การอบรม Machine Learning for Data Analytics ณ ห้องประชุม 501 ชั้น 5 สมาคมประกันวินาศภัยไทย (30 สิงหาคม 2562)
Instructor

• ผศ.ดร.สุรณพีร์ ภูมิวุฒิสาร
  (ผู้อำนวยการบัณฑิตศึกษา คณะวิทยาการและเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีมahanakhon)

• Education
  • วศบ. วิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย
  • M.Sc., Information Science, University of Pittsburgh, USA
  • Ph.D Computer Science and Engineering, University of New South Wales, Australia

• Teaching Experience
  • Machine Learning
  • Mathematics for Research

• Contact
  • suronape@mut.ac.th
AGENDA

Lecture (9.00 – 10.30)
- Machine Learning Overview

Lecture (10.45 – 12.00)
- Applied Analytics through Case Studies

Lecture (13.00 – 16.00)
- Demo – Insurance Case
- Model Assessment
Big Data Analytics

- We are entering the era of **big data**
- Big Data Analytics create “**Business Value**” in today’s world (better decision)

**Recent Big Success Story**

Google: $109.65B, Facebook: $40.65B, Amazon: $177.87B, Netflix: $11.69B

Analytics is for everyone whether “Big or Not”

Picture: Internet

Picture: [https://www.promptcloud.com/blog/datafication-era-of-big-data](https://www.promptcloud.com/blog/datafication-era-of-big-data)
Big Data Analytics

- The process of discovering actionable insights through the analysis from the data”

**Example – NETFLIX**

- **What is data?**  ข้อมูลการรับชมภาพยนตร์ของสมาชิก
- **What is insight?**  ภาพยนตร์เรื่องไหน ประเภทใด ที่เป็นที่สนใจที่ผู้ชม
- **What is action?**  ระบบแนะนำโปรแกรมหนัง/การสร้างภาพยนตร์
Types of Big Data Analytics

https://www.ecapitaladvisors.com/blog/analytics-maturity/analytics-maturity/
Descriptive Analytics

• “What happened?”

• “Describe” or “Summarize” raw data and make it interpretable by humans.

• Example
  • Google Analytics

Picture: https://www.itopclass.com/blog/google-analytics-%E0%B8%84%E0%B8%97%E0%B8%A5%E0%B8%B7%E0%B8%99%E0%B8%97%E0%B9%86
Diagnostic Analytics

• Why did it happen?

• Underline why something happened in the past and help locate the root cause of the problem

• Example
  • factors influence medical expenses for insurance company

Which factors influence medical expenses charged per year?

> str(insurance)
'data.frame': 1333 obs. of 7 variables:
$ age : int 19 19 29 32 31 46 37 37 60 ...
$ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 1 ...
$ bmi : num 27.9 33.6 33 22.7 28.9 25.7 35.4 27.7 28.8 25.8 ...
$ children: int 0 1 3 0 0 1 3 2 0 ...
$ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 ...
$ region : Factor w/ 4 levels "northeast","northwest",...
$ expenses: num 16855 1726 4449 21984 3867 ...
Predictive Analytics

• “What will happen?”

• Make predictions about the future

• Example
  • Southwest Airlines

What might happen in the future

For example, Southwest Airlines analyses sensor data on their planes in order to identify patterns that indicate a potential malfunction, thus allowing the airlines to the necessary repairs before its schedule.

Picture: https://www.edureka.co/blog/big-data-analytics/
Prescriptive Analytics

- “What should we do about this?”

- Uses optimization and simulation algorithms to advice on possible outcomes

- Example
  - Google Maps – best route
Cognition/ Self Learning Analytics

• What don’t I know?

• Cognition is all about thinking, understanding, learning and remembering. Cognitive computing is all about creating analytics technology that mimics the human brain's ability to perform these functions.

• Example
  • Google’s self-driving car

Picture: adapt from https://www.edureka.co/blog/big-data-analytics/
Data Analytics Process

- **Data collection**: gather data an algorithm will use
- **Data preparation**: prepare the data
- **Model Building**: chose a learning algorithm to represent the data in the form of a model.
- **Model evaluation**: evaluate the algorithm learned
- **Model improvement**: improve a performance of the model

Source: Applied Analytics through case studies
What is Machine Learning?

