Practices when Applying Machine Learning

(วิธีปฏิบัติเมื่อประยุกต์ใช้ Machine Learning)

Krittameth Teachasrisaksakul

krittameth.teacha@gmail.com

บทนำ

Machine learning เป็นหนึ่งในแขนงวิชาที่ หลักการปฏิบัติทั่วไป (rules of thumb) มีแนวโน้มปรากฎออกมาก่อนหลักการทางทฤษฎี

บทเรียนนี้ เกี่ยวกับหลักการทางทฤษฎีที่สำคัญที่สุดบางข้อ ที่เป็นรากฐาน หรืออธิบายวิธีปฏิบัติ (practical heuristics):

- Bias-variance trade-off (การต้องเลือกอย่างใดอย่างหนึ่ง, ได้อย่างหนึ่งก็ต้องเสียอีกอย่างหนึ่ง)
- ชุดข้อมูล training / cross-validation / test เกี่ยวข้องกันอย่างไร?
- เมื่อเราใช้โมเดลที่มี variance สูงที่มีแนวโน้มเกิด overfitting จะควบคุมมันได้อย่างไร?

Debug (แก้ปัญหา,จุดบกพร่องในโปรแกรม) learning algorithm

สมมติ เรา implement regularized linear regression เพื่อทำนายราคาบ้าน ดังนี้:

$$J(\theta) = \frac{1}{2m}\left[\underbrace{\Sigma_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2}_{} + \underbrace{\lambda\Sigma_{j=1}^{m}\theta_j^2}_{}\right]$$

แต่เมื่อเราทดสอบ hypothesis กับข้อมูลบ้านชุดใหม่ จะพบว่า hypothesis มี error สูงมากเมื่อทำนายค่าเราควรทำอย่างไรต่อไป?

Debug (แก้ปัญหา,จุดบกพร่องในโปรแกรม) learning algorithm

สมมติ เรา implement regularized linear regression เพื่อทำนายราคาบ้าน ดังนี้:

$$J(\theta) = \frac{1}{2m}\left[\Sigma_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\Sigma_{j=1}^{m}\theta_j^2\right]$$

แต่เมื่อเราทดสอบ hypothesis กับข้อมูลบ้านชุดใหม่ จะพบว่า hypothesis มี error สูงมากเมื่อทำนายค่าเราควรทำอย่างไรต่อไป?

- เก็บตัวอย่างข้อมูล training เพิ่ม
- ลองใช้ชุด features ที่เล็กลง

Debug (แก้ปัญหา,จุดบกพร่องในโปรแกรม) learning algorithm

สมมติ เรา implement regularized linear regression เพื่อทำนายราคาบ้าน ดังนี้:

$$J(\theta) = \frac{1}{2m}\left[\Sigma_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\Sigma_{j=1}^{m}\theta_j^2\right]$$

แต่เมื่อเราทดสอบ hypothesis กับข้อมูลบ้านชุดใหม่ จะพบว่า hypothesis มี error สูงมากเมื่อทำนายค่าเราควรทำอย่างไรต่อไป?

- เก็บตัวอย่างข้อมูล training เพิ่ม
- ลองใช้ชุด features ที่เล็กลง
- ลองเพิ่ม features ใหม่
- ลองเพิ่ม polynomial features (เช่น x_1^2, x_2^2, x_1x_2 เป็นต้น)

Debug (แก้ปัญหา,จุดบกพร่องในโปรแกรม) learning algorithm

สมมติ เรา implement regularized linear regression เพื่อทำนายราคาบ้าน ดังนี้:

$$J(\theta) = \frac{1}{2m}\left[\Sigma_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\Sigma_{j=1}^{m}\theta_j^2\right]$$

แต่เมื่อเราทดสอบ hypothesis กับข้อมูลบ้านชุดใหม่ จะพบว่า hypothesis มี error สูงมากเมื่อทำนายค่าเราควรทำอย่างไรต่อไป?

- เก็บตัวอย่างข้อมูล training เพิ่ม
- ลองใช้ชุด features ที่เล็กลง
- ลองเพิ่ม features ใหม่
- ลองเพิ่ม polynomial features (เช่น x_1^2, x_2^2, x_1x_2 เป็นต้น)
- ลองลดค่า λ
- ลองเพิ่มค่า λ

Machine Learning Diagnostic

Diagnostic: คือ การทดสอบที่เราสามารถ run เพื่อหา insight (ความเข้าใจลึกซึ้ง) ว่าสิ่งไหนที่ได้ผล / ไม่ได้ผลกับ learning algorithm และหา guidance (คำแนะนำ / แนวทาง) ว่า เพิ่ม performance / สมรรถภาพ ของมันได้ดีที่สุดอย่างไร

Diagnostics ต้องใช้เวลาเพื่อ implement แต่เป็นการใช้เวลาที่ดี

Question

ข้อใดต่อไปนี้เกี่ยวกับ diagnostics เป็นจริง / ถูกต้อง ? (วงทุกข้อที่ถูก)

(i) มันยากที่จะบอกว่าอะไรจะได้ผลดี เพื่อปรับปรุง learning algorithm ดังนั้น วิธีที่ดีที่สุด คือ ทำตาม gut feeling (ลางสังหรณ์/สัญชาตญาณ) และลองดูว่าวิธีไหนได้ผล

(ii) Diagnostics สามารถให้ guidance (คำแนะนำ/แนวทาง) ว่าสิ่งใดอาจได้ผลมากกว่า ที่จะลองปรับปรุง learning algorithm

(iii) Diagnostics อาจใช้เวลามากที่จะ implement และลองใช้ แต่เป็นการใช้เวลาที่ดี

(iv) Diagnostic บางครั้งสามารถบอกได้ว่า การกระทำใด (การเปลี่ยนแปลง learning algorithm) ที่น่าจะไม่ทำให้ performance ของมันดีขึ้นมาก

Question

ข้อใดต่อไปนี้เกี่ยวกับ diagnostics เป็นจริง / ถูกต้อง ? (วงทุกข้อที่ถูก)

(i) มันยากที่จะบอกว่าอะไรจะได้ผลดี เพื่อปรับปรุง learning algorithm ดังนั้น วิธีที่ดีที่สุด คือ ทำตาม gut feeling (ลางสังหรณ์/สัญชาตญาณ) และลองดูว่าวิธีไหนได้ผล

(ii) Diagnostics สามารถให้ guidance (คำแนะนำ/แนวทาง) ว่าสิ่งใดอาจได้ผลมากกว่า ที่จะลองปรับปรุง learning algorithm

(iii) Diagnostics อาจใช้เวลามากที่จะ implement และลองใช้ แต่เป็นการใช้เวลาที่ดี

(iv) Diagnostic บางครั้งสามารถบอกได้ว่า การกระทำใด (การเปลี่ยนแปลง learning algorithm) ที่น่าจะไม่ทำให้ performance ของมันดีขึ้นมาก

วิธีปฏิบัติเมื่อประยุกต์ใช้ Machine Learning

Evaluating a Hypothesis

(การประเมินผล Hypothesis)

Krittameth Teachasrisaksakul

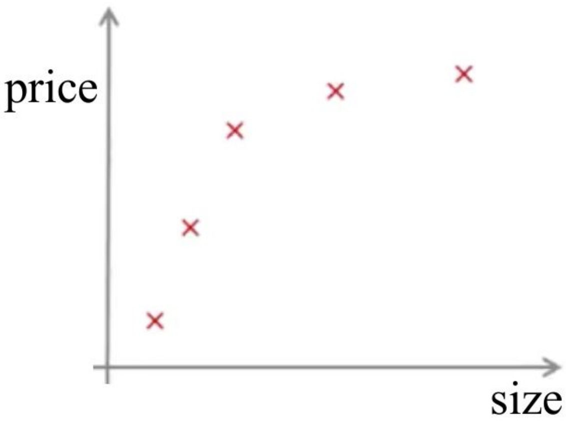
krittameth.teacha@gmail.com

แรงจูงใจ : Motivation

การประเมินผล

Hypothesis:

เมื่อเรา fit parameter ของ learning algorithm ⟶ เราจะเลือก parameter เพื่อ minimize training error (ทำให้ training error น้อยที่สุด)

