

COMP 540 HW 01

Lyu Pan (lp28), Yuhui Tong (yt30)

January 17, 2018

0. Background Refresher

0.0 Samplers

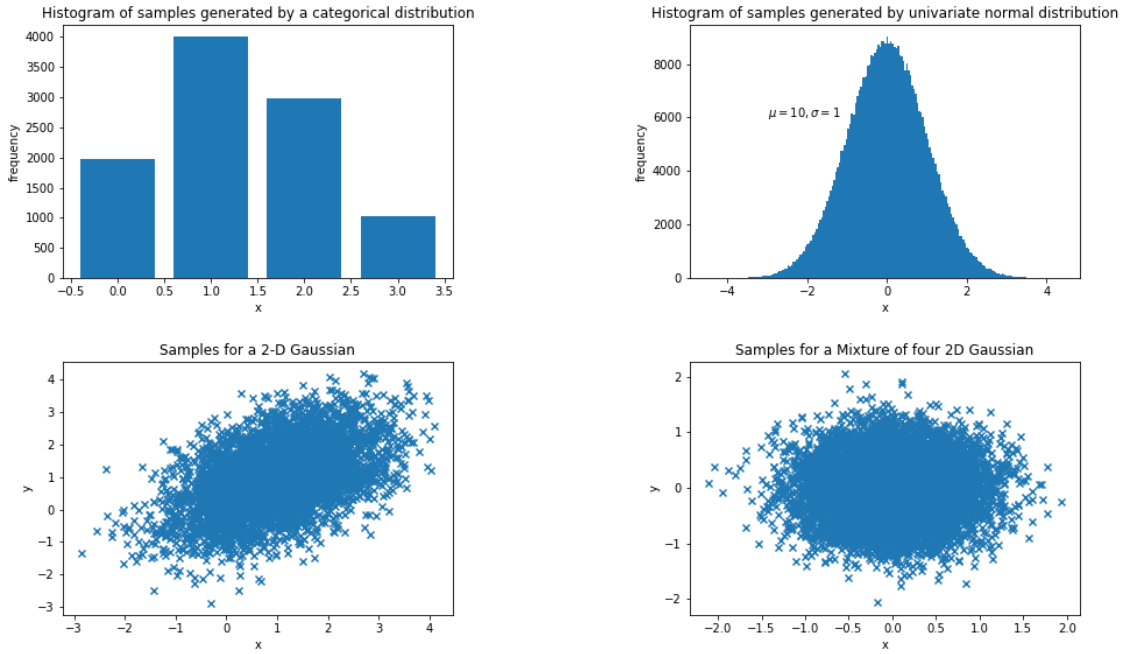


Figure 1: Visualization of four distributions

Figure 1 visualize four distributions. Specifically, for a mixture distribution of four 2D Gaussians, the probability that a sample from this distribution lies within the unit circle centered at $(0.1, 0.2)$ is $p = 0.83736$.

0.1 Prove two independent Poisson random variables are also Poisson variable.

Proof:

Given two independent random variables $X_1 \sim P(\lambda_1)$ and $X \sim P(\lambda_2)$. i.e.,

$$P(X_1 = m) = e^{-\lambda_1} \frac{\lambda_1^m}{m!}, \quad m = 1, 2, \dots \quad (1)$$

and

$$P(X_2 = n) = e^{-\lambda_2} \frac{\lambda_2^n}{n!}, \quad n = 0, 1, 2, \dots \quad (2)$$

Denote sum of them are

$$X = X_1 + X_2 \quad (3)$$

then the probability distribution of X is:

$$\begin{aligned} P(X = k) &= \sum_{m+n=k} P(X_1 = m, X_2 = n) \\ &\stackrel{X_1, X_2 \text{ indep. R.V.}}{=} \sum_{m+n=k} P(X_1 = m) P(X_2 = n) \\ &= \sum_{m+n=k} e^{-\lambda_1} \frac{\lambda_1^m}{m!} e^{-\lambda_2} \frac{\lambda_2^n}{n!} \\ &= \frac{1}{k!} e^{-(\lambda_1 + \lambda_2)} \sum_{m+n=k} \frac{k!}{m!n!} \lambda_1^m \lambda_2^n \\ &= \frac{1}{k!} e^{-(\lambda_1 + \lambda_2)} (\lambda_1 + \lambda_2)^k \end{aligned} \quad (4)$$

that is, $X = X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$. Q.E.D.