“the field of study that gives computers the ability to learn without being explicitly programmed” Arthur Samuel (1959)
How does the Machine Learn?

- How Machine Learns?

  - Training Data

  - Input → Model \( (h) \) → Output

- Housing Price Prediction

  Choose algorithm to build the model

  Picture: internet
Types of Machine Learning

• **Supervised learning** (h = predictive model)
  • "correct answers" given
    • Regression problems (predict continuous output)
    • Classification problems (predict discrete output)

• **Unsupervised learning** (h = descriptive model)
  • "no correct answers" given
Classification or Regression problems?

• You’re running a company, and you want to develop learning algorithms to address each of two problems
  • **Problem 1**: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
  
  • **Problem 2**: You’d like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.
## Recent Success Stories

### Examples

<table>
<thead>
<tr>
<th>Application</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Estate</td>
<td>Size, #Bedrooms, Location</td>
<td>Price of the House</td>
</tr>
<tr>
<td>Spam Classification</td>
<td>Keywords</td>
<td>Spam/not Spam</td>
</tr>
<tr>
<td>Self-Driving Car</td>
<td>Image, Radar Info.</td>
<td>Position of other Cars</td>
</tr>
<tr>
<td>Face Recognition</td>
<td>Image</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>Audio</td>
<td>Text</td>
</tr>
<tr>
<td>Game Application</td>
<td>User Profile, Ad Info.</td>
<td>Click/ not Click</td>
</tr>
</tbody>
</table>

Source: Google
Types of Machine Learning (con’t)
The Limits of Machine Learning

• Machine learning, at this time, is not in any way a substitute for a human brain.
  • little flexibility to extrapolate outside of the strict parameters it learned and knows no common sense

Source: Google Twitter
**Common Notation**

\[ x = \text{input variable/features } x = \{x_1, x_2, \ldots, x_n\} \]

- Numeric – year, price, mileage
- Categorical (or Nominal) – model, color, transmission

\[ y = \text{output variable/target feature } y = \{y_1, y_2, \ldots, y_n\} \]

Training examples = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}

\[ m = \text{number of training examples} \]

<table>
<thead>
<tr>
<th>Size</th>
<th>House Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1,000,000</td>
</tr>
<tr>
<td>200</td>
<td>2,000,000</td>
</tr>
<tr>
<td>300</td>
<td>3,000,000</td>
</tr>
</tbody>
</table>

**Features**

<table>
<thead>
<tr>
<th>year</th>
<th>model</th>
<th>price</th>
<th>mileage</th>
<th>color</th>
<th>transmission</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>SEL</td>
<td>21992</td>
<td>7413</td>
<td>Yellow</td>
<td>AUTO</td>
</tr>
<tr>
<td>2011</td>
<td>SEL</td>
<td>20995</td>
<td>10926</td>
<td>Gray</td>
<td>AUTO</td>
</tr>
<tr>
<td>2011</td>
<td>SEL</td>
<td>19995</td>
<td>7351</td>
<td>Silver</td>
<td>AUTO</td>
</tr>
<tr>
<td>2011</td>
<td>SEL</td>
<td>17809</td>
<td>11613</td>
<td>Gray</td>
<td>AUTO</td>
</tr>
<tr>
<td>2012</td>
<td>SE</td>
<td>17500</td>
<td>8367</td>
<td>White</td>
<td>MANUAL</td>
</tr>
<tr>
<td>2010</td>
<td>SEL</td>
<td>17495</td>
<td>25125</td>
<td>Silver</td>
<td>AUTO</td>
</tr>
<tr>
<td>2011</td>
<td>SEL</td>
<td>17000</td>
<td>27393</td>
<td>Blue</td>
<td>AUTO</td>
</tr>
<tr>
<td>2010</td>
<td>SEL</td>
<td>16995</td>
<td>21026</td>
<td>Silver</td>
<td>AUTO</td>
</tr>
<tr>
<td>2011</td>
<td>SES</td>
<td>16995</td>
<td>32655</td>
<td>Silver</td>
<td>AUTO</td>
</tr>
</tbody>
</table>
AGENDA