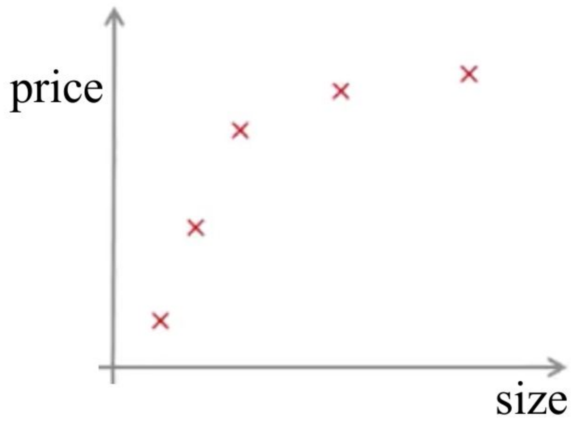


$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

แรงจูงใจ : Motivation

การประเมินผล **Hypothesis:**

เมื่อเรา fit parameter ของ learning algorithm ⟶ เราจะเลือก parameter เพื่อ minimize training error (ทำให้ training error น้อยที่สุด)



$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

ดังนั้น ล้มเหลวที่จะ generalize กับตัวอย่างใหม่ที่ไม่อยู่ในชุดข้อมูล training set

โดยมาก มี features จำนวนมาก เช่น

- x_1 = ขนาดพื้นที่บ้าน
- x_2 = จำนวนห้องนอน
- x_3 = จำนวนชั้น
- x_4 = อายุบ้าน
- x_5 = รายได้เฉลี่ยของเพื่อนบ้าน
- x_6 = ขนาดพื้นที่ห้องครัว

การประเมินผล Hypothesis

ชุดข้อมูลของเรา

Size	Price
2104	400
1600	330
2400	369
1416	232
3000	540
1985	300
1534	315
1427	199
1380	212
1494	243

โดยแบ่ง training examples เป็น 2 กลุ่ม:
(a) Training set: 70% ของข้อมูลทั้งหมด
(b) Test set: 30% ที่เหลือ

การประเมินผล Hypothesis

ชุดข้อมูลของเรา

	Size	Price	
70%	2104	400	Training set
	1600	330	
	2400	369	
	1416	232	
	3000	540	
	1985	300	
	1534	315	
30%	1427	199	Testing set
	1380	212	
	1494	243	

Training set $\longrightarrow$

$$(x^{(1)}, y^{(1)})$$
$$(x^{(2)}, y^{(2)})$$
$$\vdots$$
$$(x^{(m)}, y^{(m)})$$

Testing set $\longrightarrow$

$$(x^{(1)}_{test}, y^{(1)}_{test})$$
$$(x^{(2)}_{test}, y^{(2)}_{test})$$
$$\vdots$$
$$(x^{(m_{test})}_{test}, y^{(m_{test})}_{test})$$

Question

สมมติ implementation ของ linear regression (ที่ไม่ใช้ regularization) เกิด overfitting อย่างมากเมื่อ train กับชุดข้อมูล training set ในกรณีนี้ เราจะคาดว่าอะไรจะเกิดขึ้น

(i) training error $J(\theta)$ ต่ำ test error $J_{\text{test}}(\theta)$ ต่ำ

(ii) training error $J(\theta)$ ต่ำ test error $J_{\text{test}}(\theta)$ สูง

(iii) training error $J(\theta)$ สูง test error $J_{\text{test}}(\theta)$ ต่ำ

(iv) training error $J(\theta)$ สูง test error $J_{\text{test}}(\theta)$ สูง

Question

สมมติ implementation ของ linear regression (ที่ไม่ใช้ regularization) เกิด overfitting อย่างมากเมื่อ train กับชุดข้อมูล training set ในกรณีนี้ เราจะคาดว่าอะไรจะเกิดขึ้น

(i) training error $J(\theta)$ ต่ำ test error $J_{\text{test}}(\theta)$ ต่ำ

(ii) training error $J(\theta)$ ต่ำ test error $J_{\text{test}}(\theta)$ สูง

(iii) training error $J(\theta)$ สูง test error $J_{\text{test}}(\theta)$ ต่ำ

(iv) training error $J(\theta)$ สูง test error $J_{\text{test}}(\theta)$ สูง

Guidance สำหรับ Linear Regression

กระบวนการ Training / Testing:

- เรียนรู้ parameter θ จาก ข้อมูล training หรือ 70% ของมัน (เพื่อทำให้ training error $J(\theta)$ น้อยที่สุด)
- คำนวณ test set error:

$$J_{\textbf{test}}(\theta) = \underbrace{\frac{1}{2m_{\textbf{test}}}\Sigma_{i=1}^{m_{\textbf{test}}}\left(h_\theta(x^{(i)}_{\textbf{test}}) - y^{(i)}_{\textbf{test}}\right)^2}_{\text{Mean squared error}}$$

Guidance สำหรับ Logistic Regression

กระบวนการ Training / Testing:

- เรียนรู้ parameter θ จาก ข้อมูล training หรือ 70% ของมัน (เพื่อทำให้ training error $J(\theta)$ น้อยที่สุด)
- คำนวณ test set error:

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

Guidance สำหรับ Logistic Regression

กระบวนการ Training / Testing:

- เรียนรู้ parameter θ จาก ข้อมูล training หรือ 70% ของมัน (เพื่อทำให้ training error $J(\theta)$ น้อยที่สุด)
- คำนวณ test set error:
- หรือ ใช้เทคนิค 0/1 misclassification error

ทำนาย ไม่ตรงกับ ข้อมูล output จริง ⟶ นับ 1

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

บอกสัดส่วนของ test data ที่ถูกแยกประเภทผิด (misclassified)

$$\boxed{\text{Test error} = \frac{1}{m_{\text{test}}} \Sigma_{i=1}^{m_{\text{test}}} \text{err}\left(h_\theta(x_{\text{test}}^{(i)}), y_{\text{test}}^{(i)}\right)}$$

วิธีปฏิบัติเมื่อประยุกต์ใช้ Machine Learning

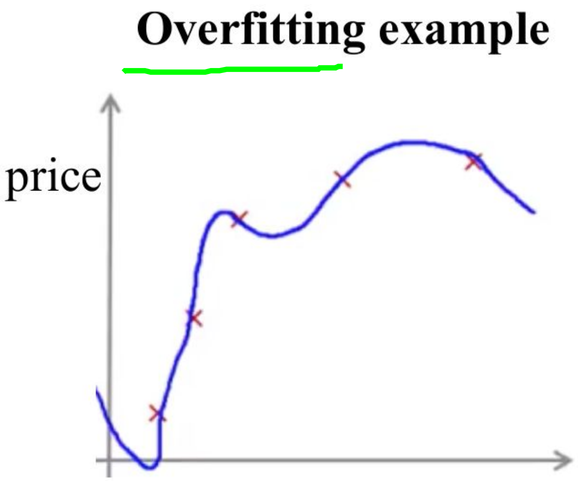
การเลือก Model และชุดข้อมูล

Training / Validation / Test Set

Krittameth Teachasrisaksakul

krittameth.teacha@gmail.com

การเลือก Model : ความเข้าใจพื้นฐาน



$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

เมื่อ parameter $\theta_0, \theta_1, ..., \theta_4$ ถูก fit กับชุดข้อมูล training set

training error $J(\theta)$ (ความผิดพลาดของชุดข้อมูล training set) ⟶ มีแนวโน้มจะต่ำกว่า generalization error (ความผิดพลาดของชุดข้อมูลอื่นๆ)

แม้จะ fit กับข้อมูล training set ได้ดี แต่อาจ overfit และ ทำให้ค่าที่ทำนาย (prediction) ของ test set แย่ / ผิดเยอะ / ไม่แม่นยำ

Figure source: Andrew Ng's Machine Learning course

ปัญหาการเลือก Model (Model Selection Problem)

