The Coders' Club

Machine Learning: G1

Week 6: Assignment

Topics:
Advice for Applying Machine Learning
Machine Learning System Design

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

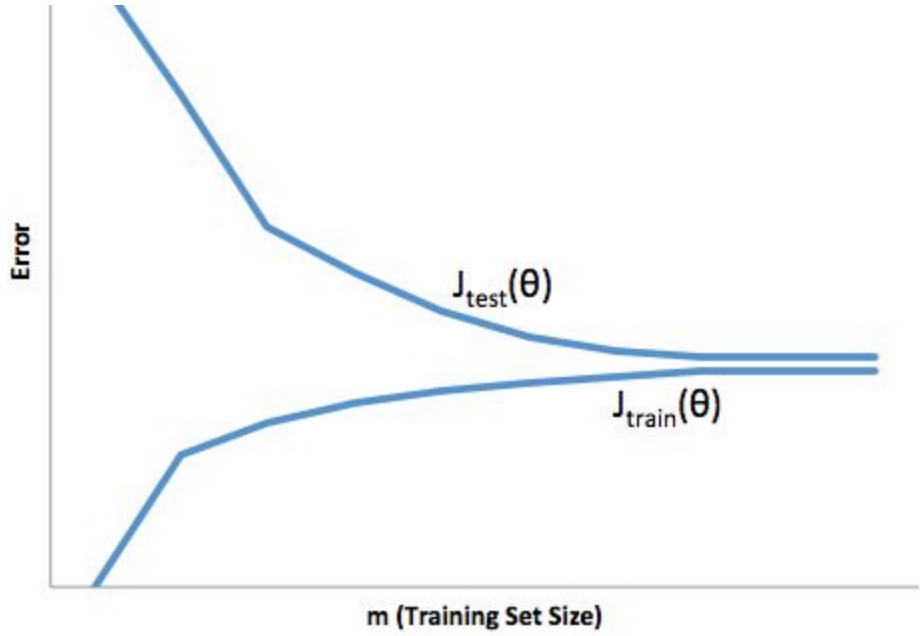
- AI From the Data Center to the Edge – An Optimized Path Using Intel® Architecture (Intel AI)
  https://software.intel.com/en-us/ai/courses/data-center-to-edge

- 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

Advice for Applying Machine Learning

Q.1. You train a learning algorithm, and find that it has unacceptably high error on the test set. You plot the learning curve, and obtain the figure below. Is the algorithm suffering from high bias, high variance, or neither?



- High bias
- High variance
- Neither

Q.2. Suppose you have implemented regularized logistic regression to classify what object is in an image (i.e., to do object recognition). However, when you test your hypothesis on a new set of images, you find that it makes unacceptably large errors with its predictions on the new images. However, your hypothesis performs **well** (has low error) on the training set. Which of the following are promising steps to take? Check all that apply.

- Get more training examples.
- Use fewer training examples.
- Trying using a smaller set of features.
- Try adding polynomial features.

Q.3. Suppose you have implemented regularized logistic regression to predict what items customers will purchase on a web shopping site. However, when you test your hypothesis on a new set of customers, you find that it makes unacceptably large errors in its predictions. Furthermore, the hypothesis performs **poorly** on the training set. Which of the following might be promising steps to take? Check all that apply.

- Try evaluating the hypothesis on a cross validation set rather than the test set.
- Try adding polynomial features.
- Try decreasing the regularization parameter λ.
- Use fewer training examples.

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

- The performance of a learning algorithm on the training set will typically be better than its performance on the test set.

- Suppose you are training a regularized linear regression model. The recommended way to choose what value of regularization parameter λ to use is to choose the value of λ which gives the lowest **cross validation** error.

- Suppose you are training a regularized linear regression model. The recommended way to choose what value of regularization parameter λ to use is to choose the value of λ which gives the lowest **test set** error.

- Suppose you are training a regularized linear regression model. The recommended way to choose what value of regularization parameter λ to use is to choose the value of λ which gives the lowest **training set** error.

Q.5. Which of the following statements are true? Check all that apply.

- A model with more parameters is more prone to overfitting and typically has higher variance.

- When debugging learning algorithms, it is useful to plot a learning curve to understand if there is a high bias or high variance problem.

- If a learning algorithm is suffering from high bias, only adding more training examples may **not** improve the test error significantly.

- If a neural network has much lower training error than test error, then adding more layers will help bring the test error down because we can fit the test set better.

Machine Learning System Design

Q.1. You are working on a spam classification system using regularized logistic regression. "Spam" is a positive class (y = 1) and "not spam" is the negative class (y = 0). You have trained your classifier and there are m=1000 examples in the cross-validation set.
The chart of predicted class vs. actual class is:

	Actual Class: 1	Actual Class: 0
Predicted Class: 1	85	890
Predicted Class: 0	15	10

For reference:

- Accuracy = (true positives + true negatives) / (total examples)
- Precision = (true positives) / (true positives + false positives)
- Recall = (true positives) / (true positives + false negatives)
- F_1 score = (2 * precision * recall) / (precision + recall)

What is the classifier's precision (as a value from 0 to 1)?

Q.2. Suppose a massive dataset is available for training a learning algorithm. Training on a lot of data is likely to give good performance when two of the following conditions hold true. Which are the two?

- The features x contain sufficient information to predict y accurately. (For example, one way to verify this is if a human expert on the domain can confidently predict y when given only x).
- We train a learning algorithm with a large number of parameters (that is able to learn/represent fairly complex functions).
- We train a learning algorithm with a small number of parameters (that is thus unlikely to overfit).
- When we are willing to include high order polynomial features of x (such as x_1^2, x_2^2, x_1x_2, etc.).

Q.3. Suppose you have trained a logistic regression classifier which is outputting $h_\theta(x)$. Currently, you predict 1 if $h_\theta(x) \geq$ threshold, and predict 0 if $h_\theta(x) <$ threshold, where currently the threshold is set to 0.5.
Suppose you **increase** the threshold to 0.7. Which of the following are true? Check all that apply.

- The classifier is likely to have unchanged precision and recall, but lower accuracy.
- The classifier is likely to now have lower precision.
- The classifier is likely to have unchanged precision and recall, but higher accuracy.
- The classifier is likely to now have lower recall.

Q.4. Suppose you are working on a spam classifier, where spam emails are positive examples (y=1) and non-spam emails are negative examples (y=0). You have a training set of emails in which 99% of the emails are non-spam and the other 1% is spam. Which of the following statements are true? Check all that apply.

- A good classifier should have both high precision and high recall on the cross validation set.
- If you always predict non-spam (output y=0), your classifier will have an accuracy of 99%.
- If you always predict non-spam (output y=0), your classifier will have 99% accuracy on the training set, but it will do much worse on the cross validation set because it has overfit the training data.
- If you always predict non-spam (output y=0), your classifier will have 99% accuracy on the training set, and it will likely perform similarly on the cross validation set.

Q.5. Which of the following statements are true? Check all that apply.

- It is a good idea to spend a lot of time collecting a **large** amount of data before building your first version of a learning algorithm.

- After training a logistic regression classifier, you **must** use 0.5 as your threshold for predicting whether an example is positive or negative.

- The “error analysis” process of manually examining the examples which your algorithm got wrong can help suggest what are good steps to take. (e.g. developing new features to improve your algorithm’s performance).

- If your model is underfitting the training set, then obtaining more data is likely to help.

- Using a **very large** training set makes it unlikely for a model to overfit the training data.