Sample Questions

Linear Regression – 1

- We want to predict the number of AAs of a student by looking at his/her previous semester's performance.
- In the table below, x denotes the number of AAs that students got in the first year and y denotes the number of AAs that students receives in the next year.
- Our hypothesis $is\ h_{\theta(x)} = \theta_0 + \theta_1 x$, and we use m to denote the number of training examples

x	y
3	4
2	1
4	3
0	1

Linear Regression – 1

- For the training set given above, what is the value of *m*?
- What is J(0,1)? Recall
- Using $\theta_0 = -1, \theta_1 = 2$ in the linear regression hypothesis find $h_\theta(6)$?

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2.$$

Linear Regression – 2

- Let Suppose we want to minimize some function $f(\theta_0,\theta_1)$. Which of the following are true?
  - If θ_0 and θ_1 are initialized so that $\theta_0 = \theta_1$, then by symmetry (because we do simultaneous updates to the two parameters), after one iteration of gradient descent, we will still have $\theta_0 = \theta_1$.
  - Even if the learning rate α is very large, every iteration of gradient descent will decrease the value of $f(\theta_0,\theta_1)$.
  - If θ_0 and θ_1 are initialized at a local minimum, then one iteration will not change their values.
  - If the learning rate is too small, then gradient descent may take a very long time to converge

Linear Regression – 2

Suppose you are able to find $J(\theta_0\,\theta_1)=0$ for some θ_0 and θ_1 for some linear regression problem. Which of the following statements are true?

- For this to be true, we must have θ_0 =0 and θ_1 =0 so that $h_\theta(x)=0$
- For this to be true, we must have $y^{(i)}=0$ for every value of $i=1,2,\ldots,m$
- Our training set can be fit perfectly by a straight line, i.e., all of our training examples lie perfectly on some straight line.
- Gradient descent is likely to get stuck at a local minimum and fail to find the global minimum.

Linear Regression – 3

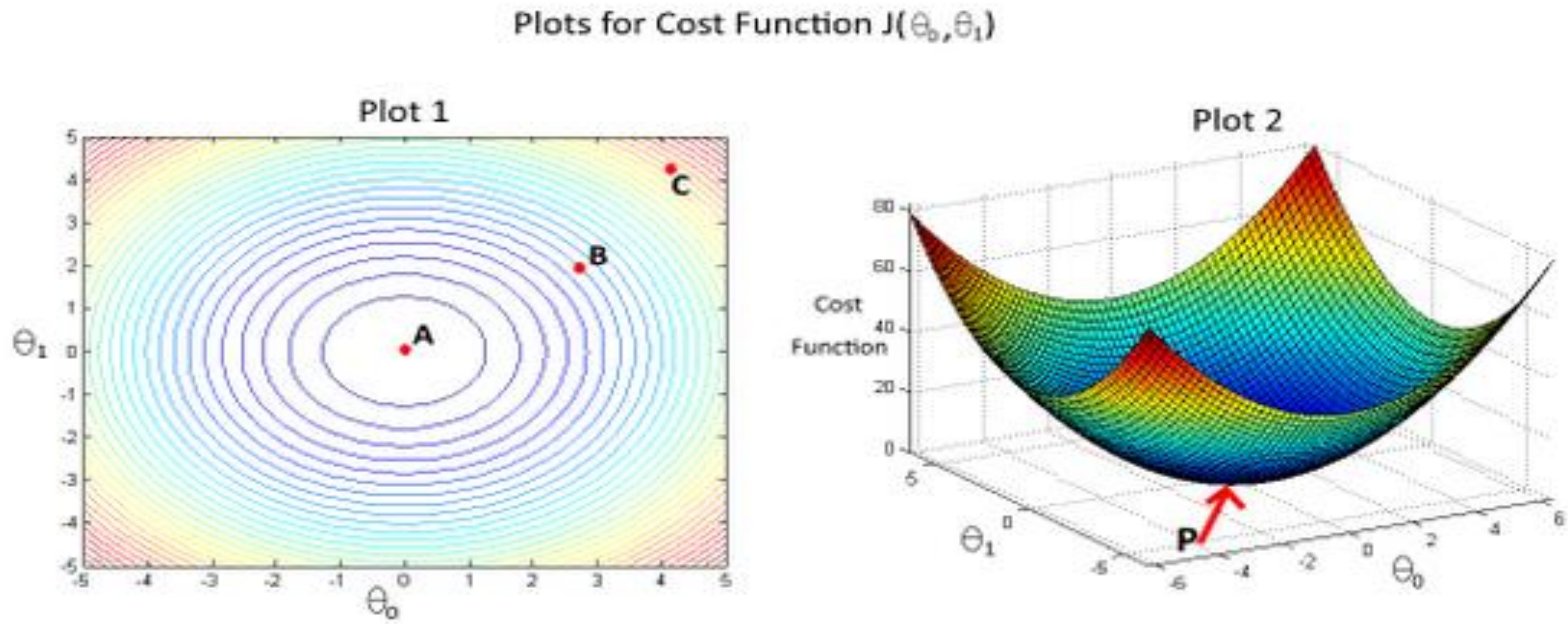
- Q6. Consider the following training set of m=4 training examples. What are the values of θ_0 and θ_1 that you would expect to obtain upon running gradient descent on the linear regression model $h\theta(x)=\theta_0+\theta_0\,\theta_1 x$.