0.2 Proof question 2

$$\begin{aligned} p(x_1, x_0) &= p(x_1 | x_0) p(x_0) \\ &= \alpha e^{-\frac{(x_1 - x_0)^2}{2\sigma^2}} \alpha_0 e^{-\frac{(x_0 - \mu_0)^2}{2\sigma_0^2}} \end{aligned} \quad (5)$$

The probability distribution of X_1 is,

$$\begin{aligned} p(x_1) &= \int_{-\infty}^{\infty} p(x_1, x_0) dx_0 \\ &= \int_{-\infty}^{\infty} \alpha e^{-\frac{(x_1 - x_0)^2}{2\sigma^2}} \alpha_0 e^{-\frac{(x_0 - \mu_0)^2}{2\sigma_0^2}} dx_0 \\ &= D \int_{-\infty}^{\infty} e^{-A[x_0 - B]^2 + C} dx_0 \end{aligned} \quad (6)$$

where A, B, C, D

$$D = \alpha \alpha_0 e^{-\frac{\mu_0^2}{2\sigma_0^2}} e^{-\frac{x_1^2}{2\sigma^2}} \quad (7)$$

$$A = \frac{\sigma_0^2 + \sigma^2}{2\sigma_0^2 \sigma^2}, \quad (8)$$

$$B = \frac{x_1 \sigma_0^2 + \mu_0 \sigma^2}{\sigma_0^2 + \sigma^2}, \quad (9)$$

$$C = \frac{(x_1\sigma_0^2 + \mu_0\sigma^2)^2}{2\sigma_0^2\sigma^2(\sigma_0^2 + \sigma^2)}. \quad (10)$$

Making the substitution $t = \sqrt{2A}(x_0 - B)$ gives

$$\begin{aligned} \int_{-\infty}^{\infty} e^{-A[x_0-B]^2+C} dx_0 &= \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}+C} d\frac{t}{\sqrt{2A}} \\ &= \frac{1}{\sqrt{2A}} e^{-C} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt \\ &= \sqrt{\frac{\pi}{A}} e^{-C} \end{aligned} \quad (11)$$

where in the second step we used $\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt = \sqrt{2\pi}$.

Substituting the above equation to the Eq. (6) yields

$$\begin{aligned} p(x_1) &= D\sqrt{\frac{\pi}{A}} e^{-C} \\ &= \alpha\alpha_0 e^{-\frac{\mu_0^2}{2\sigma_0^2}} e^{-\frac{x_1^2}{2\sigma^2}} \sqrt{\frac{\pi}{A}} e^{-C} \end{aligned} \quad (12)$$

0.3 question 4 eigenvalues

$$A = \begin{bmatrix} 13 & 5 \\ 2 & 4 \end{bmatrix}$$

$$\begin{aligned} A\mathbf{X} &= \lambda\mathbf{X} \\ (A - \lambda I)\mathbf{X} &= 0 \\ \begin{bmatrix} 13-\lambda & 5 \\ 2 & 4-\lambda \end{bmatrix} \mathbf{X} &= 0 \end{aligned}$$

$$\lambda_1 = 3, \mathbf{X}_1 = \begin{pmatrix} -1/2 \\ 1 \end{pmatrix}$$

$$\lambda_2 = 14, \mathbf{X}_2 = \begin{pmatrix} 5 \\ 1 \end{pmatrix}$$

0.4 question 5, matrix multiplication

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}, \text{ we have } (A+B)^2 \neq A^2 + 2AB + B^2$$

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, A, B \neq 0 \text{ and we have } AB = 0$$

0.5 question 6

$$\begin{aligned} A^T A &= (I - 2uu^T)^T(I - 2uu^T) \\ &= (I - 2uu^T)(I - 2uu^T) \\ &= I - 2uu^T - 2uu^T + 4u(u^T u)u^T \\ &= I \end{aligned} \tag{13}$$

0.6 convex function

0.6.1

for $x \geq 0$

$$f(x) = 3x^3 \tag{14}$$

$$f'(x) = 3x^2 \tag{15}$$

$$f''(x) = 6x \geq 0 \tag{16}$$

so, $f(x) = x^3$ is convex.