สมมติ อยากเลือก degree ของพหุนาม (polynomial) ที่จะใช้กับข้อมูล

$$
\begin{array}{lll}
1. \quad h_\theta(x) = \theta_0 + \theta_1 x & \longrightarrow \Theta^{(1)} & \longrightarrow J_{\text{test}}(\Theta^{(1)}) \\
2. \quad h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 & \longrightarrow \Theta^{(2)} & \longrightarrow J_{\text{test}}(\Theta^{(2)}) \\
3. \quad h_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_3 x^3 & \longrightarrow \Theta^{(3)} & \longrightarrow J_{\text{test}}(\Theta^{(3)}) \\
\quad \vdots & & \quad \vdots \\
10. \quad h_\theta(x) = \theta_0 + \theta_1 x + \theta_{10} x^{10} & \longrightarrow \Theta^{(10)} & \longrightarrow J_{\text{test}}(\Theta^{(10)})
\end{array}
$$

วิธีเลือก model ที่ดีที่สุด : ลองใช้ polynomial degree แต่ละค่า ⟶ ดู error (ความผิดพลาด)

ปัญหาการเลือก Model (Model Selection Problem)

สมมติ อยากเลือก degree ของพหุนาม (polynomial) ที่จะใช้กับข้อมูล

$$
\begin{array}{lll}
1. \quad h_\theta(x) = \theta_0 + \theta_1 x & \longrightarrow \Theta^{(1)} & \longrightarrow J_{\text{test}}(\Theta^{(1)}) \\
2. \quad h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 & \longrightarrow \Theta^{(2)} & \longrightarrow J_{\text{test}}(\Theta^{(2)}) \\
3. \quad h_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_3 x^3 & \longrightarrow \Theta^{(3)} & \longrightarrow J_{\text{test}}(\Theta^{(3)}) \\
\quad \vdots & & \quad \vdots \\
10. \quad h_\theta(x) = \theta_0 + \theta_1 x + \theta_{10} x^{10} & \longrightarrow \Theta^{(10)} & \longrightarrow J_{\text{test}}(\Theta^{(10)})
\end{array}
$$

สมมติ เลือก $\theta_0 + \ldots + \theta_5 x^5$

นี่ดูเหมือนจะเหมาะสม แต่มันน่าจะเป็น optimistic estimate (ค่าประมาณ) ของ generalization error

ก็คือ degree ที่ถูกเลือกของ polynomial ถูกใช้กับชุดข้อมูล test set

ปัญหาการเลือก Model (Model Selection Problem)

ชุดข้อมูลของเรา

วิธีเลือก model ที่ดีที่สุด : แบ่งชุดข้อมูลเป็น 3 ชุดย่อย (set)

	Size	Price	
60%	2104	400	Training set
	1600	330	
	2400	369	
	1416	232	
	3000	540	
	1985	300	
20%	1534	315	Cross validation set (CV)
	1427	199	
20%	1380	212	Test set
	1494	243	

$$(x^{(1)}, y^{(1)})$$
$$(x^{(2)}, y^{(2)})$$
$$\vdots$$
$$(x^{(m)}, y^{(m)})$$

$$(x_{cv}^{(1)}, y_{cv}^{(1)})$$
$$(x_{cv}^{(2)}, y_{cv}^{(2)})$$
$$\vdots$$
$$(x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})$$

$$(x_{test}^{(1)}, y_{test}^{(1)})$$
$$(x_{test}^{(2)}, y_{test}^{(2)})$$
$$\vdots$$
$$(x_{test}^{(m_{test})}, y_{test}^{(m_{test})})$$

Train / Validation / Test Error

คำนวณ error (ความผิดพลาด) ของ 3 set (ชุดข้อมูลย่อย) ที่ต่างกัน

Training error:

$$J_{\textbf{train}}(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Cross validation error:

$$J_{\textbf{cv}}(\theta) = \frac{1}{2m_{\textbf{cv}}}\Sigma_{i=1}^{m_{\textbf{cv}}}\left(h_\theta(x_{\textbf{cv}}^{(i)}) - y_{\textbf{cv}}^{(i)}\right)^2$$

Test error:

$$J_{\textbf{test}}(\theta) = \frac{1}{2m_{\textbf{test}}}\Sigma_{i=1}^{m_{\textbf{test}}}\left(h_\theta(x_{\textbf{test}}^{(i)}) - y_{\textbf{test}}^{(i)}\right)^2$$

ปัญหาการเลือก Model (Model Selection Problem)

สมมติ เราอยากเลือก degree ของพหุนาม (polynomial) ที่จะใช้กับข้อมูล

(1) (2)

$$
\begin{array}{llllll}
1. & h_\theta(x) = \theta_0 + \theta_1 x & \longrightarrow \min\limits_\theta J(\theta) & \longrightarrow \Theta^{(1)} & \longrightarrow & J_{\mathbf{cv}}(\Theta^{(1)}) \\
2. & h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 & & \longrightarrow \Theta^{(2)} & \longrightarrow & J_{\mathbf{cv}}(\Theta^{(2)}) \\
3. & h_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_3 x^3 & & \longrightarrow \Theta^{(3)} & \longrightarrow & J_{\mathbf{cv}}(\Theta^{(3)}) \\
 & \vdots & & & & \vdots \\
10. & h_\theta(x) = \theta_0 + \theta_1 x + \theta_{10} x^{10} & & \longrightarrow \Theta^{(10)} & \longrightarrow & J_{\mathbf{cv}}(\Theta^{(10)})
\end{array}
$$

ขั้นตอน

(1) สำหรับแต่ละ polynomial degree ⟶ optimize/ปรับค่า parameters ใน Θ โดยใช้ **training set**

(2) เลือก polynomial degree d ที่มี error $J_{CV}(\Theta^{(d)})$ ต่ำสุด โดยใช้ **cross validation set**

(3) ประมาณค่า generalization error $J_{test}(\Theta^{(d)})$ โดยใช้ **test set**

เมื่อ $\Theta^{(d)} = \Theta$ ของ polynomial ที่มี error ต่ำสุด

ปัญหาการเลือก Model (Model Selection Problem)

สมมติ เราอยากเลือก degree ของพหุนาม (polynomial) ที่จะใช้กับข้อมูล

$$
\begin{array}{lllll}
1. & h_\theta(x) = \theta_0 + \theta_1 x & \longrightarrow \min\limits_\theta J(\theta) & \longrightarrow \Theta^{(1)} & \longrightarrow J_{cv}(\Theta^{(1)}) \\
2. & h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 & & \longrightarrow \Theta^{(2)} & \longrightarrow J_{cv}(\Theta^{(2)}) \\
3. & h_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_3 x^3 & & \longrightarrow \Theta^{(3)} & \longrightarrow J_{cv}(\Theta^{(3)}) \\
 & \vdots & & & \vdots \\
10. & h_\theta(x) = \theta_0 + \theta_1 x + \theta_{10} x^{10} & & \longrightarrow \Theta^{(10)} & \longrightarrow J_{cv}(\Theta^{(10)})
\end{array}
$$

สมมติ เลือก 4th order degree polynomial ก็คือ $J_{CV}(\Theta^{(4)})$

แล้ว ประมาณค่า generalization error ของ test set ก็คือ $J_{test}(\theta^{(4)})$

Question

พิจารณา กระบวนการเลือก model ที่เราเลือก degree ของ polynomial โดยใช้ชุดข้อมูล cross validation สำหรับ model สุดท้าย (ที่มี parameters θ) เราอาจคาดว่า $J_{\mathrm{CV}}(\theta)$ น้อยกว่า $J_{\mathrm{test}}(\theta)$ เพราะ

(i) degree ของ polynomial ที่ถูกเลือกถูก fit กับชุดข้อมูล cross validation set

(ii) degree ของ polynomial ที่ถูกเลือกถูก fit กับชุดข้อมูล test set

(iii) cross validation set โดยปกติจะ เล็กกว่า test set

(iv) cross validation set โดยปกติจะ ใหญ่กว่า test set

Question

พิจารณา กระบวนการเลือก model ที่เราเลือก degree ของ polynomial โดยใช้ชุดข้อมูล cross validation สำหรับ model สุดท้าย (ที่มี parameters θ) เราอาจคาดว่า $J_{\mathrm{CV}}(\theta)$ น้อยกว่า $J_{\mathrm{test}}(\theta)$ เพราะ