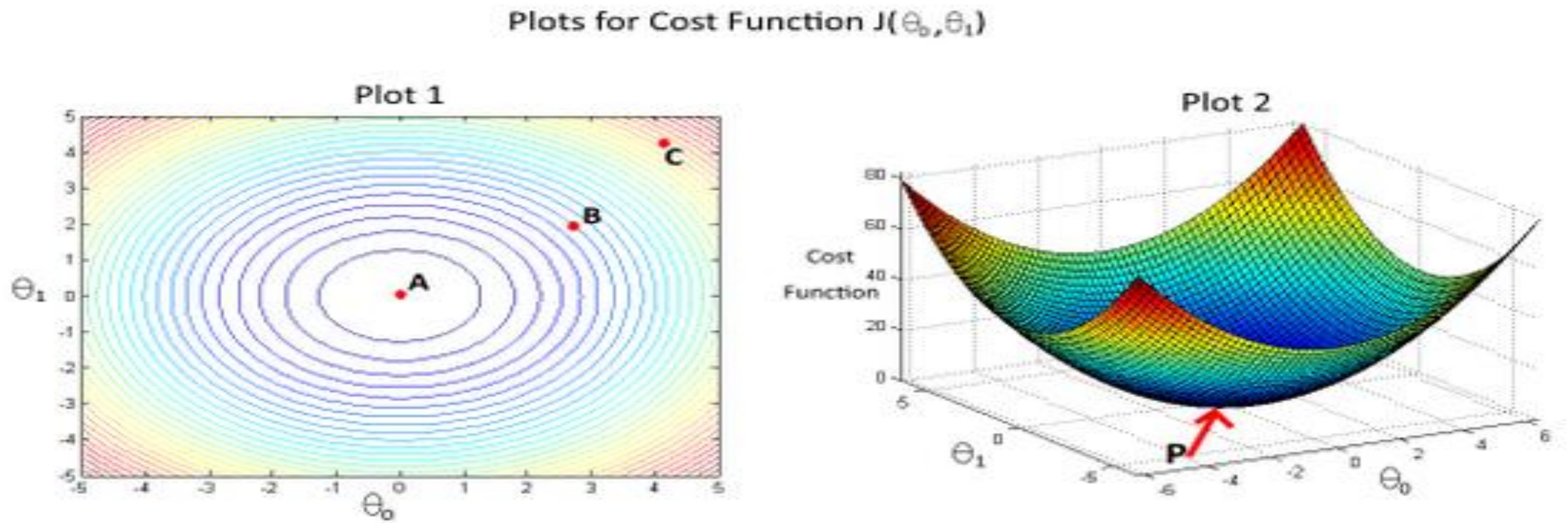
x	y
1	0.5
2	1
4	2
0	0

- ○ $\theta_0 = 1, \theta_1 = 0.5$
- ○ $\theta_0 = 0.5, \theta_1 = 0.5$
- ○ $\theta_0 = 0.5, \theta_1 = 0$
- ○ $\theta_0 = 0, \theta_1 = 0.5$
- ○ $\theta_0 = 1, \theta_1 = 1$

Linear Regression – 4

Q7. In the following you have a plot of the cost function $J(\theta_0,\theta_1)$ (Figure 2) and its contour plot in (Figure 1). Which of the following statements are true.





- If we start from point B, gradient descent with a well-chosen learning rate will eventually help us reach at or near point A, as the value of cost function $J(\theta_0,\theta_1)$ is maximum at point A.
- If we start from point B, gradient descent with a well-chosen learning rate will eventually help us reach at or near point A, as the value of cost function $J(\theta_0,\theta_1)$ is minimum at A.
- If we start from point B, gradient descent with a well-chosen learning rate will eventually help us reach at or near point C, as the value of cost function $J(\theta_0,\theta_1)$ is minimum at point C.
- Point P (The global minimum of plot 2) corresponds to point C of Plot 1.
- Point P (the global minimum of plot 2) corresponds to point A of Plot 1.

Linear Regression – 5

- Suppose that for some linear regression problem (say, predicting housing prices as in the lecture), we have some training set, and for our training set we managed to find some θ_0,θ_1 such that J (θ_0,θ_1)=0.
- Which of the following are true? Check all that apply.

