The Coders' Club

Machine Learning: G1

Week 3: Assignment

Topics:

- Logistic Regression
- Regularization

Some Additional Courses:

- Machine Learning Onramp (MathWorks)
  https://www.mathworks.com/learn/tutorials/machine-learning-onramp.html

- Deep Learning Onramp (MathWorks)
  https://www.mathworks.com/learn/tutorials/deep-learning-onramp.html

- AI From the Data Center to the Edge – An Optimized Path Using Intel® Architecture (Intel AI)
  https://software.intel.com/en-us/ai/courses/data-center-to-edge

- Machine Learning (Intel)
  https://software.intel.com/en-us/ai/courses/machine-learning

- Deep Learning (Intel)
  https://software.intel.com/en-us/ai/courses/deep-learning

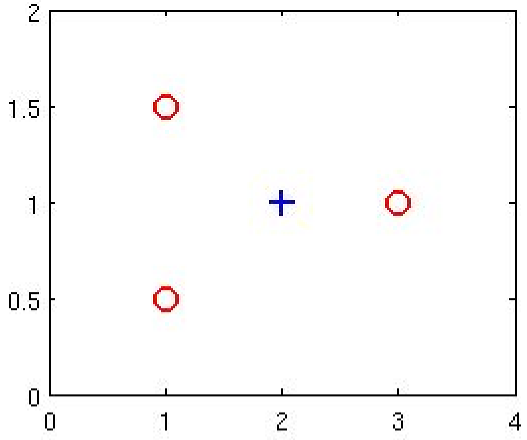
Logistic Regression

Q.1. Suppose that you have trained a logistic regression classifier, and it outputs on a new example x, a prediction $h_\theta(x)=0.4$. This means (check all that apply)

- Our estimate for $P(y=1 \mid x;\theta)$ is 0.6
- Our estimate for $P(y=0 \mid x;\theta)$ is 0.4
- Our estimate for $P(y=0 \mid x;\theta)$ is 0.6
- Our estimate for $P(y=1 \mid x;\theta)$ is 0.4

Q.2. Suppose you have the following training set, and fit a logistic regression classifier $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

x_1	x_2	y
1	0.5	0
1	1.5	0
2	1	1
3	1	0



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

- Adding polynomial features (e.g., instead using $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1 x_2 + \theta_5 x_2^2)$) could increase how well we can fit the training data.
- J(θ) will be a convex function, so gradient descent should converge to the global minimum.
- The positive and negative examples cannot be separated using a straight line. So, gradient descent will fail to converge.
- Because the positive and negative examples cannot be separated using a straight line, linear regression will perform as well as logistic regression on this data.

Q.3. For logistic regression, the gradient is given by

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)}.$$

Which of these is a correct gradient descent update for logistic regression with a learning rate of α? Check all that apply.

- $\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{1+e^{-\theta^T x^{(i)}}} - y^{(i)} \right) x_j^{(i)}$ (simultaneously update for all j).
- $\theta := \theta - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(\theta^T x - y^{(i)} \right) x^{(i)}$.
- $\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)}$ (simultaneously update for all j).
- $\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x^{(i)}$ (simultaneously update for all j).

Q.4. Which of the following statements are true? Check all that apply.

- Linear regression always works well for classification if you classify by using a threshold on the prediction made by the linear regression.

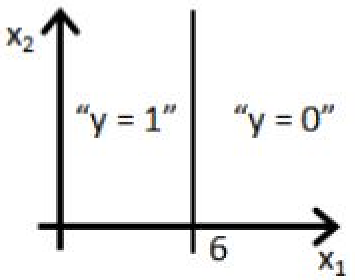
- The sigmoid function g(z) = $\frac{1}{1+e^{-z}}$ is never greater than 1.

- For logistic regression, sometimes gradient descent will converge to a local minimum (and fail to find the global minimum). This is the reason we prefer more optimized advanced algorithms such as fminunc (conjugate gradient/BFGS/L-BFGS/etc.)

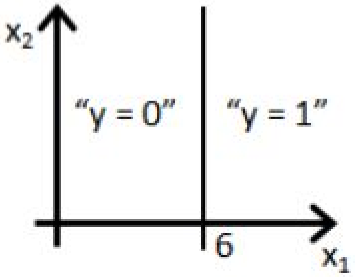
- The cost function J(θ) for logistic regression trained with m ≥ 1 examples is always greater than or equal to zero.

Q.5. Suppose you train a logistic classifier $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$. Suppose $\theta_0 = 6$, $\theta_1 = -1$, $\theta_2 = 0$. Which of the following figures represents the decision boundary found by your classifier?

