Machine Learning Lecture by Andrew Ng on Coursera

Week 1

Introduction

- What is Machine Learning?
- Supervised Learning
- Unsupervised Learning

Linear Regression with One Variable

- Model (Hypothesis)
- Cost Function
- Gradient Descent

Week 2

Linear Regression with Multiple Variable

- Multiple Features
- Feature Scaling
- Learning Rate
- Features and Polymomial Regression
- Computing Parameters Analytically - Normal Equation

Week 3

Logistic Regression

- Classification
- Logistic Hypothesis
- Decision Boundary
- Cost Function
- Simplified Cost Function and Gradient Descent
- Advanced Optimization
- Multiclass Classification - One-vs-all

Regularization

- The Problem of Overfitting
- Cost Function
- Regularized Linear Regression

Week 4

Neural Networks: Representation

- Non-linear Hypothesis
- Neurons and the Brain
- **Model Representation** $x = a^{(1)} \rightarrow g(a^{(1)} * \Theta^{(1)}) = a^{(2)} \rightarrow g(a^{(2)} * \Theta^{(2)}) = a^{(3)} \rightarrow ... \rightarrow g(a^{(l-1)} * \Theta^{(l-1)}) = a^{(l)} = h_{\Theta}(x)$
- Example - AND, ! AND !, OR -> XOR
- Example - Multclass Classification

Week 5

Neural Networks: Learning

- Cost Function - NN multiclass classification
- Backpropagation

Backpropagation in Practice

- Unrolling Parameters
- Gradient Checking
- Random Initialization
- Putting It Together

Week 6

Advice for Applying Machine Learning

- ***Not-satisfiying result! -> What to Try Next?***
  - More training examples
  - Trying smaller sets of features
  - Trying additional sets of features
  - Trying polynomial features
  - Increasing or decreasing λ

Training set / Cross Validation set / Test set

- Evaluating a hypothesis with a separate test set
  - Check overfit, generalization
  - train:test = 70%:30%
- Model selection with another separate cross validation set
  - Compare different models (# of features, degree of polynomial, and λ)
  - train:cv:test = 60%:20%:20%

Model selection details: $J(\Theta; \lambda)$

- Number of parameters($|\Theta|$) and Bias/Variance
  - bias(underfit): $J_{CV}(\theta) \approx J_{train}(\theta) \gg 0$, not enough parameters for task
  - variance(overfit): $J_{CV}(\theta) \gg J_{train}(\theta)$, too many parameters for task
- Regualization and Bias/Variance
  - Remind contribution of λ to J, $J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\sum_{j=1}^{m}\theta_j^2$
  - $\lambda \approx 0 \implies$ `maybe overfit`
  - $\lambda \gg 0 \implies$ `maybe less overfit`
- Learning Curves, Error x ***training set size***
  - High bias, low training size: $J_{train}(\Theta)$ low and $J_{CV}(\Theta)$ high
  - High bias, large training size: both $J_{train}(\Theta)$ and $J_{CV}(\Theta)$ high
  - High variance, low training size: $J_{train}(\Theta)$ low and $J_{CV}(\Theta)$ high
  - High variance, large training size: $J_{train}(\Theta)$ OK and $J_{CV}(\Theta)$ **keep decreasing**
- Summary
  - Select best combo Θ, λ, and right amount of data by checking J_{CV}
  - And also check J_{test} for good generalization

Machine Learning System Design (with Spam classifier)

- Prioritizing What to Work On
- Error Analysis
- Error Metrics Skewed Classes
- Precision and Recall trade off

Week 7

- Support Vector Machine

Week 8

Unsupervised Learning

- Clustering
- K-Means Algorithm
- Optimization Objective
- Random Initialization
- Choosing the Number of Clusters

Dimensionality Reduction

- Motivation
- Motivation I: Data Compression
- Motivation II: Visualization

- Principal Component Analysis
- Principal Component Analysis Problem Formulation
- Principal Component Analysis Problem Algorithm
- Applying PCA
- Reconstructin from Compressed Representation
- Choosing the Number of Principal Components
- Advice for Applying PCA

Week 9

Anomaly Detection

Recommender Systems

Week 10

Large Scale Machine Learning

- Gradient Descent with Large Datasets
- Learning with Large Datasets
- Stochastic Gradient Descent
- Mini-Batch Gradient Descent
- Stochastic Gradient Descent Convergence
- Advanced Topics
- Online Learning
- Map Reduce and Data Parallelism

Week 11

Application Example: Photo OCR

- Photo OCR
- Problem Description and Pipeline
- Sliding Windows
- Getting Lots of Data and Artificial Data
- Ceiling Analysis: What Part of the Pipeline to Work on Next

Conclusion

- Summary and Thank You