Lecture (9.00 – 10.30)
- Machine Learning Overview

Lecture (10.45 – 12.00)
- Applied Analytics through Case Studies

Lecture (13.00 – 16.00)
- Demo – Insurance Case
- Model Assessment
Case Studies – Supervised (Regression)

- Understanding Regression
  - specifying the relationship between one or more numeric independent variables (the predictors) and a single numeric dependent variable (the value to be predicted)

- More example
  - Predict housing price based on width and #bedroom
Linear Regression

- **Hypothesis Function**
  \[ h_\theta(x) = \theta_0 + \theta_1 x \]

- **Cost Function**
  - Measure the accuracy of the hypothesis function
  \[ J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \]
Linear Regression (2)

• Choose $\theta_0, \theta_1$ to minimize the cost function

• Parameter Learning - Gradient Descent
  - Start with some parameters, for example $\theta_0, \theta_1$
  - Keep changing $\theta_0, \theta_1$ until we end up at a minimum

• Gradient Descent Algorithm
  Repeat until convergence{
  
  \[ \theta_i := \theta_i - \alpha \frac{\partial J(\theta_i)}{\partial \theta_i} \]

  } 

\[ \alpha = \text{learning rate} \]

Source: Andrew Ng Coursera
Example – Regression Problem (How much/How many?)

• Predict medical expense each person charged to the insurance plan for the year
  กลุ่มตัวอย่าง 1,338 ตัวอย่าง
tัวแปรต้น 6 ตัว ตัวแปรตามที่เราสนใจจะทำนาย 1 ตัว (expenses)

```r
> str(insurance)
'data.frame': 1338 obs. of 7 variables:
$ age : int 19 18 28 33 37 31 10 39 50 ...
$ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
$ bmi : num 27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...
$ children: int 0 1 3 0 0 0 1 3 2 0 ...
$ smoker: Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
$ region: Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2$
$ expenses: num 16885 1726 4449 21984 3867 ...
```

• Prediction Model
  • \( \text{expenses} = 139.0053 - 32.6181 \times \text{age} + 3.7307 \times \text{age}^2 + 678.6017 \times \text{children} + 119.7715 \times \text{BMI} - 496.7690 \times \text{sexmale} - 997.9355 \times \text{BMI}^30 + 13404.5952 \times \text{smokeryes} + 19810.1534 \times \text{bmi}^30:\text{smokeryes} \)
Case Studies – Supervised (Classification)

• Predict whether a student gets admitted or rejected into a university

• Predict whether a microship pass a quality test
Example – Classification Problem (A or B or..)

- Diagnosing breast cancer by determining whether the mass is likely to be malignant or benign.

กลุ่มตัวอย่าง 569 ตัวอย่าง
ตัวแปรตัว 31 ตัว ตัวแปรตามที่เราสนใจจะทำนาย 1 ตัว (diagnosis)

### Code Snippet

```r
> wbcd <- read.csv("wisc_bc_data.csv", stringsAsFactors = FALSE)
> str(wbcd)
'data.frame': 569 obs. of 32 variables:
  $ id : Int 87139402 8910251 905520 868871 9012568 906539 925291 87880 86 2989 89827 ... 
  $ diagnosis : chr "B" "B" "B" "B" ... 
  $ radius_mean : num 12.3 10.6 11 11.3 15.2 ... 
  $ texture_mean : num 12.4 18.9 16.8 13.4 13.2 ... 
  ... 
```

(Credit จาก demographics statistics from the US Census Bureau)
K-Nearest Neighbor Classification (KNN)

- Uses an example's k-nearest neighbors to classify unlabeled examples.

