

$$\Rightarrow P(X_1 = x_1) = \sqrt{2\pi\sigma_0^2\sigma^2/\sigma^2+\sigma_0^2} \cdot \alpha_0\alpha_1 \cdot \exp\left\{-\frac{1}{2\sigma_0^2\sigma^2}\left[\frac{\sigma^2\sigma_0^2\mu_0^2 + \sigma^2\sigma_0^2x_1^2 - 2\sigma^2\sigma_0^2\mu_0x_1}{\sigma^2+\sigma_0^2}\right]\right\}$$

$$= \sqrt{2\pi\sigma_0^2\sigma^2/\sigma^2+\sigma_0^2} \cdot \alpha_0\alpha_1 \cdot \exp\left\{-\frac{x_1-\mu_0)^2}{2(\sigma^2+\sigma_0^2)^2}\right\}$$

Thus, $\alpha = \sqrt{\frac{2\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2}}\,\alpha_0\alpha_1$

$\mu_1 = \mu_0$

$\sigma_1 = \sqrt{\sigma^2+\sigma_0^2}$

(4) Since $P(A|BC) > P(A|B)$

$\Rightarrow \frac{P(ABC)}{P(BC)} - \frac{P(AB)}{P(B)} > 0 \Rightarrow P(AB)\cdot P(BC) - P(ABC)\cdot P(B) < 0$ ①

If we want to get $P(A|BC^c) < P(A|B)$

we need $\frac{P(AB)-P(ABC)}{P(B)-P(BC)} < \frac{P(AB)}{P(B)}$

$\Leftrightarrow [P(AB)-P(ABC)]\cdot P(B) - [P(B)-P(BC)]\cdot P(AB) < 0$

$\Leftrightarrow P(AB)\cdot P(BC) - P(ABC)\cdot P(B) < 0$ which is equal to ①

Thus, $P(A|BC^c) < P(A|B)$

(5) $u = [1\ 2]^T$, $v = [2\ 3]^T$, $M = uv^T$

$\Rightarrow M = \begin{bmatrix}1\\2\end{bmatrix} \times [2\ 3] = \begin{bmatrix}2 & 3\\4 & 6\end{bmatrix}$

$M-\lambda I = \begin{bmatrix}2 & 3\\4 & 6\end{bmatrix} - \begin{bmatrix}\lambda & 0\\0 & \lambda\end{bmatrix} = \begin{bmatrix}2-\lambda & 3\\4 & 6-\lambda\end{bmatrix}$

$\det\begin{bmatrix}2-\lambda & 3\\4 & 6-\lambda\end{bmatrix} = (2-\lambda)(6-\lambda) - 4\times 3 = \lambda^2 - 8\lambda = \lambda(\lambda-8)$

$\Rightarrow$ eigenvalues are $\lambda = 0$, $\lambda = 8$

when $\lambda = 0$, $(M-\lambda I)X = \begin{pmatrix}2 & 3\\4 & 6\end{pmatrix} X = 0 \Rightarrow X = \begin{pmatrix}3\\-2\end{pmatrix}$

when $\lambda = 8$, $(M-\lambda I)X = \begin{pmatrix}-6 & 3\\4 & -2\end{pmatrix} X = 0 \Rightarrow X = \begin{pmatrix}1\\2\end{pmatrix}$

(6)

Let λ be an eigenvalue, v be an eigenvector,

By definition: $Av = \lambda v$

multiply both sides with v^T,

since A is positive semi-definite,

$\Rightarrow v^T A v = \lambda v^T v \geq 0$

we know $v^T v \geq 0$

$\Rightarrow \lambda \geq 0$

(7) ① $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$

Since $AB = 0$, $BA = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \neq 0$

$\Rightarrow (A+B)^2 = A^2 + AB + BA + B^2 \neq A^2 + 2AB + B^2$

② $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, $B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, we can see $AB = 0$