(i) degree ของ polynomial ที่ถูกเลือกถูก fit กับชุดข้อมูล cross validation set

(ii) degree ของ polynomial ที่ถูกเลือกถูก fit กับชุดข้อมูล test set

(iii) cross validation set โดยปกติจะ เล็กกว่า test set

(iv) cross validation set โดยปกติจะ ใหญ่กว่า test set

k-fold Cross Validation

เมื่อข้อมูลมีไม่เพียงพอ เรามักจะเลือก model ที่ทำงานได้ดีที่สุดกับข้อมูลทั้งหมด ไม่ใช่เพียง 20% หรือ 30% ของข้อมูล ดังนั้น การใช้ k-fold cross validation สมเหตุสมผลมากกว่า:

1. แบ่ง S เป็น ชุดย่อย subset k ชุด ที่ disjoint (แยกจากกัน / ไม่มีสมาชิกซ้ำกัน) S_1, ..., S_k (ค่าปกติที่พบได้บ่อยของ k คือ 5 กับ 10)
2. สำหรับแต่ละ model M_i สำหรับ $j = 1,..., k$
   a. Train (ฝึก / สร้าง) M_i โดยใช้ $S_{train} = S_1, ..., S_{j-1}, S_{j+1}, ..., S_k$ (ทุก subset ยกเว้น S_j)
   b. Test (ทดสอบ) M_i ที่ฝึกแล้ว โดยใช้ $S_{CV} = S_j$
3. เลือก M_i ที่มี average error (ความผิดพลาดเฉลี่ย) ต่ำสุด เมื่อใช้กับ S_{CV} over k folds
4. ฝึก M_i ซ้ำ โดยใช้ S ทั้งหมด

(บางครั้ง วิธีนี้ เรียกว่า 'leave-one-out' cross validation)

k-fold Cross Validation

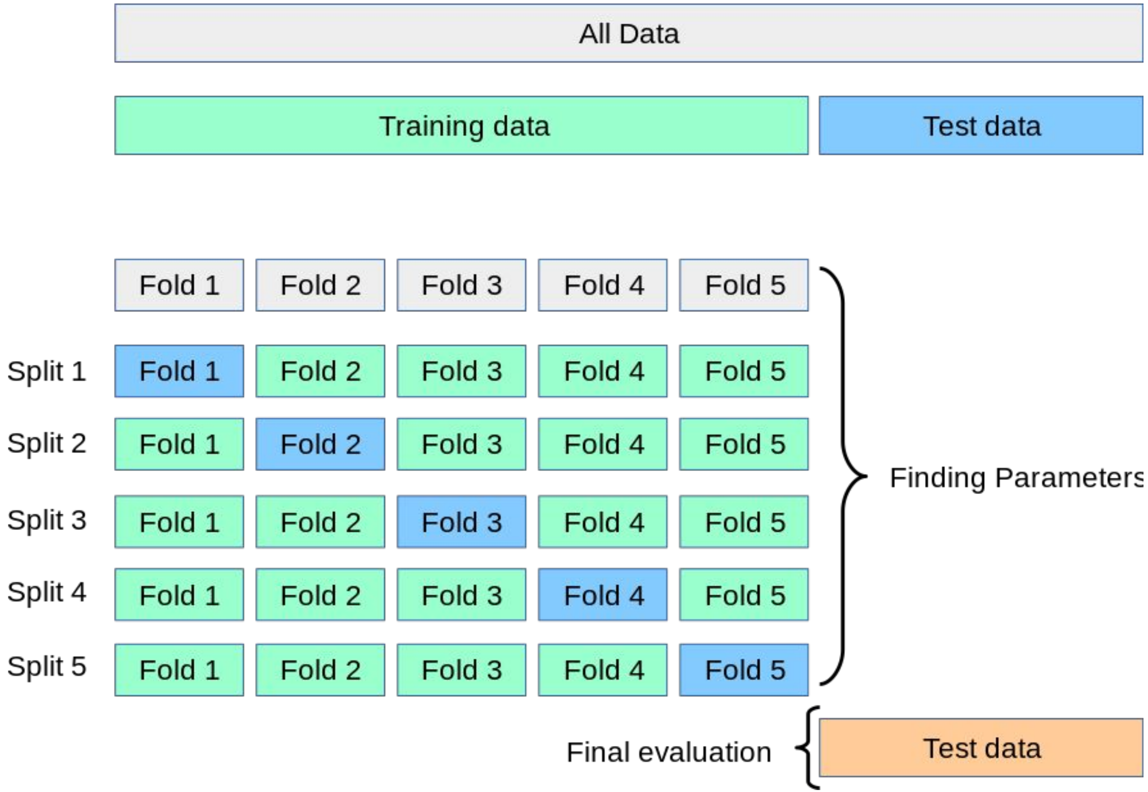


Figure source:
https://scikit-learn.org/stable/modules/cross_validation.html

k-fold Cross Validation

ระวัง ว่าแบ่งข้อมูลอย่างไร !

- โดยปกติ การแบ่งตัวอย่าง (examples) เป็น k folds เป็นแบบ random uniform
- แต่ บางครั้ง training items เกี่ยวข้องกัน (เช่น การ crop (ตัดส่วนภาพ) วัตถุเดียวกัน ในภาพเดียวกัน หลายๆครั้ง)
- การแบ่งตัวอย่างแบบสุ่ม (randomized partitioning) จะทำให้ตัวอย่างที่เกี่ยวข้องกันอยู่ใน fold ต่างกัน และทำให้เกิดประสิทธิภาพที่ประเมินค่าสูงเกินไป (overestimated performance)
- ในกรณีนี้ จำเป็นต้องจัดตัวอย่างที่เกี่ยวข้องให้อยู่ใน fold เดียวกัน

Feature Selection : การเลือก Feature

เป็นกรณีพิเศษของ model selection (การเลือก model)

สมมติ เรามี features หลายตัว เช่น $n \gg m$

feature แต่ละตัว: มี 2 ทางเลือก :
ใช้ / ไม่ใช้ feature นี้
$2 \times 2 \times \ldots \times 2$ (n ตัว) $= 2^n$

(กรณีนี้เป็นไปได้ในบางแขนง / domains เช่น การวิเคราะห์ biological sequence, ที่หาส่วน / segments ของ ลำดับ DNA (DNA sequence) ใน genome ซึ่งควบคุม biological function, หรือบางครั้งใน text classification / การแยกประเภทข้อความ)

ดังนั้น เราจะมองปัญหานี้เป็น การเลือก model ก็คือ จากกลุ่มย่อย (subset) ของ features 2^n กลุ่ม $\longrightarrow$ หา subset ที่มี cross validation performance ดีที่สุด

การลองใช้ subset ทั้งหมด 2^n กลุ่ม เป็นไปไม่ได้ ยกเว้น n มีค่าน้อยมาก ดังนั้น เราต้องการวิธีที่มีประสิทธิภาพดีกว่านี้

Feature Selection : การเลือก Feature

วิธี forward search ของ feature selection

1. ตั้งค่าเริ่มต้น (Initialize) $F := \emptyset$
2. For $j \in \{1, \ldots, \mathrm{n}\}$
    2.1. For $i \in \{1, \ldots, \mathrm{n}\} \backslash F$
        2.1.1. ให้ $F_i := F \cup \{i\}$
        2.1.2. Train (ฝึก/สร้าง) โมเดลโดยใช้ F_i กับชุดข้อมูล cross validation และหา cross validation error
3. ให้ $F := F \cup \{i\}$ เมื่อ i เป็น feature ที่มี cross validation error ต่ำสุด
4. เลือก ชุด feature (feature set) ที่มี cross validation error ต่ำสุด over all tests

Feature Selection

เมื่อมี method ที่ห่อ/wrap learning algorithm ของเรา, การส่ง feature set ที่ต่างกัน ในแต่ละ iteration เรียกว่า 'wrapper model selection'

วิธี wrapper อีกวิธี คือ backward search ซึ่งเริ่มจาก $F := \{1,2,\ldots, \mathrm{n}\}$ และทำซ้ำๆ (iteratively) ทิ้ง feature ที่ ซึ่งให้ข้อมูล (informative) น้อยสุดออกไป