- Blind Tasting Experience Example

<table>
<thead>
<tr>
<th>Ingredient</th>
<th>Sweetness</th>
<th>Crunchiness</th>
<th>Food type</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple</td>
<td>10</td>
<td>9</td>
<td>fruit</td>
</tr>
<tr>
<td>bacon</td>
<td>1</td>
<td>4</td>
<td>protein</td>
</tr>
<tr>
<td>banana</td>
<td>10</td>
<td>1</td>
<td>fruit</td>
</tr>
<tr>
<td>carrot</td>
<td>7</td>
<td>10</td>
<td>vegetable</td>
</tr>
<tr>
<td>celery</td>
<td>3</td>
<td>10</td>
<td>vegetable</td>
</tr>
<tr>
<td>cheese</td>
<td>1</td>
<td>1</td>
<td>protein</td>
</tr>
</tbody>
</table>

Source: Machine Learning with R
K-Nearest Neighbor Classification (KNN) (2)

- Is tomato a fruit or vegetable?

- Let's calculate the distance between the tomato (sweetness = 6, crunchiness = 4), and the green bean (sweetness = 3, crunchiness = 7).

\[
\text{dist(tomato, green bean)} = \sqrt{(6 - 3)^2 + (4 - 7)^2} = 4.2
\]

<table>
<thead>
<tr>
<th>Ingredient</th>
<th>Sweetness</th>
<th>Crunchiness</th>
<th>Food Type</th>
<th>Distance to the tomato</th>
</tr>
</thead>
<tbody>
<tr>
<td>grape</td>
<td>8</td>
<td>5</td>
<td>fruit</td>
<td>sqrt((6 - 8)^2 + (4 - 5)^2) = 2.2</td>
</tr>
<tr>
<td>green bean</td>
<td>3</td>
<td>7</td>
<td>vegetable</td>
<td>sqrt((6 - 3)^2 + (4 - 7)^2) = 4.2</td>
</tr>
<tr>
<td>nuts</td>
<td>3</td>
<td>6</td>
<td>protein</td>
<td>sqrt((6 - 3)^2 + (4 - 6)^2) = 3.6</td>
</tr>
<tr>
<td>orange</td>
<td>7</td>
<td>3</td>
<td>fruit</td>
<td>sqrt((6 - 7)^2 + (4 - 3)^2) = 1.4</td>
</tr>
</tbody>
</table>
K-NN and ML Algorithms

• Classification algorithms using nearest neighbor are considered lazy learning algorithms

  • Heavily reliance on the training instances rather than an abstracted model (also known as instance-based learning)

  • no parameters are learned about the data (non-parametric learning methods)

```r
> wc <- read.csv("wisc_bc_data.csv", stringsAsFactors = FALSE)
> str(wc)
'data.frame': 569 obs. of 32 variables:
$ id : int 87139402 8910251 905520 868871 9012568 906530 9235291 87880 86
2989 89827 ... 
$ diagnosis : chr "B" "B" "B" "B" ... 
$ radius_mean : num 12.3 10.6 11 11.3 15.2 ... 
$ texture_mean : num 12.4 18.9 16.8 13.4 13.2 ... 
$ perimeter_mean : num 78.8 69.3 70.9 73 97.7 ... 
$ area_mean : num 464 346 373 385 712 ... 
$ smoothness_mean : num 0.1028 0.0969 0.1077 0.1164 0.0796 ... 
$ compactness_mean : num 0.0698 0.1147 0.078 0.1336 0.0693 ... 
$ concavity_mean : num 0.0399 0.0639 0.0305 0.0464 0.0339 ... 
$ symmetry_mean : num 0.196 0.192 0.171 0.177 0.172 ... 
$ dimension_mean : num 0.0595 0.0649 0.0634 0.0607 0.0554 ... 
```

### Cell Contents

- N / Row Total
- N / Col Total
- N / Table Total

**Total Observations in Table:** 100

<table>
<thead>
<tr>
<th>wc_test_labels</th>
<th>Benign</th>
<th>Malignant</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign</td>
<td>61</td>
<td>0</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>1.000</td>
<td>0.000</td>
<td>0.610</td>
</tr>
<tr>
<td></td>
<td>0.968</td>
<td>0.000</td>
<td>0.610</td>
</tr>
<tr>
<td></td>
<td>0.610</td>
<td>0.000</td>
<td>0.610</td>
</tr>
<tr>
<td>Malignant</td>
<td>2</td>
<td>37</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>0.051</td>
<td>0.949</td>
<td>0.390</td>
</tr>
<tr>
<td></td>
<td>0.032</td>
<td>1.000</td>
<td>0.370</td>
</tr>
<tr>
<td></td>
<td>0.020</td>
<td>0.370</td>
<td>0.370</td>
</tr>
<tr>
<td>Column Total</td>
<td>63</td>
<td>37</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>0.630</td>
<td>0.370</td>
<td>0.370</td>
</tr>
</tbody>
</table>