Linear Regression – 5

- For this to be true, we must have $\theta_0 = 0$ and $\theta_1 = 0$ so that $h_\theta(x) = 0$.
- This is not possible: By definition of $J(\theta_0, \theta_1)$, it is not possible for there to exist θ_0 and θ_1 so that $J(\theta_0, \theta_1) = 0$.
- For these values of θ_0 and θ_1 that satisfy $J(\theta_0, \theta_1) = 0$, we have that $h_\theta\left(x^{(i)}\right) = y^{(i)}$ for every training example $(x^{(i)}, y^{(i)})$.
- We can perfectly predict the value of y even for new examples that we have not yet seen. (e.g., we can perfectly predict prices of even new houses that we have not yet seen.)

Linear Regression – 6

Which of the following are reasons for using feature scaling?

- It speeds up gradient descent by making it require fewer iterations to get to a good solution.
- It is necessary to prevent gradient descent from getting stuck in local optima.
- It speeds up solving for θ using the normal equation.
- It prevents the matrix X^TX (used in the normal equation) from being non-invertable (singular/degenerate).

Linear Regression – 7

- You run gradient descent for 15 iterations with α=0.3 and compute $J(\theta)$ after each iteration.
- You find that the value of $J(\theta)$ **decreases** quickly then levels off.
- Based on this, which of the following conclusions seem most plausible?
  - Rather than use the current value of α, it'd be more promising to try a larger value of α(say α=1.0).
  - α=0.3 is an effective choice of learning rate.
  - Rather than use the current value of α, it'd be more promising to try a smaller value of α(say α=0.1).

Linear Regression – 8

- Suppose you want to use an advanced optimization algorithm to minimize the cost function for logistic regression with parameters θ_0 and θ_1.
- You write the following code :

$$function\ [jVal, gradient] = \ costFunction(theta)$$

$$\text{jVal} = \%\ \text{code to compute J}(theta)$$

$$gradient(1)\ =\ \ CODE\#1\ \%\ derivative\ for\ \ theta\_0$$

$$gradient(2)\ =\ \ CODE\#2\ \%\ derivative\ for\ \ theta\_1$$

- What should CODE#1 and CODE#2 above compute?

Linear Regression – 8

- CODE#1 and CODE#2 should compute $J(\theta)$
- CODE#1 should be θ_1 and CODE#2 should be θ_2 .
- CODE#1 should compute
  - $\frac{1}{m}\sum_{i=1}^{m}\left[(h_\theta(x^{(i)}) - y^{(i)}).\, x_0^{(i)}\right](= \frac{\partial}{\partial\theta_0}J(\theta))$
- CODE#2 should compute
  - $\frac{1}{m}\sum_{i=1}^{m}\left[(h_\theta(x^{(i)}) - y^{(i)}).\, x_1^{(i)}\right](= \frac{\partial}{\partial\theta_1}J(\theta))$
- None of them.

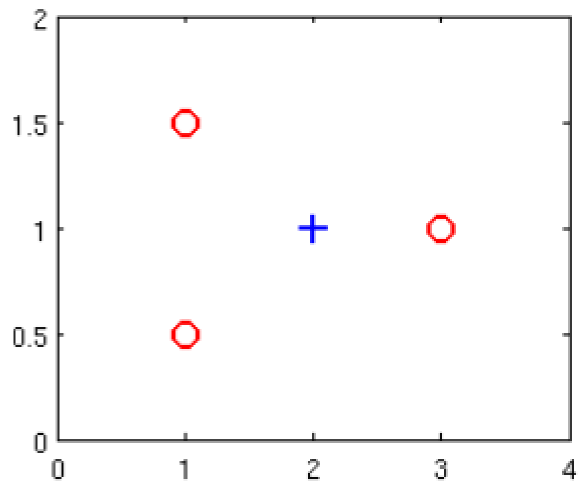
Logistic Regression – 1

- Suppose that you have trained a logistic regression classifier, and it outputs on a new example x a prediction $h_\theta(x) = 0.2$. This means (check all that apply):
- Our estimate for $P(y=1|x; \theta)$ is 0.2
- Our estimate for $P(y=0|x; \theta)$ is 0.8
- Our estimate for $P(y=1|x; \theta)$ is 0.8
- Our estimate for $P(y=0|x; \theta)$ is 0.2

Logistic Regression – 2

- Suppose you have the following training set, and fit a logistic regression classifier
- $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

X1	X2	Y
1	0.5	0
1	1.5	0
2	1	1
3	1	0



Logistic Regression – 2

Which of the following are true? Check all that apply.