Wrapper method ทำงานได้ดี แต่ใช้เวลาทำงานนาน เช่น ต้องเรียก optimization algorithm ทั้งหมด $O(kn^2)$ ครั้ง เพื่อทำ forward selection กับ k-fold cross validation

Filter Feature Selection

วิธี Filter feature selection ใช้ heuristic เพื่อเลือกว่า subset ของ features ไหนที่ควรลอง

ตัวอย่าง: อาจเรียงลำดับ features โดยใช้ตัววัดความสัมพันธ์ (relatedness) กับ output ที่ต้องการ แล้วเพิ่ม features ในลำดับนั้น ถ้ามันทำให้ generalization ดีขึ้น

ตัววัดความสัมพันธ์ (relatedness) ที่ทั่วไปที่สุดตัวหนึ่ง สำหรับ discrete features (features ที่มีค่าไม่ต่อเนื่อง) เรียกว่า 'mutual information' เขียนแทนด้วย $\mathbf{MI}(X_i, Y)$ ระหว่าง feature X_i กับ target Y:

$$\mathbf{MI}(X_i, Y) = \Sigma_{x_i \in X} \Sigma_{y_i \in Y} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)}$$

วิธีปฏิบัติเมื่อประยุกต์ใช้ Machine Learning

Diagnostic Bias vs. Variance

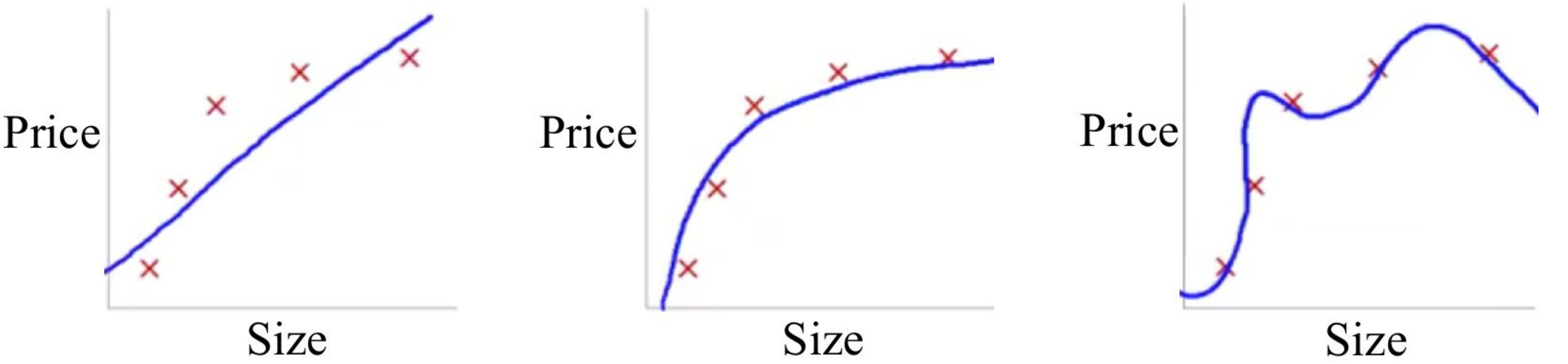
Krittameth Teachasrisaksakul

krittameth.teacha@gmail.com

ทบทวน: Bias / Variance

ความสัมพันธ์ระหว่าง degree d ของ polynomial และการเกิด underfitting หรือ overfitting

ในอุดมคติ เราอยากจะหาจุดตรงกลางระหว่าง 2 สถานการณ์นี้



$\theta_0 + \theta_1 x$	$\theta_0 + \theta_1 x + \theta_2 x^2$	$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$
High bias (underfit)	'Just Right'	High variance (overfit)
degree = 1	degree = 2	degree = 4

Figure source: Andrew Ng's Machine Learning course

Bias / Variance

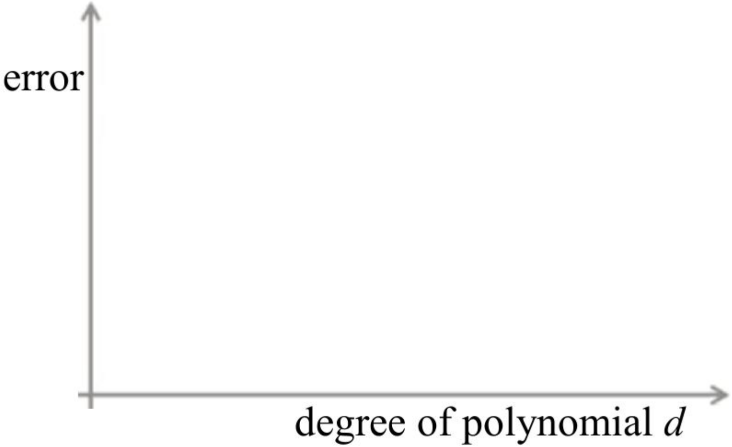
Training

$$J_{\textbf{train}}(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

error:

Cross Validation error:

$$J_{\textbf{CV}}(\theta) = \frac{1}{2m_{\textbf{CV}}}\Sigma_{i=1}^{m_{\textbf{CV}}}\left(h_\theta(x_{\textbf{CV}}^{(i)}) - y_{\textbf{CV}}^{(i)}\right)^2$$



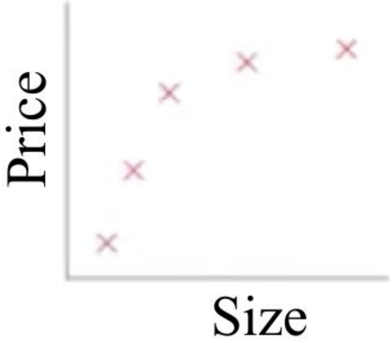


Figure source: Andrew Ng's Machine Learning course

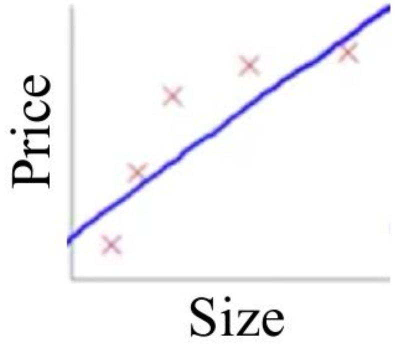
Bias / Variance

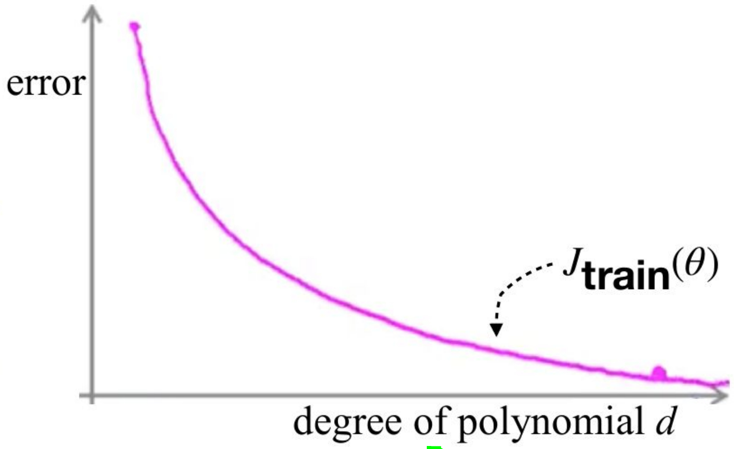
Training error:

$$J_{\textbf{train}}(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Cross Validation error:

$$J_{\textbf{cv}}(\theta) = \frac{1}{2m_{\textbf{cv}}}\Sigma_{i=1}^{m_{\textbf{cv}}}\left(h_\theta(x_{\textbf{cv}}^{(i)}) - y_{\textbf{cv}}^{(i)}\right)^2$$