**Heatmap Diagram:**

- Green: Malignant
- Blue: Benign
Case Studies - Unsupervised Learning

• Discover hidden patterns or a structures of data.
  • "no correct answers" - no target to learn
  • no single feature is more important than any other

• Example
  • Teen market segments
    • เป็นการจัดกลุ่มวัยรุ่นที่ใช้งาน Social Media ตามลักษณะที่คล้ายๆกัน เช่น อาชีพ จำนวนเพื่อนความชอบส่วนตัว
Organizing a Data Science Conference

• To facilitate professional networking and collaboration, you planned to seat people in groups according to one of three research specialties: computer and/or database science, math and statistics, and machine learning.

Source: Machine Learning with R
Clustering

• Clustering is an unsupervised machine learning task that automatically divides the data into **clusters**, or groups of similar items.

• Clustering is guided by the principle that items inside a cluster should be very similar to each other.

• Clustering is used for knowledge discovery rather than prediction.
  • It provides an insight into the natural groupings found within data.
K-means Clustering

- **K-means clustering** partition training examples into a pre-specified number of clusters.

- **K-means** consists of 2 phases
  - Cluster assignment
    - assign each example to one of the \( K \) clusters
  - Updates cluster centroids
    - adjusting centroids of clusters

* The process of updating and assigning occurs several times until changes no longer improve the cluster fit.
Iteration 1

• Cluster assignment

Euclidean distance $d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$
Iteration 1

• Update cluster centroids

Centroid (red) = \left( \frac{1+3+5+6+11}{5}, \frac{7+6+8+6+5}{5} \right)
Iteration 2

• The same process of iteration 1
  • Since no change in cluster assignment occurs, K-means stops at iteration 2.

Cluster assignment

Update cluster centroids
AGENDA

Lecture (9.00 – 10.30)
- Machine Learning Overview

Lecture (10.45 – 12.00)
- Applied Analytics through Case Studies

Lecture (13.00 – 16.00)
- Demo – Insurance Case
- Model Assessment
Insurance Case Study

- Which factors influence medical expenses charged per year?
- How many expenses should the insurance company should charge individual customers per year?

```r
> str(insurance)
'data.frame': 1338 obs. of 7 variables:
$ age : int 19 18 28 33 32 31 46 37 37 60 ...
$ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...  
$ bmi : num 27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...  
$ children: int 0 1 3 0 0 0 1 3 2 0 ...
$ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...  
$ region : Factor w/ 4 levels "northeast","northwest",...: 4 3 3 2 2 3 3 2 1 2$
$ expenses: num 16885 1726 4449 21984 3867 ... 
```
Evaluating Model Performance

• Evaluate how well a learned hypothesis can be expected to perform on new training examples
  • A hypothesis may have a low error for the training examples but still be inaccurate to predict output
  • A hypothesis may not even have a low error for the training examples and therefore be inaccurate to predict output

Bias versus Variance

- Every model has bias and variance error components
  - **Bias** component comes from erroneous assumptions of chosen learning algorithms (underfitting problems)
  - **Variance** component comes from sensitivity to change in a model (overfitting problems)

- **Bias and Variance Trade-off**
  - Bias and Variance are inversely related to each other
    - Balancing between both components is an ideal solution (low bias and low variance)
Example

- **Regression Problems**

\[ h_{\theta}(x) = \theta_0 + \theta_1 x \]

\[ h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 \]

\[ h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 \]

- **Classification Problems**
Evaluating a Model

• Splitting training examples into 2 parts – train and test datasets
  • randomly 70-30 or 80-20 into train and test dataset
Splitting Training Examples into 2 Parts