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

$= I^2 - 2uu^T I - 2uu^T I + 4uu^T uu^T$

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

$\Rightarrow A$ is orthogonal

c9) ① $f''(x) = (x^3)'' = (3x^2)' = 6x \geq 0$, for $x \geq 0$

$\Rightarrow f(x)$ is convex for $x \geq 0$

② Let $h(x) = \max(f(x), g(x))$, $\lambda \in (0, 1)$

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

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

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

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

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

③ Since $f(x)$ and $g(x)$ are convex on S

$\Rightarrow f''(x) \geq 0$, $g''(x) \geq 0$

$\Rightarrow (f+g)''(x) = f''(x) + g''(x) \geq 0$

$\Rightarrow f+g$ is convex on S

④ f and g have minimum within S at the same point

$\Rightarrow f'(x) = g'(x) = 0 \Rightarrow x = x_0$

Since f and g are convex and non-negative on S

$\Rightarrow f$ and g are decreasing when $x < x_0$, non-decreasing when $x > x_0$

$$\left.\begin{array}{l} \Rightarrow f'(x) < 0,\ g'(x) < 0,\ x < x_0 \\ \quad f'(x) \geq 0,\ g'(x) \geq 0,\ x \geq x_0 \end{array}\right\} \Rightarrow f'(x) \cdot g'(x) \geq 0$$

Thus $(fg)''(x) = [(fg)'(x)]' = (f'g + g'f)' = f''g + f'g' + g''f + gf'$

$$= f''g + g''f + 2f'g' \geq 0$$

$\Rightarrow fg$ is convex on S

c10) $H(p_1, \dots, p_k) = -\sum_{i=1}^{k} p_i \log(p_i)$

$G(p_1, \dots, p_k) = \sum_{i=1}^{k} p_i = 1$

Let $\frac{\partial}{\partial p_i}\left[H(p) + \lambda(G-1)\right] = 0$

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

$\Rightarrow \log p_1 = \log p_2 = \dots = \log p_i = \lambda - 1$

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

$\Rightarrow p_1 = p_2 = \dots = p_k = \frac{1}{k}$

$H(p_1, \dots, p_k) = -\sum_{i=1}^{k} \frac{1}{k} \log(\frac{1}{k}) = \log k$ is the highest entropy.

Problem 1

c11) denote $W = \begin{pmatrix} \frac{w^{(1)}}{2} & & & \\ & \frac{w^{(2)}}{2} & & \\ & & \ddots & \\ & & & \frac{w^{(m)}}{2} \end{pmatrix}_{m \times m}$

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

$$= \left((x^{(1)}\theta^{(1)} - y^{(1)}) \cdot \frac{w^{(1)}}{2} \quad (x^{(2)}\theta^{(2)} - y^{(2)}) \cdot \frac{w^{(2)}}{2} \quad \dots \quad (x^{(m)}\cdot\theta^{(m)} - y^{(m)}) \cdot \frac{w^{(m)}}{2}\right)_{1 \times m} \cdot (X_{m \times d}\,\theta_{d \times 1} - y_{m \times 1})_{m \times 1}$$

$$= (x^{(1)}\theta^{(1)} - y^{(1)}) \cdot \frac{w^{(1)}}{2} \cdot (x^{(1)}\cdot\theta^{(1)} - y^{(1)}) + (x^{(2)}\theta^{(2)} - y^{(2)}) \cdot \frac{w^{(2)}}{2} \cdot (x^{(2)}\cdot\theta^{(2)} - y^{(2)})$$

$$+ \dots + (x^{(m)}\theta^{(m)} - y^{(m)}) \cdot \frac{w^{(m)}}{2} \cdot (x^{(m)}\theta^{(m)} - y^{(m)})$$

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

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

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

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

$$= \theta^T X^T w X\theta - \theta^T X^T w y - y^T w X\theta - y^T w y$$

$$= \theta^T X^T w X\theta - 2\theta^T X^T w y - y^T w y$$

$$\Rightarrow \frac{J(\theta)}{d\theta} = 2X^T w X\theta - 2X^T w y \overset{\text{set}}{=} 0$$

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

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

(3) The algorithm is as follow:

① randomly choose a θ as the initialization

② calculate the weight matrix w using given bandwidth parameter τ, the diagonal element of w is:

$$w^{(i)} = \exp\left[-\frac{(x - x^{(i)})^T (x - x^{(i)})}{2\tau^2}\right] \quad i = 1, 2, \cdots, m$$

③ update θ: $(0 < j \leq d)$

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j} = \theta_j - \alpha \sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)} - y^{(i)}) x_j^{(i)}$$

repeat until covergence

This would be a non-parametric method, because for different dataset X, the size of parameter $w^{(i)}$ is different.