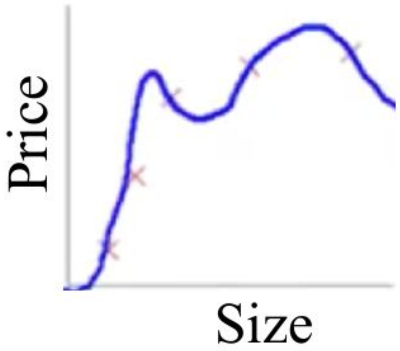


Figure source: Andrew Ng's Machine Learning course

J_{train} : training error ลดลง เมื่อเพิ่ม degree d ของพหุนาม

J_{CV} : cross validation error ลดลง เมื่อเพิ่ม d จนถึงจุดหนึ่ง และเพิ่มขึ้น เมื่อเพิ่ม d ต่อ (cross validation error เป็นเส้นโค้งแบบ convex)

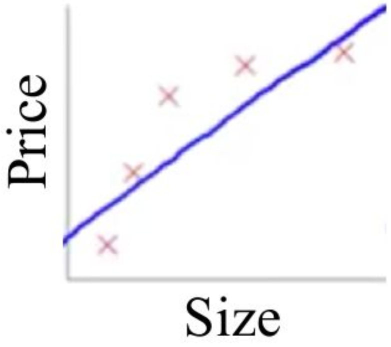
Bias / Variance

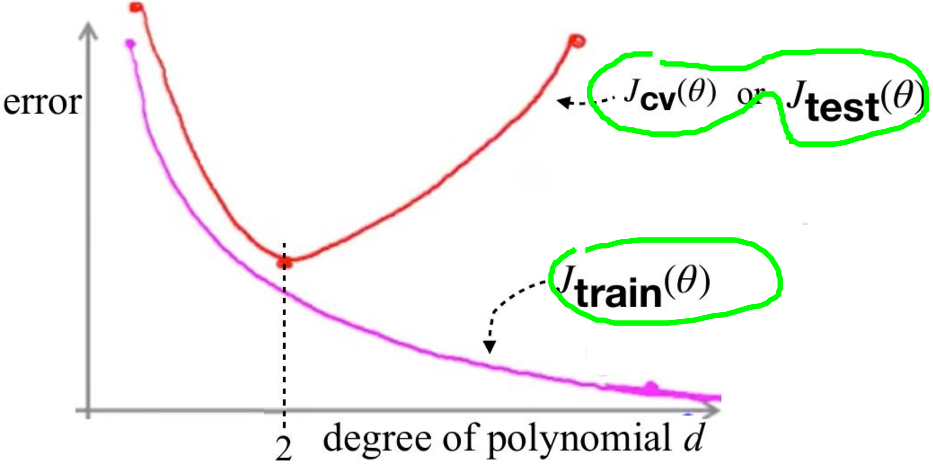
Training error:

$$J_{\textbf{train}}(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Cross Validation error:

$$J_{\textbf{cv}}(\theta) = \frac{1}{2m_{\textbf{cv}}}\Sigma_{i=1}^{m_{\textbf{cv}}}\left(h_\theta(x_{\textbf{cv}}^{(i)}) - y_{\textbf{cv}}^{(i)}\right)^2$$





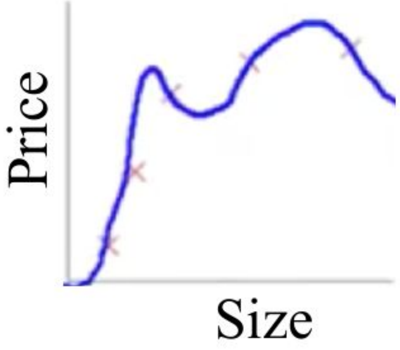


Figure source: *Andrew Ng's Machine Learning course*

พิจารณา bias vs. variance

สมมติ learning algorithm ทำงานได้แย่กว่าที่เราคาดไว้ ก็คือ $J_{\mathrm{CV}}(\theta)$ หรือ $J_{\mathrm{test}}(\theta)$ สูง

นี่เป็นปัญหาเกี่ยวกับ bias หรือ variance ?

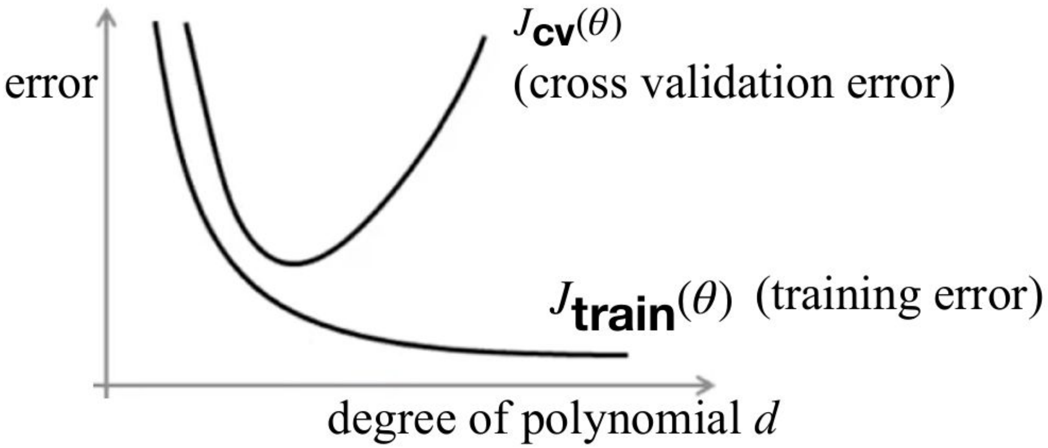
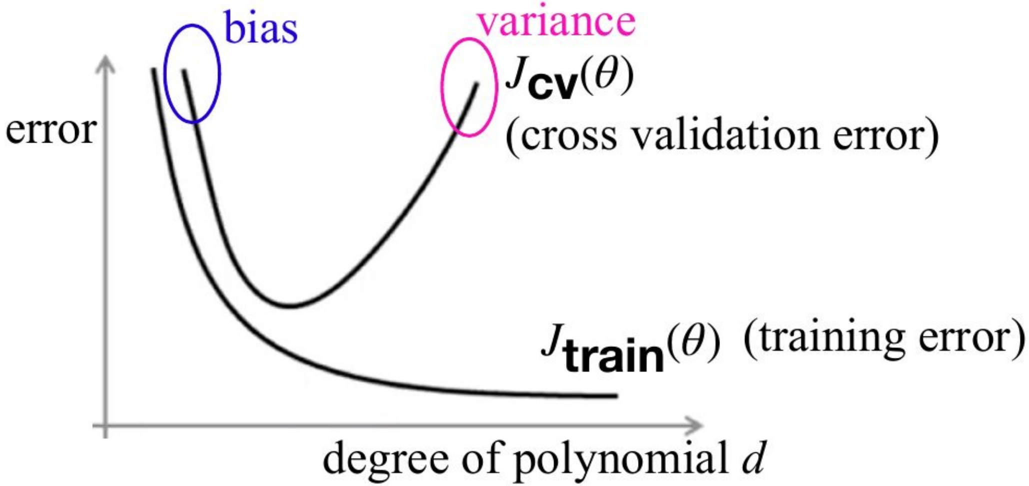


Figure source: Andrew Ng's Machine Learning course

พิจารณา bias vs. variance

สมมติ learning algorithm ทำงานได้แย่กว่าที่เราคาดไว้ ก็คือ $J_{\text{CV}}(\theta)$ หรือ $J_{\text{test}}(\theta)$ สูง

นี่เป็นปัญหาเกี่ยวกับ bias หรือ variance ?



Bias (underfit)

- $J_{\text{train}}(\theta)$ จะสูง
- $J_{\text{cv}}(\theta) \approx J_{\text{train}}(\theta)$

Variance (overfit)

- $J_{\text{train}}(\theta)$ จะต่ำ
- $J_{\text{cv}}(\theta) \gg J_{\text{train}}(\theta)$

Figure source: *Andrew Ng's Machine Learning course*

Question

สมมติเรามีปัญหาการแยกประเภท (classification problem) (misclassification) error มีนิยามว่า

$$(1/m)\Sigma_1^m \mathbf{err}\left(h_\theta(x^{(i)}), y^{(i)}\right)$$

และ cross validation (misclassification) error มีนิยามคล้ายๆกัน โดยใช้ ตัวอย่างจากชุดข้อมูล cross-validation

$$(x_{\mathbf{cv}}^{(1)}, y_{\mathbf{cv}}^{(1)}), \ldots, (x_{\mathbf{cv}}^{(m_{\mathbf{cv}})}, y_{\mathbf{cv}}^{(m_{\mathbf{cv}})})$$

สมมติ training error เป็น 0.10 และ cross validation error เป็น 0.30 ปัญหาใดที่น่าจะเกิดกับ algorithm มากที่สุด?

