6주차 - Evaluating a Learning Algorithm

만약 나의 ML 알고리즘이 예측을 잘 못하면 어떻게 해야할까?
일반적으로 다음과 같은 방법을 임의로 취할 수 있다.

```
* Getting more training examples
* Trying smaller sets of features
* Trying additional features
* Trying polynomial features
* Increasing or decreasing λ
```

하지만, 상황에 따라서 위 경우 중 한 가지만 한번 해보는 데에도 6개월이 걸릴 수 있을 만큼 시간소모가 크다.
따라서 ML 알고리즘을 평가하는 방법을 이용하여 유용한 방법을 택하는 과정을 통해 그 시간소모를 크게 줄이는 것이 좋다.

Evaluating a Hypothesis

우선 overfitting이 발견되는 경우, 주어진 데이터를 2부분(training set, test set)으로 나눌 수 있다. 먼저, training set으로 모델을 학습시킨 뒤에 test set으로 새로운 비용함수를 구함으로서 알고리즘을 평가할 수 있다. (일반적으로 training : test = 7 : 3)
단, data set을 나눌 때는 우선 전체 데이터를 random하게 shuffle한 뒤에 시도한다.

```
1. Learn Θθ and minimize Jtrain(θ) using the training set
2. Compute the test set error Jtest(θ)
```

- test set error

선형 회귀의 경우,
훈련된 Theta로 단순히 test set에 대해서 비용함수를 구하면 test set error를 구할 수있다.

1. For linear regression: $J_{test}(\Theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} (h_\Theta(x_{test}^{(i)}) - y_{test}^{(i)})^2$

로지스틱 회귀의 경우,
선형 회귀와 마찬가지로 test set에 대해서 기존 비용함수를 적용하여 구한 값이 test set error가 되지만,
대안적 방법으로 misclassification error를 계산하는 방법이 있다. (좀 더 해석이 용이하다는 장점이 있다)

1. 기존 비용함수를 이용한 test set error

$$J_{test}(\theta) = -\frac{1}{m_{test}} \sum_{i=1}^{m_{test}} y_{test}^{(i)} \log h_\theta(x_{test}^{(i)}) + (1 - y_{test}^{(i)}) \log h_\theta(x_{test}^{(i)})$$

2. misclassification error

2. For classification ~ Misclassification error (aka 0/1 misclassification error):

$$err(h_\Theta(x), y) = \begin{matrix} 1 & \text{if } h_\Theta(x) \geq 0.5 \text{ and } y = 0 \text{ or } h_\Theta(x) < 0.5 \text{ and } y = 1 \\ 0 & otherwise \end{matrix}$$

This gives us a binary 0 or 1 error result based on a misclassification. The average test error for the test set is:

$$\text{Test Error} = \frac{1}{m_{test}} \sum_{i=1}^{m_{test}} err(h_\Theta(x_{test}^{(i)}), y_{test}^{(i)})$$

This gives us the proportion of the test data that was misclassified.

참고 - 왜 error function에는 regularization term이 없을까?
https://www.coursera.org/learn/machine-learning/discussions/weeks/6/threads/om5LldI7EeeeDgqph8eWrA

Model selection and Train/Validation/Test Sets

모델의 hypothesis를 평가하는데는 훈련 데이터와 테스트 데이터가 있으면 되지만,
여러가지 모델을 정하는 경우에는 validation이라는 데이터 셋이 따로 필요하다.
예를들어 d차원 polynomial hypothesis 중에서 모델 하나를 고른다고 생각해보자.
먼저 각 d마다 해당 모델에 대하여 Train set으로 Theta를 구하고 Validation Set으로 cross validation(CV 그냥 validation이라고도 함) error를 구하고 가장 작은 error를 낸 모델을 고른다.
그리고 그 모델에 대한 평가는 Test set의 error로 하게 된다.
Validation set를 test set으로 활용하면 안되는 이유는 model selection을 할 때 d라는 parameter가 validation set에 fit되기 때문이다.
그렇게 fitting된 모델에 test set으로 validation set을 사용하면 공정하지 않은 평가가 나오게 된다.
주로 총 데이터에서 각 data set이 가지는 비율은 Train : Validation : Test = 6 : 2 : 2이다.