Problem 2

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

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

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

$E(\theta) = (X^T X)^{-1} X^T \cdot E(Y)$

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

$= \theta^*$

Thus, for the least squares estimator $E(\theta) = \theta^*$

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

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

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

Since $(X^T X)^{-1}$ is idempotent,

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

Problem 3: Implementing linear regression and regularized linear regression

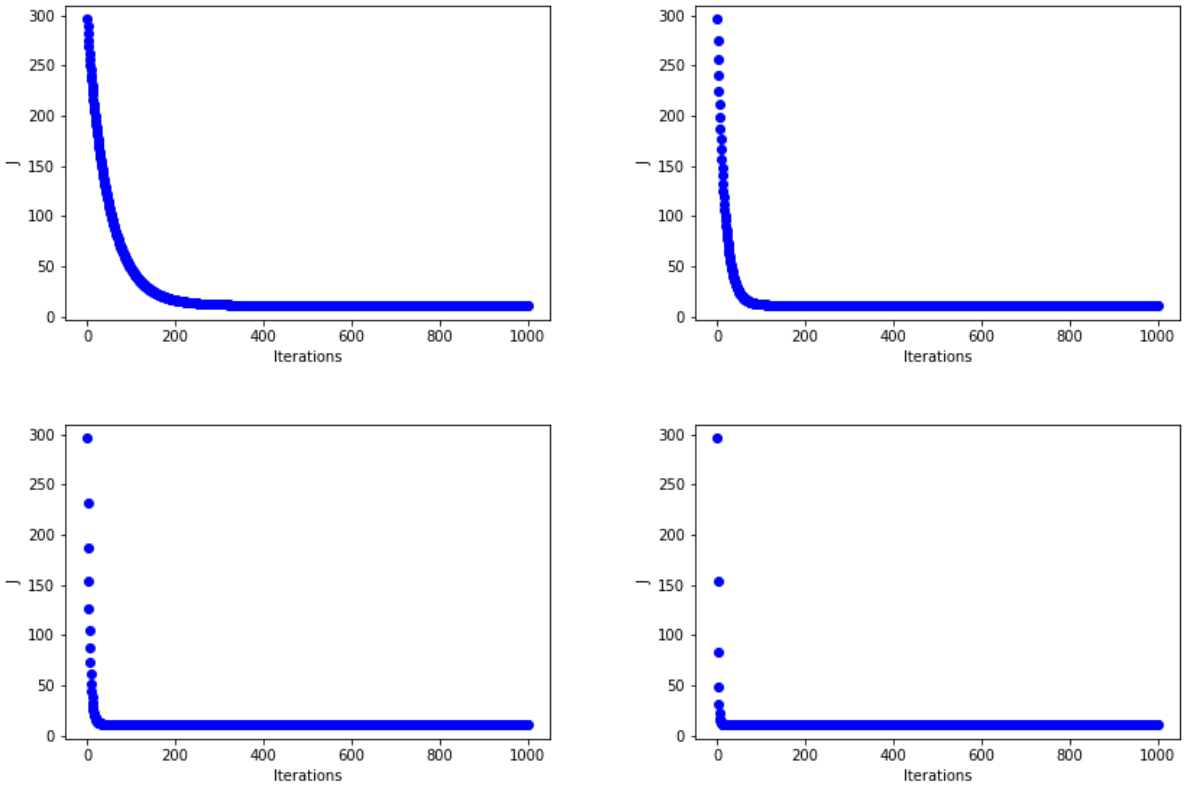
Problem 3.1.A3

The result we got from our code is:

```
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
```