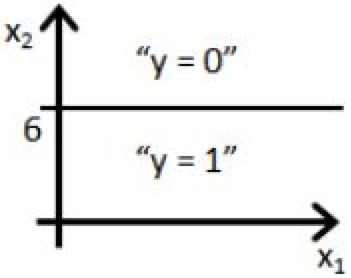
A)



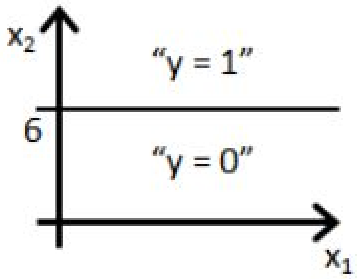
B)



C)



D)



Regularization

Q.1. You are training a classification model with logistic regression. Which of the following statements are true? Check all that apply.

- Introducing regularization to the model always results in equal or better performance on the training set.
- Introducing regularization to the model always results in equal or better performance on examples not in the training set.
- Adding a new feature to the model always results in equal or better performance on the training set.
- Adding many new features to the model helps prevent overfitting on the training set.

Q.2. Suppose you ran logistic regression twice, once with λ=0, and once with λ=1.
One of the times, you got

$$\theta = \begin{bmatrix} 74.81 \\ 45.05 \end{bmatrix}$$

And the other time you got

$$\theta = \begin{bmatrix} 1.37 \\ 0.41 \end{bmatrix}$$

However, you forgot which value of λ corresponds to which value of θ. Which one do you think corresponds to λ=1?

- $\theta = \begin{bmatrix} 74.81 \\ 45.05 \end{bmatrix}$
- $\theta = \begin{bmatrix} 1.37 \\ 0.41 \end{bmatrix}$

Q.3. Which of the following statements about regularization are true? Check all that apply.

- 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 λ=0).

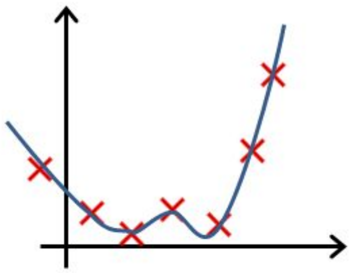
- Using a very large value of λ cannot hurt the performance of your hypothesis; the only reason we do not set λ to be too large is to avoid numerical problems.

- 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.

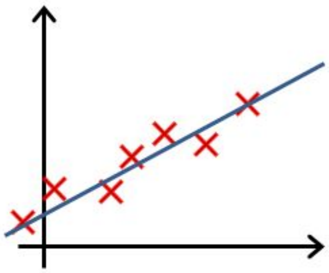
- Using too large a value of λ can cause your hypothesis to overfit the data; this can be avoided by reducing λ.

Q.4. In which of the following figures do you think the hypothesis has overfit the training set?

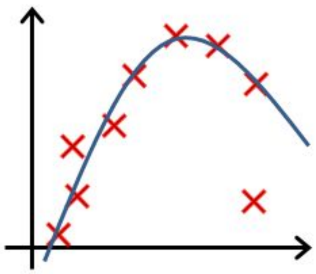
A)



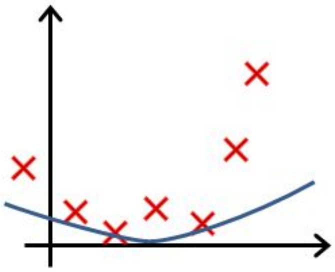
B)



C)

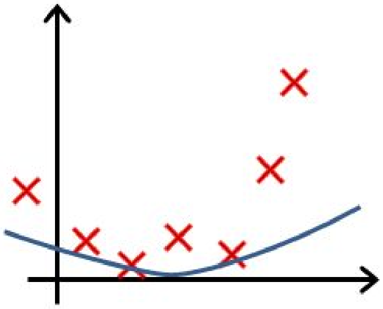


D)

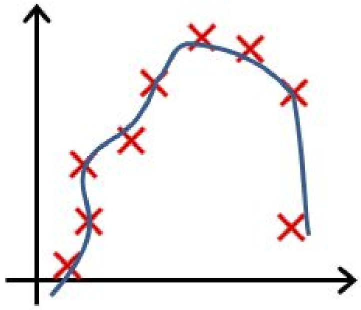


Q.5. In which one of the following figures do you think the hypothesis has underfit the training set?

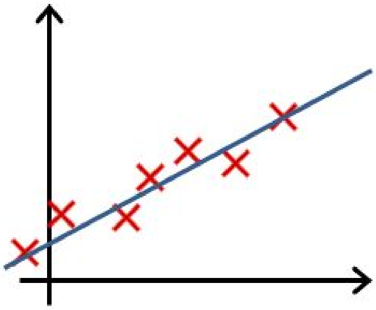
A)



B)



C)



D)