- The hypothesis is fit on the training set, and its performance is evaluated on the test set:
  - Learn $\theta$ and minimize $J_{\text{train}}(\theta)$ using a training set
  - Compute test error $J_{\text{test}}(\theta)$
    - For linear regression
      \[
      J_{\text{test}}(\theta) = \frac{1}{m_{\text{test}}} \sum_{i=1}^{m_{\text{test}}} (h_{\theta}(x_{\text{test}}^{(i)}) - y_{\text{test}}^{(i)})^2
      \]
    - For KNN
      - simple approximately misclassification error
        \[
        \frac{1}{n} \sum_{i=1}^{n} \text{Err}_i \quad \text{Err}_i = I(y_i \neq \hat{y}_i)
        \]
Problems with Splitting Data into 2 parts

• Consider the case of underfitting
• Model Selection
  • select the appropriate level of flexibility for a model

\[ h_\delta(x) = \theta_0 + \theta_1 x \]
\[ h_\delta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 \]
\[ h_\delta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 \]
\[ h_\delta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 \]
\[ \vdots \]
\[ h_\delta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \ldots + \theta_{10} x^{10} \]

How does the model generalize?
Evaluating a Model

• Splitting training examples into 3 parts – train, cross validation and test datasets (Machine Learning World – not for statistics!)
  • randomly 60-20-20 or 50-25-25 into train, cross validation and test dataset
Splitting Training Examples into 3 parts

- The hypothesis is fit on the training set, and its performance is evaluated on the test set:
  - Optimize parameters \( \theta \) using training set for each polynomial degree
    - Goal: \( \min J(\theta) = J_{\text{train}}(\theta) = \frac{1}{m_{\text{train}}} \sum_{i=1}^{m_{\text{train}}} (h_\theta(x_{\text{train}}^{(i)}) - y_{\text{train}}^{(i)})^2 \)
  - Find polynomial degree \( d \) with the least error using cross validation set
    \( J_{\text{cv}}(\theta) = \frac{1}{m_{\text{cv}}} \sum_{i=1}^{m_{\text{cv}}} (h_\theta(x_{\text{cv}}^{(i)}) - y_{\text{cv}}^{(i)})^2 \)
  - Estimate the generalization error using test set wit \( J_{\text{test}}(\theta^{(i)}), \theta^{(i)} = \theta \text{ from polynomial with lowest error} \)
    \( J_{\text{test}}(\theta) = \frac{1}{m_{\text{test}}} \sum_{i=1}^{m_{\text{test}}} (h_\theta(x_{\text{test}}^{(i)}) - y_{\text{test}}^{(i)})^2 \)
Model Improvement

• Advanced strategies may be needed to improve a model’s performance.

• Some Strategies
  • Collect more examples
  • Try smaller set of features
  • Try adding relevant features
  • Try higher degree of features
  • Increasing regularization parameter
  • decreasing regularization parameter

-> **Goal**: perform a test to gain insight what is/isn’t work with a learning algorithms so as to correctly improve the model’s performance

-> **Diagnose Bias or Variance**
Bias or Variance Problem?

- Suppose our learning hypothesis is bad predictions: bias or variance problem?
  - High bias (underfitting):
    - both $J_{\text{train}}(\theta)$ and $J_{\text{cv}}(\theta)$ will be high

- High variance (overfitting):
  - $J_{\text{train}}(\theta)$ will be low and $J_{\text{cv}}(\theta)$ will be high 
    (muchmore than $J_{\text{train}}(\theta)$)

Source: Figure by Andrew Ng
Overcome Overfitting

• Reduce a number of features
  • Select only relevant features
  • Decrease degree of polynomial

• Regularization
  • Keep all features (each contributes a bit to predict output) but constrains or regularizes the parameters to small values (many cases towards zero but not zero).
  • Regularized Linear Regression
    • Cost Function

\[
j(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2
\]
การตรวจสอบ Bias vs Variance ด้วย Learning Curve

- Learning Curve เป็นกราฟที่เราใช้สำหรับตรวจสอบ Model ว่ามีลักษณะ High Bias หรือ High Variance โดยจะทำการ plot cost ที่เกิดขึ้นใน train dataset (error_train) และ cross validation dataset (error_cv) ตามจำนวน training example ที่เพิ่มขึ้น

Model ที่ดีที่ Balance เรื่อง Bias (low) และ Variance (low)
Learning Curve – High Bias