- Adding polynomial features (e.g., instead using
  - $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1 x_2 + \theta_5 x_2^2)$
- At the optimal value of θ we will have $\mathrm{J}(\theta) \geq 0$.
- Adding polynomial features (like in (a)), would increase J(θ) because we are now summing over more terms.
- IF we train gradient descent for enough iterations, for some examples in the training set it is possible to obtain $h_\theta\left(x^{(i)}\right) > 1$.

Logistic Regression – 3

Which of the following is a correct gradient descent update for logistic regression?

- $\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)}$
- $\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$
- $\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{1+\exp\left(-\theta^T x^{(i)}\right)} - y^{(i)}\right) x_j^{(i)}$
- $\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(\theta^T x - y^{(i)}\right) x^{(i)}$

Logistic Regression – 4

Which of the following statements are true? Check all that apply.

- The one-vs-al technique allows you to use logistic regression for problems in which each y(i) comes from a fixed, discrete set of values
- The cost function J(θ) for logistic regression trained with m>1 examples is always $\mathrm{J}(\theta) \geq 0$
- For logistics regression sometimes gradiesn descent will converge to a local minimum and hence fail to find the global minimum. This is why we prefer more advanced tecniques such as fminunc (conjuguate gradient,/BFGS/etc)
- Since we train one classifier when there are two classes, we train two classifiers when there are three classes (and we do one vs all classification)

Logistic Reg’n – 5

- Suppose you train a logistic classifies $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$.
- Suppose $\theta_0 = -6, \theta_1 =$ and $\theta_2 = 1$.
- Which of the following figures represents the decision boundary found by your classifier.

Figure:

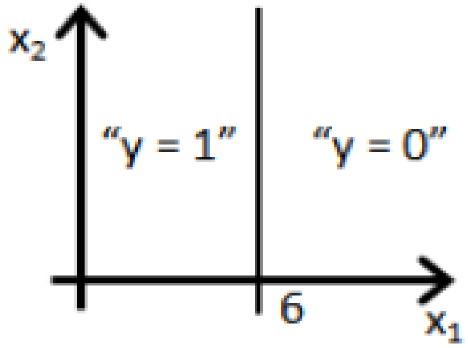


Figure:

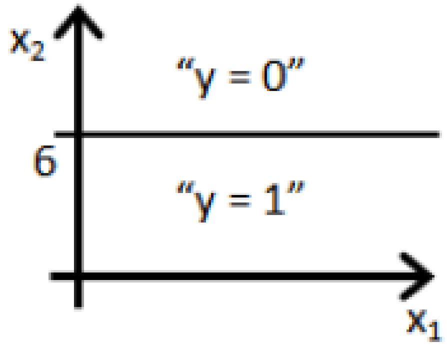


Figure:

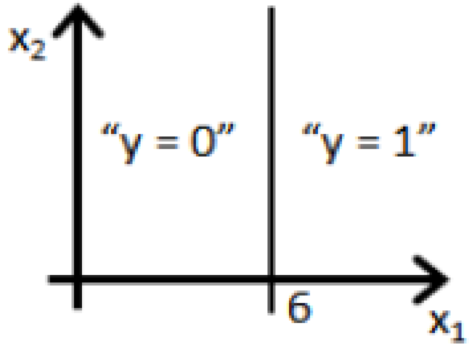
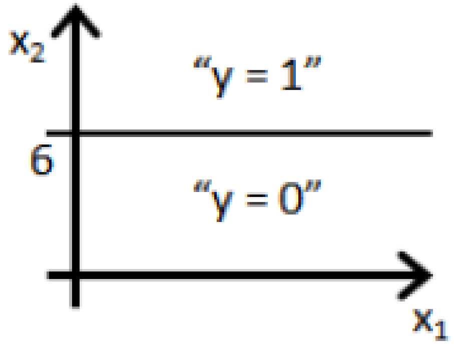


Figure:



Regularization – 1

You are training a classification model with logistic regression. Which of the following statements are true? Check all that apply.

- Adding a new feature to the model always results in equal or better performance on examples not in the training set.
- Introducing regularization to the model always results in equal or better performance on the training set.
- Adding many new features to the model makes it more likely to overfit the training set.
- Introducing regularization to the model always results in equal or better performance on examples not in the training set.

Regularization – 2

Which of the following statements about regularization are true? Check all that apply

- Because logistic regression outputs values $0 \le h_\theta(x) \le 1$, its range of output values can only be "shrunk" slightly by regularization anyway, so regularization is generally not helpful for it.
- Consider a classification problem. Adding regularization may cause your classifier to incorrectly classify some training examples (which it had correctly classified when not using regularization, i.e. when $\lambda=0$).
- Using too large a value of λ can cause your hypothesis to overfit the data; this can be avoided by reducing λ
- Using a very large value of λ cannot hurt the performance of your hypothesis; the only reason we do not set λ to be too large is to avoid numerical problems.

Regularization – 3