(i) bias สูง (overfitting)

(ii) bias สูง (underfitting)

(iii) variance สูง (overfitting)

(iv) variance สูง (underfitting)

Question

สมมติเรามีปัญหาการแยกประเภท (classification problem) (misclassification) error มีนิยามว่า

$$(1/m)\Sigma_1^m \mathbf{err}\left(h_\theta(x^{(i)}), y^{(i)}\right)$$

และ cross validation (misclassification) error มีนิยามคล้ายๆกัน โดยใช้ ตัวอย่างจากชุดข้อมูล cross-validation

$$(x_{\mathbf{cv}}^{(1)}, y_{\mathbf{cv}}^{(1)}), \ldots, (x_{\mathbf{cv}}^{(m_{\mathbf{cv}})}, y_{\mathbf{cv}}^{(m_{\mathbf{cv}})})$$

สมมติ training error เป็น 0.10 และ cross validation error เป็น 0.30 ปัญหาใดที่น่าจะเกิดกับ algorithm มากที่สุด?

(i) bias สูง (overfitting)

(ii) bias สูง (underfitting)

(iii) variance สูง (overfitting)

(iv) variance สูง (underfitting)

วิธีปฏิบัติเมื่อประยุกต์ใช้ Machine Learning

Regularization และ Bias / Variance

Krittameth Teachasrisaksakul

krittameth.teacha@gmail.com

ทบทวน : Regularization

Linear regression ที่ใช้ regularization

Model:

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$J(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\Sigma_{j=1}^{n}\theta_j^2$$

ทบทวน : Regularization

Linear regression ที่ใช้ regularization

Model: $$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$J(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\Sigma_{j=1}^{n}\theta_j^2$$

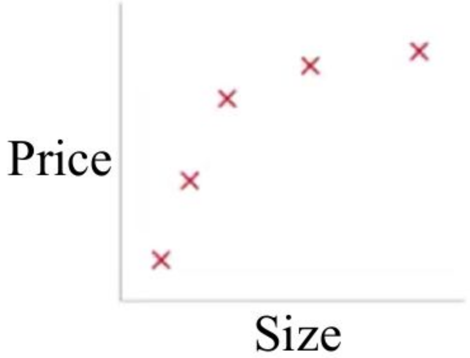
ทบทวน : Regularization

Linear regression ที่ใช้ regularization

Model:

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$J(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\Sigma_{j=1}^{n}\theta_j^2$$



λ มาก

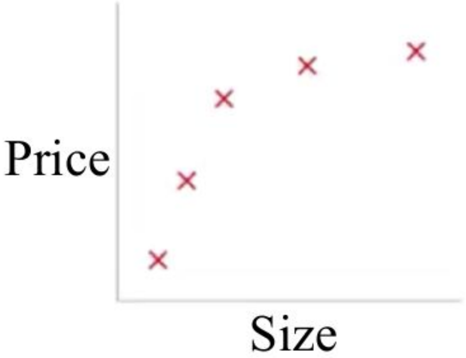
Bias สูง (underfit)

$$\lambda = 10000,\ \theta_1 \approx 0,\ \theta_2 \approx 0, \ldots$$
$$h_\theta(x) \approx \theta_0$$

λ กลาง

'พอดี'

'Just right'



λ น้อย

Variance สูง (overfit)

$$\lambda = 0$$

Figure source: Andrew Ng's Machine Learning course

ทบทวน : Regularization

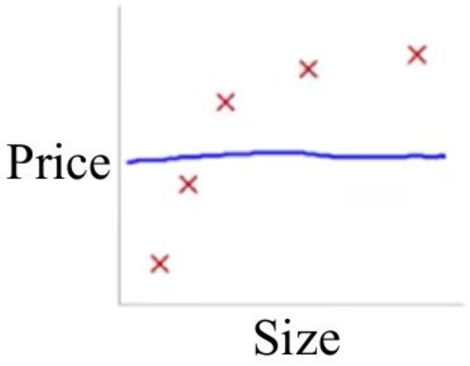
Model:

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$J(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\Sigma_{j=1}^{n}\theta_j^2$$

Linear regression ที่ใช้ regularization

เมื่อ λ เพิ่ม $\longrightarrow$ model จะเป็นเส้นตรงมากขึ้น

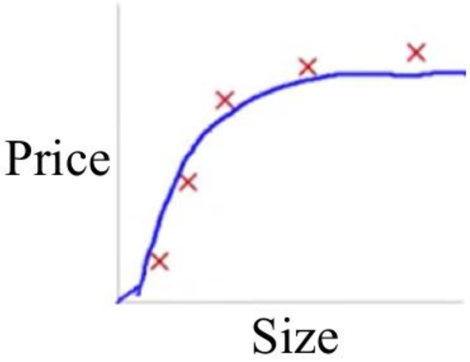


λ มาก

Bias สูง (underfit)

$\lambda = 10000,\ \theta_1 \approx 0,\ \theta_2 \approx 0, \ldots$

$h_\theta(x) \approx \theta_0$



λ กลาง

‘พอดี’

‘Just right’

เมื่อ λ เข้าใกล้ $0 \longrightarrow$

มีแนวโน้ม overfit ข้อมูล

λ น้อย

Variance สูง (overfit)

$\lambda = 0$

Figure source: Andrew Ng’s Machine Learning course

การเลือก Regularization Parameters

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$J(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2m}\Sigma_{j=1}^{n}\theta_j^2$$

การเลือก Regularization Parameters

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$J(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2m}\Sigma_{j=1}^{n}\theta_j^2$$

$$J_{\textbf{train}}(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$J_{\textbf{cv}}(\theta) = \frac{1}{2m_{\textbf{cv}}}\Sigma_{i=1}^{m_{\textbf{cv}}}\left(h_\theta(x_{\textbf{cv}}^{(i)}) - y_{\textbf{cv}}^{(i)}\right)^2$$

$$J_{\textbf{test}}(\theta) = \frac{1}{2m_{\textbf{test}}}\Sigma_{i=1}^{m_{\textbf{test}}}\left(h_\theta(x_{\textbf{test}}^{(i)}) - y_{\textbf{test}}^{(i)}\right)^2$$

การเลือก Regularization Parameters

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$J(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2m}\Sigma_{j=1}^{n}\theta_j^2$$

เลือก regularization

โดยอัตโนมัติอย่างไร?

1. Try $\lambda = 0$
2. Try $\lambda = 0.01$
3. Try $\lambda = 0.02$
4. Try $\lambda = 0.04$
5. Try $\lambda = 0.08$

$\vdots$

12. Try $\lambda = 10$

1. สร้าง list ของค่า λ

การเลือก Regularization Parameters

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$J(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2m}\Sigma_{j=1}^{n}\theta_j^2$$

เลือก regularization

โดยอัตโนมัติอย่างไร?

1. Try $\lambda = 0$ $\longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(1)}$
2. Try $\lambda = 0.01$ $\longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(2)}$
3. Try $\lambda = 0.02$ $\longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(3)}$
4. Try $\lambda = 0.04$ $\longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(4)}$
5. Try $\lambda = 0.08$ $\longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(5)}$

$\vdots$

12. Try $\lambda = 10$ $\longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(12)}$

2. สำหรับ แต่ละค่า λ (จาก list ใน ขั้นที่ 1) เรียนรู้ Θ

การเลือก Regularization Parameters

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$J(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2m}\Sigma_{j=1}^{n}\theta_j^2$$

เลือก regularization

โดยอัตโนมัติอย่างไร?