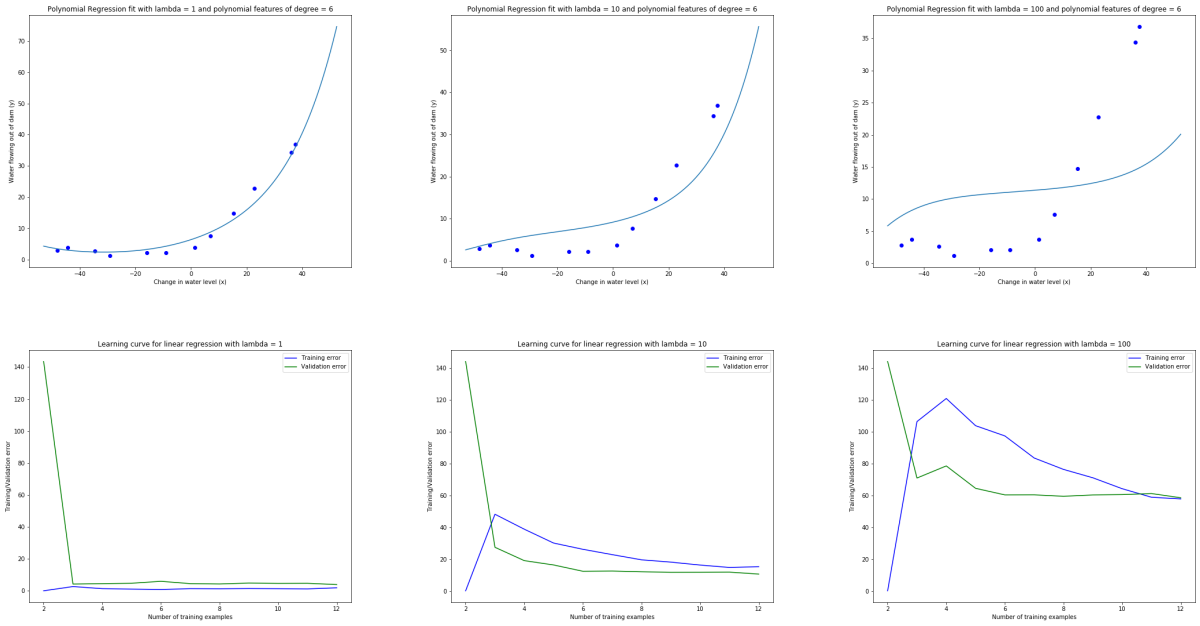
Since a negative home value is meaningless, our prediction would be:
"Predicted median value of a home with LSTAT = 5% is 298034.49, predicted median value of a home with LSTAT = 50% is 0"

Problem 3.1.B5



We choose to run experiments with learning rate equals to 0.01, 0.03, 0.1 and 0.3 according to the homework instruction. Plots at the left top, right top, left bottom and right bottom are for learning rate equals 0.01, 0.03, 0.1 and 0.3 correspondingly. From these 4 plots we can see that with the learning rate equals 0.3 we reach convergence at the fastest speed. We need less than 50 iterations to reach convergence in this case.