0.6.2

For any $\lambda \in [0, 1]$,

we have $\lambda \geq 0, 1 - \lambda \geq 0$

For any $(x_1, y_1), (x_2, y_2) \text{ on } R^2$

$$\begin{aligned} f(\lambda(x_1, y_1) + (1 - \lambda)(x_2, y_2)) &= \max(\lambda(x_1, y_1) + (1 - \lambda)(x_2, y_2)) \\ &\leq \max(\lambda(x_1, y_1)) + \max((1 - \lambda)(x_2, y_2)) \\ &= \lambda \max(x_1, y_1) + (1 - \lambda) \max(x_2, y_2) \\ &= \lambda f(x_1, y_1) + (1 - \lambda) f(x_2, y_2) \end{aligned} \tag{17}$$

So, $f(x)$ is convex.

0.6.3

Because f is convex on S , for $\lambda \in [0, 1]$ and all $x_1, x_2 \in S$, we have

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \tag{18}$$

Also,

$$g(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda g(x_1) + (1 - \lambda)g(x_2) \tag{19}$$

Let $h = f + g$,

$$\begin{aligned} h(\lambda x_1 + (1 - \lambda)x_2) &= f(\lambda x_1 + (1 - \lambda)x_2) + g(\lambda x_1 + (1 - \lambda)x_2) \\ &\leq \lambda f(x_1) + (1 - \lambda)f(x_2) + \lambda g(x_1) + (1 - \lambda)g(x_2) \\ &= \lambda(f(x_1) + g(x_1)) + (1 - \lambda)(f(x_2) + g(x_2)) \\ &= \lambda h(x_1) + (1 - \lambda)h(x_2) \end{aligned} \tag{20}$$

So, h is convex, i.e. $f + g$ is convex.

0.6.4

0.7 entropy of categorical distribution

The entropy of a categorical distribution on K values is

$$H(p) = - \sum_{i=1}^K p_i \log(p_i), \quad (21)$$

with constraint that

$$\sum_{i=1}^K p_i = 1. \quad (22)$$

Using Lagrange Multiplier, one can combine the above two equations into:

$$L(p, \lambda) = - \sum_{i=1}^K p_i \log(p_i) + \lambda \left(\sum_{i=1}^K p_i - 1 \right). \quad (23)$$

Taking derivative of all unknown variables gives:

$$\frac{\partial L}{\partial p_i} = -(\log p_i + 1) + \lambda = 0, \quad i = 1, 2, \dots \quad (24)$$

Substituting p_i with λ yields

$$p_i = \frac{1}{K}, \quad i = 1, 2, \dots, K. \quad (25)$$

Q.E.D.

1. Locally weighted linear regression.

1.1 Expression of $J(\theta)$

Define \mathbf{X} , \mathbf{W} in the following way:

$$\mathbf{X} = \begin{bmatrix} \text{---} & \text{---} & \text{---} & (x^{(1)})^T & \text{---} & \text{---} & \text{---} \\ \text{---} & \text{---} & \text{---} & (x^{(2)})^T & \text{---} & \text{---} & \text{---} \\ & & & \vdots & & & \\ \text{---} & \text{---} & \text{---} & (x^{(m)})^T & \text{---} & \text{---} & \text{---} \end{bmatrix} \iff \mathbf{X}_{i,j} = (x^{(i)})_j \quad (26)$$

$$\mathbf{W} = \begin{bmatrix} w^{(1)} & & & & \\ & w^{(2)} & & & \\ & & \ddots & & \\ & & & w^{(m)} & \end{bmatrix} \iff \mathbf{W}_{i,j} = \delta_{i,j} w^{(i)} \quad (27)$$

Substituting \mathbf{X} , \mathbf{W} into $(\mathbf{X}\theta - \mathbf{y})^T \mathbf{W} (\mathbf{X}\theta - \mathbf{y})$ yields

$$\begin{aligned}
(\mathbf{X}\theta - \mathbf{y})^T \mathbf{W} (\mathbf{X}\theta - \mathbf{y}) &= \sum_{i,j} [(\mathbf{X}\theta - \mathbf{y})^T]_{1,i} \mathbf{W}_{i,j} (\mathbf{X}\theta - \mathbf{y})_{j,1} \\
&= \sum_{i,j} \delta_{i,j} w^{(i)} [(\mathbf{X}\theta - \mathbf{y})^T]_{1,i} (\mathbf{X}\theta - \mathbf{y})_{j,1} \\
&= \sum_i w^{(i)} [(\mathbf{X}\theta - \mathbf{y})_{i,1}]^2 \\
&= \sum_i w^{(i)} \left[\left(x^{(i)} \right)^T \theta - y^{(i)} \right]^2 \\
&= \sum_i w^{(i)} \left[\theta x^{(i)} - y^{(i)} \right]^2.
\end{aligned}$$

So $J(\theta) = \frac{1}{2}(\mathbf{X}\theta - \mathbf{y})^T \mathbf{W} (\mathbf{X}\theta - \mathbf{y})$. Q.E.D.