• ถ้าเส้น error train และ cross validation มีค่าสูงทั้งคู่ Model จะมีลักษณะ High Bias เพราะ จาก error ที่เกิดขึ้น แสดงว่า Model ไม่สามารถถอดิบายข้อมูลทั้งที่นำมาทดสอบและข้อมูลกลุ่มตัวอย่างใหม่ได้
Learning Curve – High Variance

• ถ้าเส้น error จาก train มีค่าต่ำ และ error จาก cross validation มีค่าสูง Model จะมีลักษณะ High Variance เพราะ จาก error ที่เกิดขึ้น แสดงว่า Model จำรูปแบบ training examples มากเกินไป ทำให้อธิบายข้อมูลกลุ่มตัวอย่างใหม่ไม่ดี
วิธีการในการ Improve Model

<table>
<thead>
<tr>
<th>การเพิ่มจำนวน training examples</th>
<th>Fix High Bias</th>
<th>Fix High Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>การเพิ่มจำนวน input features</td>
<td>⭐</td>
<td></td>
</tr>
<tr>
<td>การเพิ่ม degree of polynomial</td>
<td>⭐</td>
<td></td>
</tr>
<tr>
<td>การลดจำนวน input features</td>
<td>⭐</td>
<td>⭐</td>
</tr>
<tr>
<td>การเพิ่มค่า regularized parameters</td>
<td>⭐</td>
<td>⭐</td>
</tr>
<tr>
<td>การลดค่า regularized parameters</td>
<td>⭐</td>
<td></td>
</tr>
</tbody>
</table>
Machine Learning Algorithms

Deep Learning
- Deep Boltzmann Machine (DBM)
- Deep Belief Networks (DBN)
- Convolutional Neural Network (CNN)
- Stacked Auto-Encoders

Ensemble
- Random Forest
- Gradient Boosting Machines (GBM)
- Bootstrapped Aggregation (Bagging)
- AdaBoost
- Stacked Generalization (Blending)
- Gradient Boosted Regression Trees (GBRT)
- Radial Basis Function Network (RBFN)

Neural Networks
- Perceptron
- Back-Propagation
- Hopfield Network
- Ridge Regression
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Elastic Net
- Least Angle Regression (LARS)
- Cubist
- One Rule (OneR)
- Zero Rule (ZeroR)
- Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
- Linear Regression
- Ordinary Least Squares Regression (OLSR)
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)
- Logistic Regression

Regularization

Bayesian
- Naive Bayes
- Averaged One-Dependence Estimators (AODE)
- Bayesian Belief Network (BBN)
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Bayesian Network (BN)

Decision Tree
- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5
- C5.0
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- Conditional Decision Trees
- MS

Principal Component Analysis (PCA)
- Partial Least Squares Regression (PLSR)
- Sammon Mapping
- Multidimensional Scaling (MDS)
- Projection Pursuit
- Principal Component Regression (PCR)
- Partial Least Squares Discriminant Analysis
- Mixture Discriminant Analysis (MDA)
- Quadratic Discriminant Analysis (QDA)
- Regularized Discriminant Analysis (RDA)
- Flexible Discriminant Analysis (FDA)
- Linear Discriminant Analysis (LDA)

k-Nearest Neighbour (kNN)
- k-Nearest Neighbours (k-NN)
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)
- Locally Weighted Learning (LWL)

Clustering
- k-Means
- k-Medians
- Expectation Maximization
- Hierarchical Clustering
Programming for Data Analytics and Demo

- R Programming Language
  - A collection of R functions that can be shared among users is called a package
  - View a list at Comprehensive R Archive Network (CRAN) http://cran.r-project.org/index.html
- Python Programming Language
- Others programming platforms such as SAS

Source: Applied Analytics through case studies

SURONAPEE PHOOMVUTHISARN, Ph.D
How to Start a Project?

• Challenges
  • Successful businesses will be defined by their ability to collect and curate the right data and apply analytics in order to make insights actionable at the point of decision and make better decisions.
Top 10 Skills ที่ Data Scientist ต้องรู้

- Source: http://www.msit.mut.ac.th
Thank you

• Question?