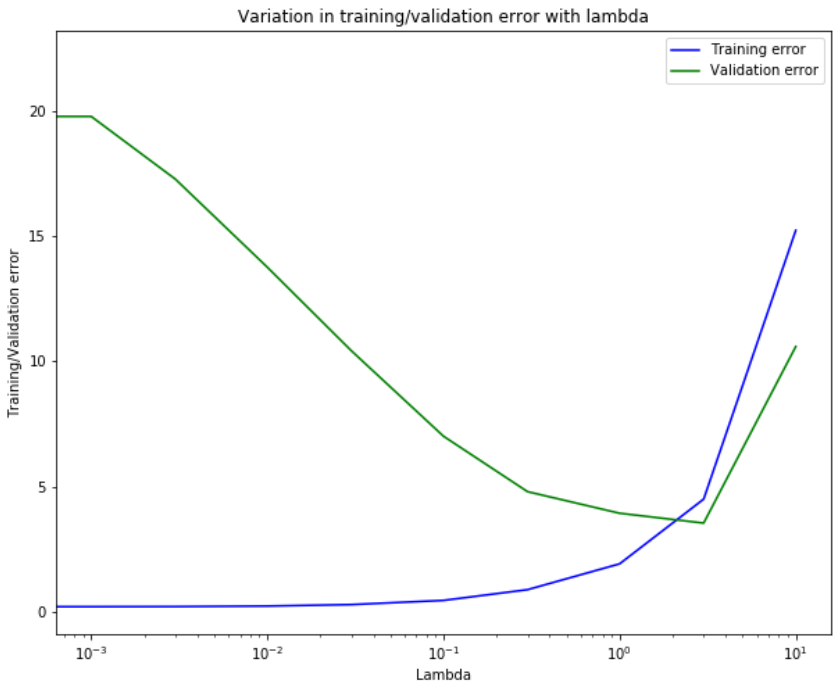
Problem 3.2.A4



Plots at the top are the polynomial regression fits while plots at the bottom are the learning curves for linear regression. From left to right we have plots for $\lambda = 1, 10, 100$ respectively.
Compared with the case where $\lambda = 0$ (figure 9 and 10 in the homework instruction), we can see the fitted line for $\lambda = 1$ looks much more like a curve (instead of going ups and downs along with the training data) and the validation error drops to a small value much faster, which indicates less overfitting issue. As λ increases, we can see that the linear regression starts to have underfitting issues, where the fitted line becomes further and further from the training points and both the training and validation errors become higher. This makes sense since the higher order terms are "less encouraged" when λ increases. When $\lambda = 100$ we can see the fitted line is basically flat, very similar to a straight line. Personally, I would choose $\lambda = 1$ in my final model.

Problem 3.2.A5

The required plot is as follows:



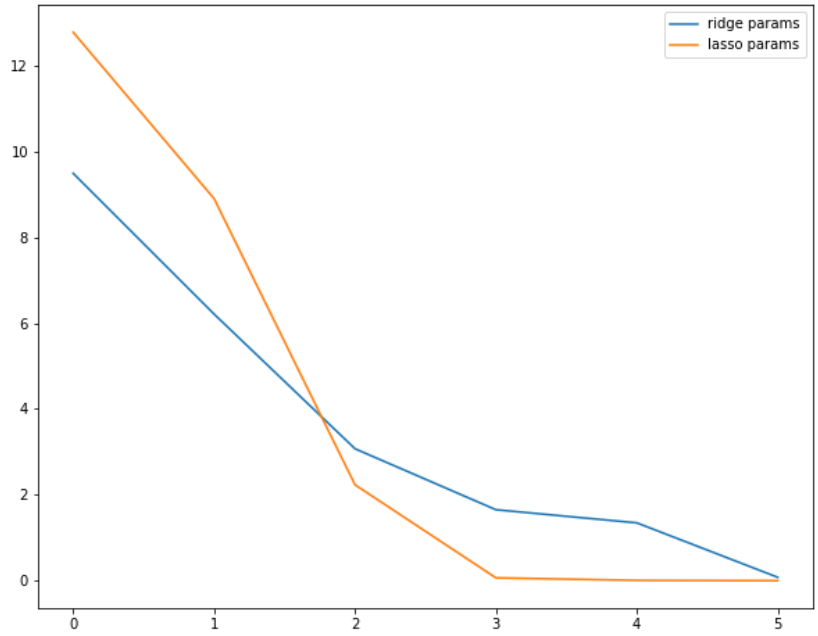
We can see that the optimal value for λ would be $\lambda = 3$.

Problem 3.2.A6

The test error for $\lambda = 3$ is 4.3976. It is also interesting to point out that for $\lambda = 1$ we will see a smaller test error of 3.0987. I guess the saying is true that our model can never be perfect!