1.2 Closed formed solution of θ

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta} &= \frac{1}{2} \frac{\partial}{\partial \theta} [(\mathbf{X}\theta - \mathbf{y})^T \mathbf{W} (\mathbf{X}\theta - \mathbf{y})] \\
&= \frac{1}{2} \frac{\partial}{\partial \theta} \text{tr} [\mathbf{y}^T \mathbf{W} \mathbf{X} \theta - \mathbf{y}^T \mathbf{W} \mathbf{y} - \theta^T \mathbf{X}^T \mathbf{W} \mathbf{X} \theta + \theta^T \mathbf{X}^T \mathbf{W} \mathbf{y}] \\
&= \mathbf{X}^T \mathbf{W} \mathbf{y} - \mathbf{X}^T \mathbf{W} \mathbf{X} \theta
\end{aligned}$$

where in the third line we used the following two equations¹:

$$\frac{\partial \text{tr}(\mathbf{A}^T \mathbf{B})}{\partial \mathbf{A}} = \mathbf{B}^T \quad (28)$$

and

$$\frac{\partial \text{tr}(\mathbf{A}^T \mathbf{B} \mathbf{A})}{\partial \mathbf{A}} = \mathbf{B} \mathbf{A} + \mathbf{B}^T \mathbf{A} \quad (29)$$

Therefore the closed form solution for θ is

$$\theta = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}. \quad (30)$$

¹Take the second equation for example:

$$\begin{aligned}
\left[\frac{\partial \text{tr}(\mathbf{A}^T \mathbf{B} \mathbf{A})}{\partial \mathbf{A}} \right]_{i,j} &= \frac{\partial A_{l,m}^T B_{m,n} A_{n,l}}{\partial A_{i,j}} = \frac{\partial A_{m,l} B_{m,n} A_{n,l}}{\partial A_{i,j}} \\
&= \delta_{m,i} \delta_{l,j} B_{m,n} A_{n,l} + A_{m,l} B_{m,n} \delta_{n,i} \delta_{l,j} \\
&= B_{i,n} A_{n,j} + A_{m,j} B_{m,i} \\
&= (\mathbf{B} \mathbf{A})_{i,j} + (\mathbf{B}^T \mathbf{A})_{i,j},
\end{aligned}$$

hence leading to $\frac{\partial \text{tr}(\mathbf{A}^T \mathbf{B} \mathbf{A})}{\partial \mathbf{A}} = \mathbf{B} \mathbf{A} + \mathbf{B}^T \mathbf{A}$

1.3 Batch gradient descent for locally weighted linear regression

The derivative of J_θ is:

$$\frac{\partial}{\partial \theta} J(\theta) = \sum_{i=1}^m w^{(i)} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}. \quad (31)$$

Therefore we have the following algorithm:

Algorithm 1: Batch gradient descent algorithm for locally weighted linear regression

Result: The estimated parameters θ for locally weighted linear regression

```

1 while  $\theta$  does not converge do
2   for every  $j$  do
3      $\theta_j = \theta_j - \alpha \sum_{i=1}^m w^{(i)} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$ ;
4   end
5 end
```

2. Properties of linear regression estimator.

2.1 Prove $E[\theta] = \theta^*$

Proof:

The following facts:

1. $y^{(i)} = \theta^{*T} x^{(i)} + \epsilon^{(i)}$,
2. $\epsilon^{(i)}, i = 1, 2, \dots, m$ are *i.i.d.* of $N(0, \sigma^2)$,

indicate that:

given fixed arbitrary $x^{(i)}$ and fixed unknown parameter θ^* , $y^{(i)}$, $1 \leq i \leq m$ are *i.i.d.* of $N(\theta^{*T} x^{(i)}, \sigma^2)$

$$p(y^{(i)} | x^{(i)}, \theta^*) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y^{(i)} - \theta^{*T} x^{(i)})^2}{2\sigma^2}}. \quad (32)$$

The least-square estimate of θ^* is θ given by

$$\theta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (33)$$

$$\text{denote: } (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \equiv \mathbf{A} \quad \mathbf{A} \mathbf{y} \quad (34)$$

The expectation of θ is

$$E[\theta] = E[\mathbf{A} \mathbf{y}] \quad (35)$$

$$= E \begin{bmatrix} \sum_j A_{1,j} y^{(j)} \\ \sum_j A_{2,j} y^{(j)} \\ \vdots \\ \sum_j A_{d+1,j} y^{(j)} \end{bmatrix}. \quad (36)$$

Since expectations has the following property (regardless of the independence of Z_k):

$$E[Z_1 + Z_2 + \dots + Z_l] = \sum_k Z_k, \quad (37)$$

we have

$$E[\theta] = E \begin{bmatrix} \sum_j A_{1,j} y^{(j)} \\ \sum_j A_{2,j} y^{(j)} \\ \vdots \\ \sum_j A_{d+1,j} y^{(j)} \end{bmatrix} = \begin{bmatrix} \sum_j A_{1,j} E[y^{(j)}] \\ \sum_j A_{2,j} E[y^{(j)}] \\ \vdots \\ \sum_j A_{d+1,j} E[y^{(j)}] \end{bmatrix} \quad (38)$$

$$= \mathbf{A} \begin{bmatrix} E[y^{(1)}] \\ E[y^{(2)}] \\ \vdots \\ E[y^{(m)}] \end{bmatrix} = \mathbf{A} \begin{bmatrix} \theta^{*T} x^{(1)} \\ \theta^{*T} x^{(2)} \\ \vdots \\ \theta^{*T} x^{(m)} \end{bmatrix} \quad (39)$$

$$= \mathbf{A} \mathbf{X} \theta^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \theta^* \quad (40)$$

$$= \theta^* \quad (41)$$

where in the second step the following equation is used: $E[y^{(i)}] = \theta^{*T} x^{(i)}$ (trivial to obtain as $y^{(i)}$ observe normal distribution). Therefore $E[\theta] = \theta^*$, implying that the estimation θ is unbiased. Q.E.D

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

Denote $\mathbf{A} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$, therefore $\theta = \mathbf{A} \mathbf{y}$.

$$Var(\theta) = Var(\mathbf{A} \mathbf{y}) \quad (42)$$

$$= Var \left(\begin{bmatrix} \sum_j A_{1,j} y^{(j)} \\ \sum_j A_{2,j} y^{(j)} \\ \vdots \\ \sum_j A_{m,j} y^{(j)} \end{bmatrix} \right) = \begin{bmatrix} \sum_j A_{1,j} Var(y^{(j)}) \\ \sum_j A_{2,j} Var(y^{(j)}) \\ \vdots \\ \sum_j A_{m,j} Var(y^{(j)}) \end{bmatrix} = \mathbf{A} \begin{bmatrix} Var(y^{(1)}) \\ Var(y^{(2)}) \\ \vdots \\ Var(y^{(m)}) \end{bmatrix}, \quad (43)$$

where in the second line we made used of the property of variance:

$$Var(\sum_k Z_k) = \sum Var(Z_k), \text{ } Z_k \text{ is independent from each other.} \quad (44)$$

Substituting $Var(y^{(i)}) = \sigma^2$ $i = 1, 2, \dots, m$, we have $Var(\theta) = \mathbf{A} \sigma^2 = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2$. Q.E.D.

3. Implementing linear regression and regularized linear regression

3.1.A1

No plot for this question.

3.1.A2

Figure 2 plots the linear fit using parameter obtained through training.

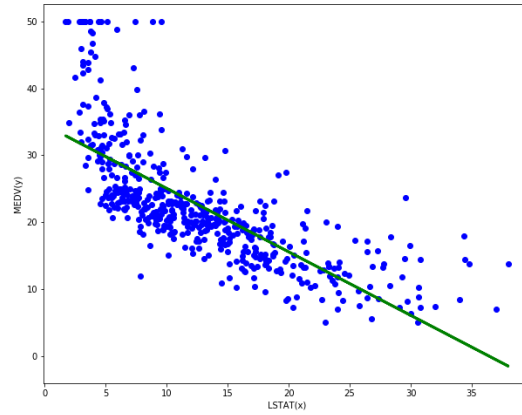


Figure 2: Fitting a linear model to the data

Problem 3.1.A3: Predicting on unseen data

For average home in Boston suburbs, we predict a median home value of 225328.063241 (USD)

3.1.B1: Feature Normalization

No plot for this question.

3.1.B2: Loss function and gradient descent

Figure 3 shows the loss function against the number of iterations.