Problem 3.2.A8

The required plot is as follows:



Compared with ridge regression, we can see that lasso penalize higher order terms harder. For lower order terms (1st and 2nd order), lasso parameters have higher magnitude, whereas for third and higher terms, ridge regression has higher magnitude. We also notice that in lasso for terms that are higher than third order the magnitude is very close to 0, which confirmed the saying in the lecture that "Lasso often puts more terms to 0".

Other plots appeared in the homework instruction

We ran the script and made all other plots that are shown (but are not required to be put in the writeup) in the homework instruction. We put them here to complete this homework writeup.

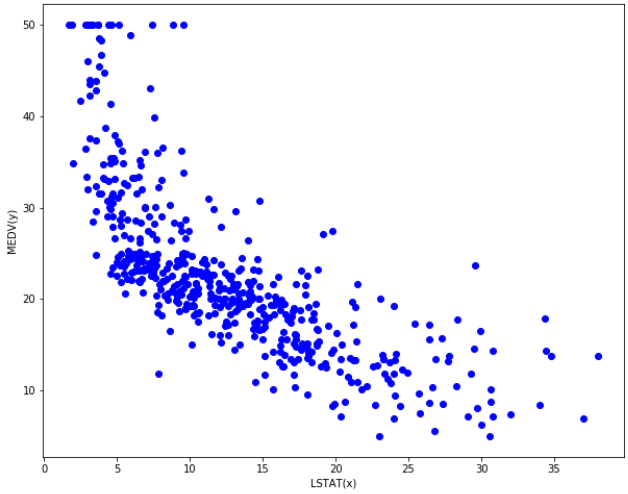


Figure 1 in homework instruction

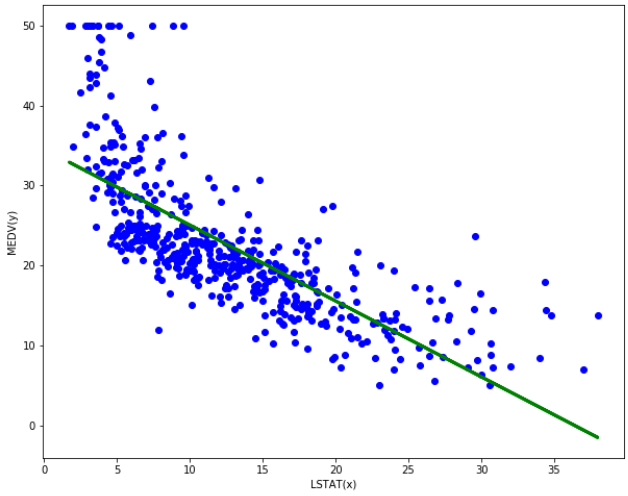


Figure 2 in homework instruction