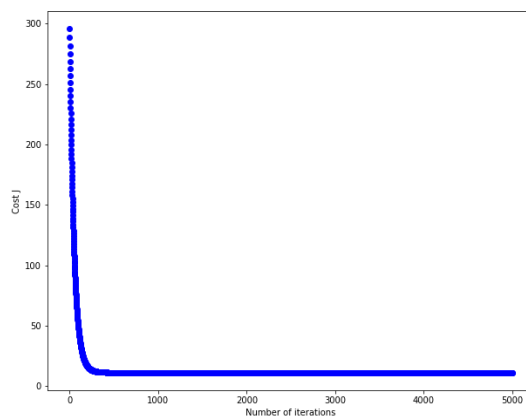


Figure 3: Fitting a linear model to the data

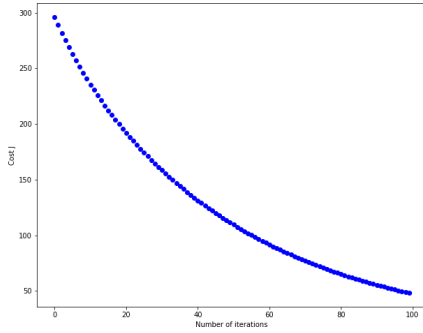
3.1.B3 Making predictions on unseen data

For average home in Boston suburbs, we predict a median home value of 225328.063241

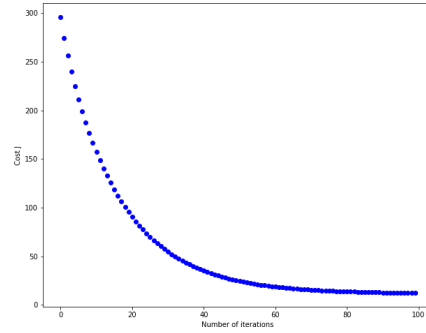
3.1.B4: Normal equations

For average home in Boston suburbs, we predict a median home value of 225328.063241, which is the same as we obtained in subsec. 3.1.B3.

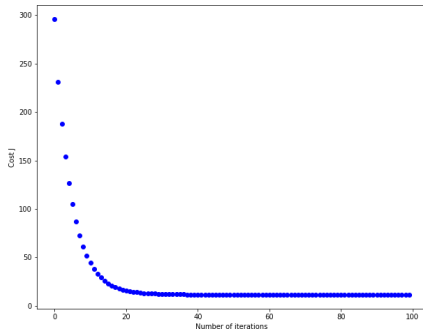
Problem 3.1.B5: Exploring convergence of gradient descent



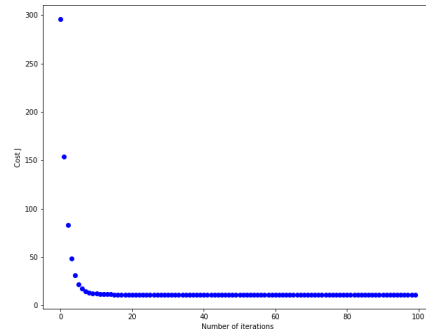
(a) learning rate $\alpha = 0.01$



(b) learning rate $\alpha = 0.03$



(c) learning rate $\alpha = 0.1$



(d) learning rate $\alpha = 0.3$

Figure 4: Convergence of gradient descent for linear regression with multiple variables using different learning rate.

$\alpha = 0.1, 0.3$ and $N_{iteration} = 80$ are good trade off between accuracy and efficiency. By observing Fig. 4d, one can easily find that small α (i.e. $\alpha = 0.01, 0.03$) leads to very slow convergence rate. $\alpha = 0.1, 0.3$ on the other hands, converges swiftly.

3.2.A1: Regularized linear regression cost function

No figures for this question.

3.2.A2: Gradient of the regularized linear regression cost function

Figure 5 shows the fitted curve of the linear model.

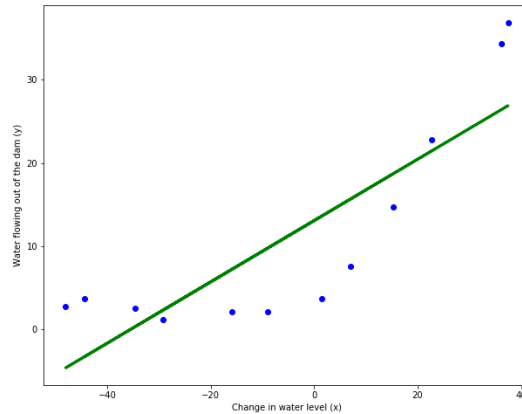


Figure 5: The fitted curve of the linear model.