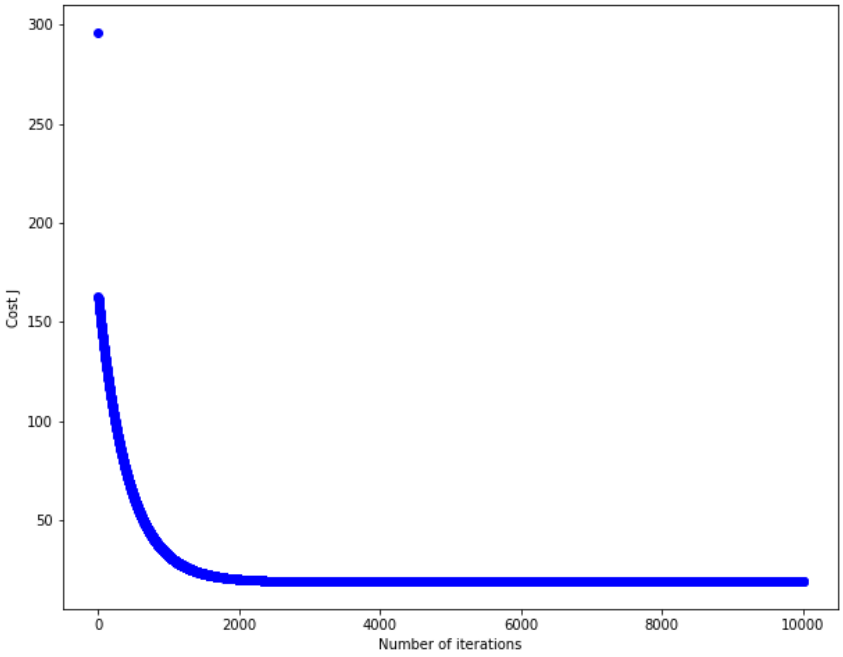


Figure 3 in homework instruction

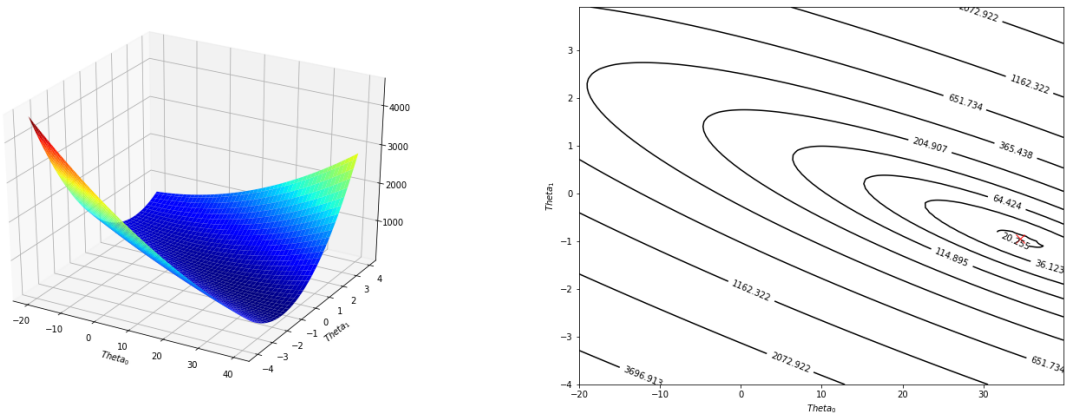


Figure 4 in homework instruction

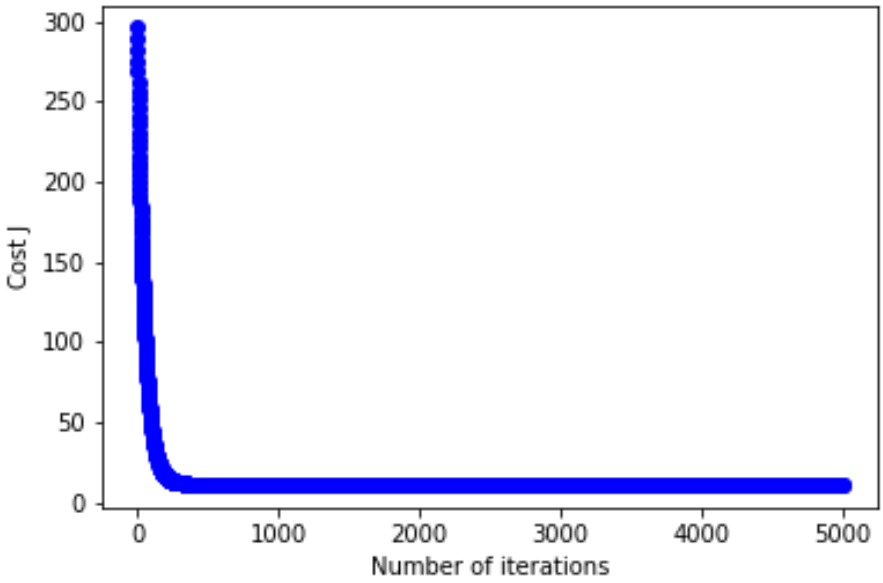


Figure 5 in homework instruction

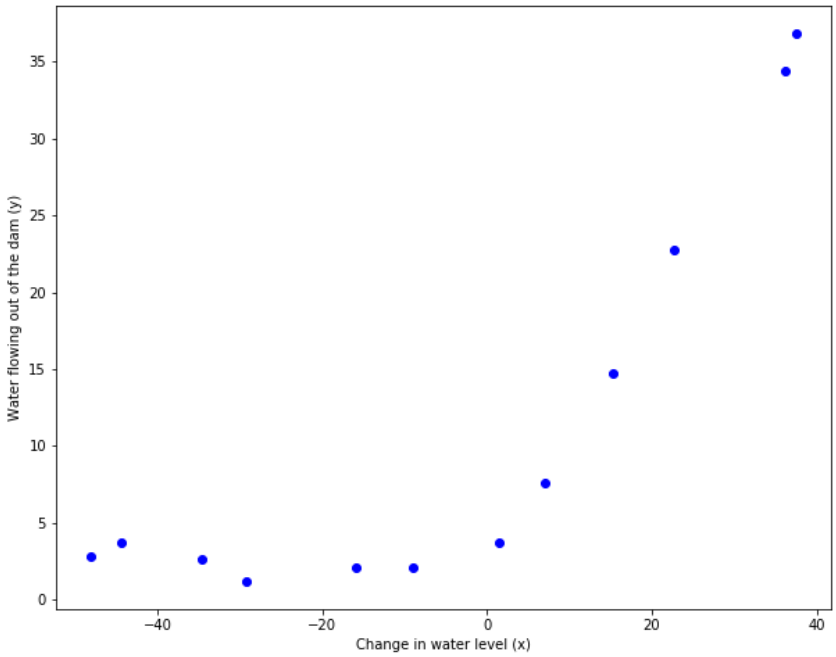


Figure 6 in homework instruction

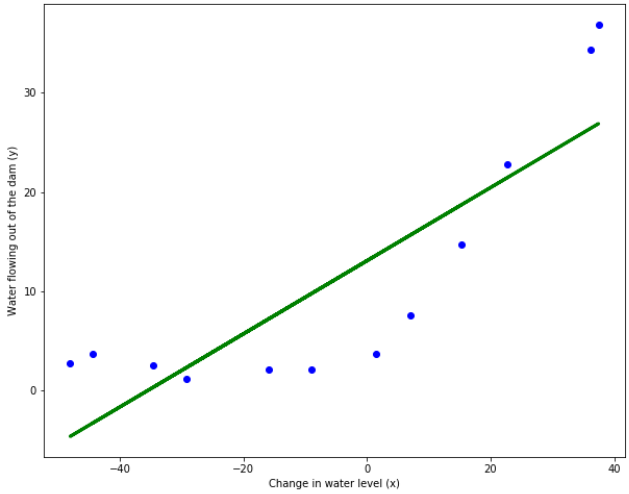


Figure 7 in homework instruction

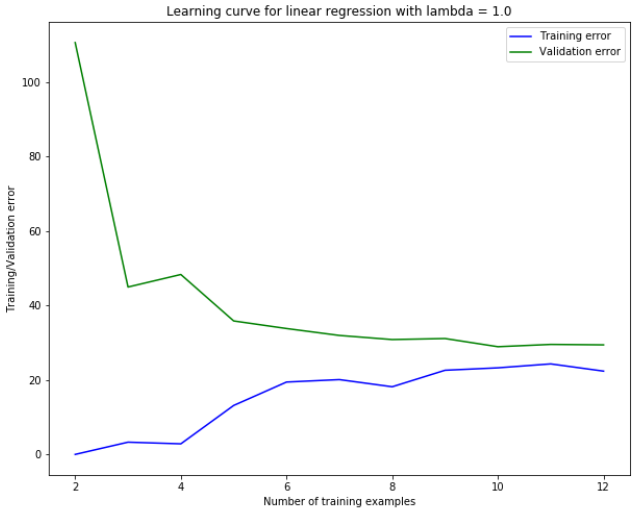


Figure 8 in homework instruction