3.2.A3: Learning curves

Figure 6 shows the learning curve of the linear model.

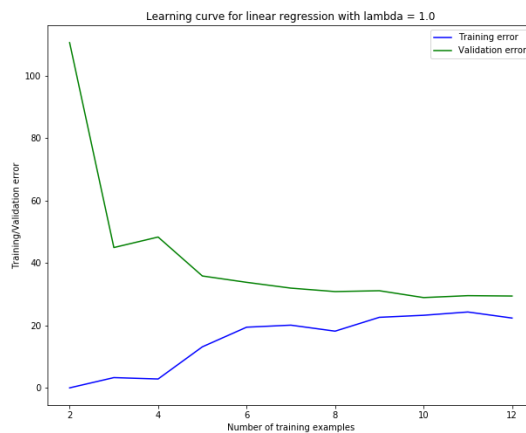
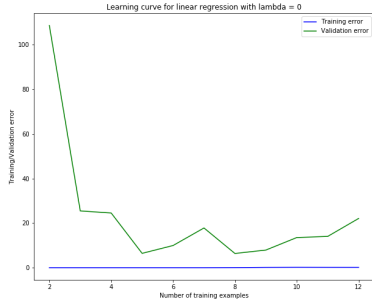


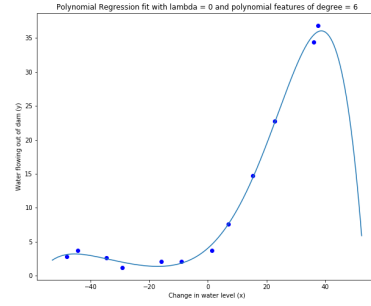
Figure 6: Learning curve of the linear model.

3.2.A4: Adjusting the regularization parameter

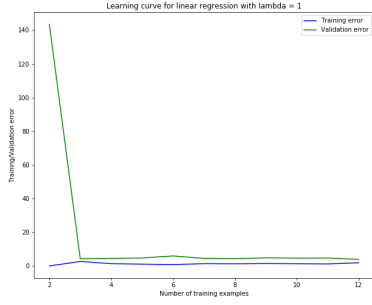
Figure 7 plots the polynomial fit and learning curves for each value of λ , from which we draw the following conclusions:



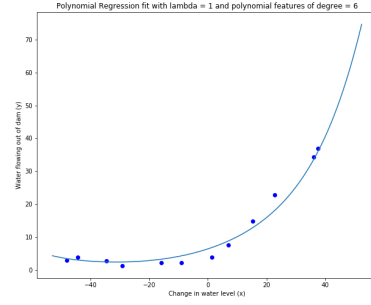
(a) learning rate $\lambda = 0$



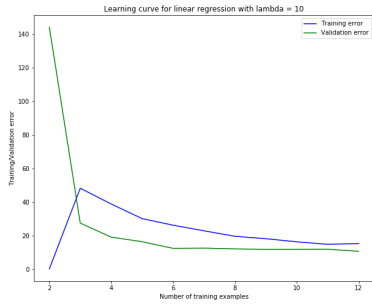
(b) Polynomial fit $\lambda = 0$



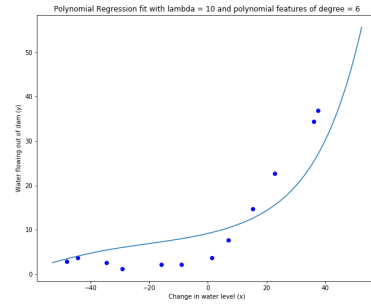
(c) learning rate $\lambda = 1$



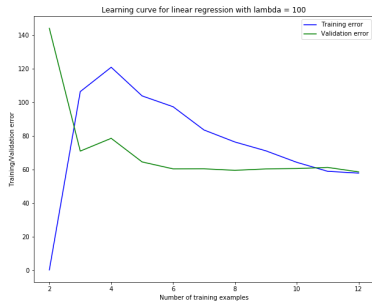
(d) Polynomial fit $\lambda = 1$



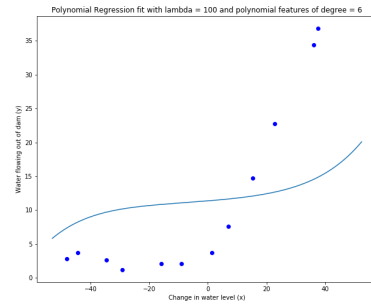
(e) learning rate $\lambda = 10$



(f) Polynomial fit $\lambda = 10$



(g) learning rate $\lambda = 100$



(h) Polynomial fit $\lambda = 100$

Figure 7: Convergence of gradient descent for linear regression with multiple variables using different learning rate.

If λ is too small, the effect of penalty term can be neglected so that the regulation will not be implemented effectively – the training model is still troubled by high-variation/over-fitting issue (as can be seen in Fig....). When λ is too large, the loss function is in fact dominated by the penalty term, which is a slightly similar to the biased issue that we have too strong assumptions on the model. Therefore, the training model turns to be under-fit. Only when λ takes appropriate value that the regulation can works effectively.

3.2.A5 Selecting λ using a validation set

Figure 8 shows the variation in training/validation error with respect of regulation parameter λ . For the current model, $\lambda = 1$ is approximately a good choice for the training. We list several reasons to justify our choice:

- The difference between training/validation error is too large when λ is small, implying an over-fitting problem – this is true because the penalty term characterized by λ is too small to regulate the effect of excessive features.
- The difference between training/validation error becomes relatively small, but the absolute training/validation error becomes too large when λ gets larger than 1. This indicates that the penalty term dominates the loss function and we actually are making a strong assumption of the target function (of similar form as that of penalty term), i.e., too much bias / not enough variation.
- $\lambda = 1$ makes a good balance to avoid either overfitting or underfitting, i.e. the validation/training error is small (prediction performance); the difference between validation/training error is small (implying the prediction model is reliable).

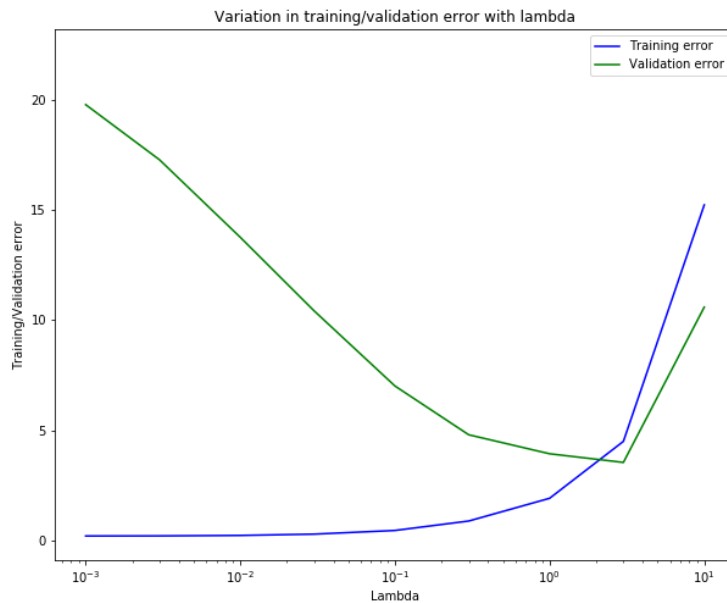


Figure 8: Variation in training/validation error with λ

3.2.A6 Computing test set error

Error when choosing best lamdba 1.0 is 30987.4826556 (USD).

3.2.A7 Plotting learning curves with randomly selected examples

Figure 9 plot the averaged learning curve for $\lambda = 1$.

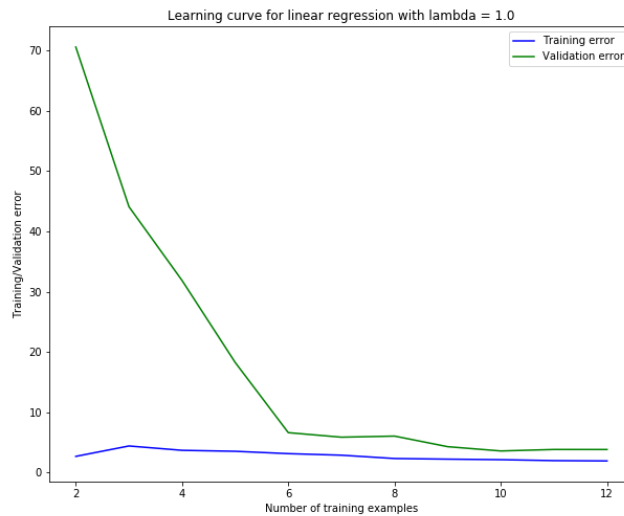


Figure 9: Averaged learning curve for $\lambda = 1$

Extra credit

Please see **bostonexp.pdf** and **bostonexp.ipynb** for our solution.