1. Try $\lambda = 0 \longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(1)} \longrightarrow J_{cv}(\theta^{(1)})$
2. Try $\lambda = 0.01 \longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(2)} \longrightarrow J_{cv}(\theta^{(2)})$
3. Try $\lambda = 0.02 \longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(3)} \longrightarrow J_{cv}(\theta^{(3)})$
4. Try $\lambda = 0.04 \longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(4)} \longrightarrow J_{cv}(\theta^{(4)})$
5. Try $\lambda = 0.08 \longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(5)} \longrightarrow J_{cv}(\theta^{(5)})$

$\vdots$

12. Try $\lambda = 10 \longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(12)} \longrightarrow J_{cv}(\theta^{(12)})$

3. คำนวณ cross validation error

$J_{\text{CV}}(\Theta)$ โดยใช้ Θ ที่เรียนรู้ (จากขั้นที่ 2)

การเลือก Regularization Parameters

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$J(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2m}\Sigma_{j=1}^{n}\theta_j^2$$

เลือก regularization โดยอัตโนมัติอย่างไร?

$$\left.\begin{array}{lllll}
1.\ \text{Try } \lambda = 0 & \longrightarrow \min_\theta J(\theta) & \longrightarrow \theta^{(1)} & \longrightarrow & J_{cv}(\theta^{(1)}) \\
2.\ \text{Try } \lambda = 0.01 & \longrightarrow \min_\theta J(\theta) & \longrightarrow \theta^{(2)} & \longrightarrow & J_{cv}(\theta^{(2)}) \\
3.\ \text{Try } \lambda = 0.02 & \longrightarrow \min_\theta J(\theta) & \longrightarrow \theta^{(3)} & \longrightarrow & J_{cv}(\theta^{(3)}) \\
4.\ \text{Try } \lambda = 0.04 & \longrightarrow \min_\theta J(\theta) & \longrightarrow \theta^{(4)} & \longrightarrow & J_{cv}(\theta^{(4)}) \\
5.\ \text{Try } \lambda = 0.08 & \longrightarrow \min_\theta J(\theta) & \longrightarrow \theta^{(5)} & \longrightarrow & J_{cv}(\theta^{(5)}) \\
\vdots & & & & \\
12.\ \text{Try } \lambda = 10 & \longrightarrow \min_\theta J(\theta) & \longrightarrow \theta^{(12)} & \longrightarrow & J_{cv}(\theta^{(12)})
\end{array}\right\}$$

4. เลือกค่า λ ที่ทำให้ error ของ cross validation set $J_{CV}(\theta)$ น้อยสุด ! เช่น $\theta^{(5)}$

การเลือก Regularization Parameters

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

$$J(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2m}\Sigma_{j=1}^{n}\theta_j^2$$

เลือก regularization

โดยอัตโนมัติอย่างไร?

$$1.\ \text{Try } \lambda = 0 \longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(1)} \longrightarrow J_{cv}(\theta^{(1)})$$

$$2.\ \text{Try } \lambda = 0.01 \longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(2)} \longrightarrow J_{cv}(\theta^{(2)})$$

$$3.\ \text{Try } \lambda = 0.02 \longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(3)} \longrightarrow J_{cv}(\theta^{(3)})$$

$$4.\ \text{Try } \lambda = 0.04 \longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(4)} \longrightarrow J_{cv}(\theta^{(4)})$$

$$5.\ \text{Try } \lambda = 0.08 \longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(5)} \longrightarrow J_{cv}(\theta^{(5)})$$

$$\vdots$$

$$12.\ \text{Try } \lambda = 10 \longrightarrow \min_\theta J(\theta) \longrightarrow \theta^{(12)} \longrightarrow J_{cv}(\theta^{(12)})$$

5. คำนวณ test error เช่น $J_{test}(\theta^{(5)})$

โดยใช้ test set เพื่อดูว่ามัน generalize (ใช้กับข้อมูลใหม่ที่ไม่เคยเจอ) ได้ดีหรือไม่

Question

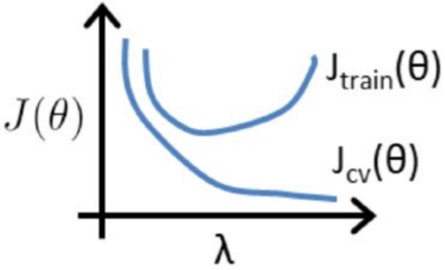
พิจารณา

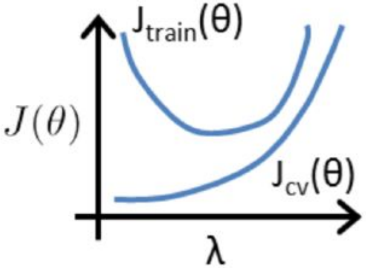
regression และ ให้

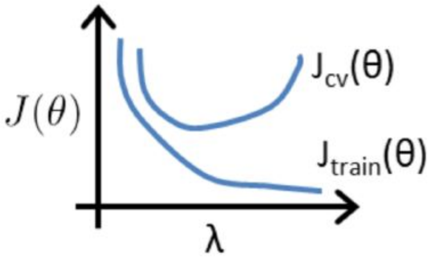
- $J(\theta) = \frac{1}{2m}\left[\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda\Sigma_{j=2}^{n}\theta_j^2\right]$
- $J_{\textbf{train}}(\theta) = \frac{1}{2m_{\textbf{train}}}\left[\Sigma_{i=1}^{m_{\textbf{train}}}\left(h_\theta(x_{\textbf{train}}^{(i)}) - y_{\textbf{train}}^{(i)}\right)^2\right]$
- $J_{\textbf{cv}}(\theta) = \frac{1}{2m_{\textbf{cv}}}\left[\Sigma_{i=1}^{m_{\textbf{cv}}}\left(h_\theta(x_{\textbf{cv}}^{(i)}) - y_{\textbf{cv}}^{(i)}\right)^2\right]$

สมมติเรา plot J_{train} และ J_{cv} เป็น function ของ regularization parameter λ

plot ใดต่อไปนี้ที่เราคาดว่าจะได้?







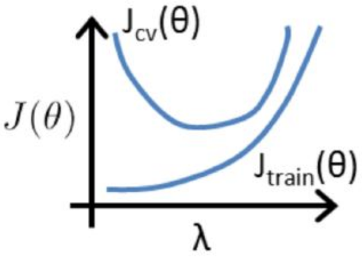


Figure source: Andrew Ng's Machine Learning course

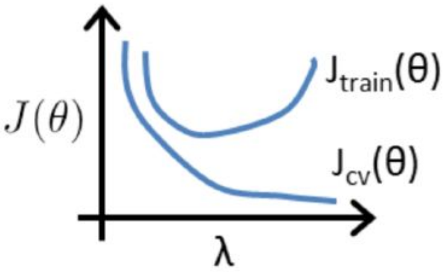
Question

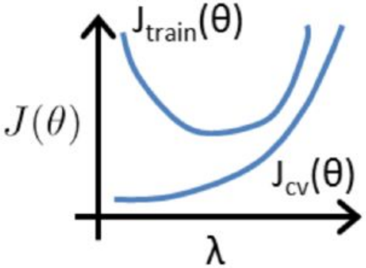
พิจารณา

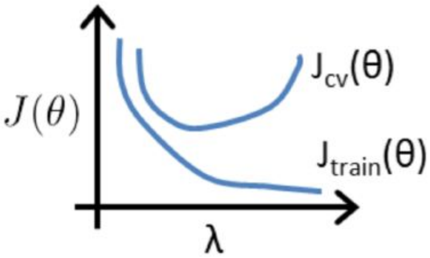
regression และ ให้

- $J(\theta) = \frac{1}{2m}\left[\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda\Sigma_{j=2}^{n}\theta_j^2\right]$
- $J_{\textbf{train}}(\theta) = \frac{1}{2m_{\textbf{train}}}\left[\Sigma_{i=1}^{m_{\textbf{train}}}\left(h_\theta(x_{\textbf{train}}^{(i)}) - y_{\textbf{train}}^{(i)}\right)^2\right]$
- $J_{\textbf{cv}}(\theta) = \frac{1}{2m_{\textbf{cv}}}\left[\Sigma_{i=1}^{m_{\textbf{cv}}}\left(h_\theta(x_{\textbf{cv}}^{(i)}) - y_{\textbf{cv}}^{(i)}\right)^2\right]$

สมมติเรา plot J_{train} และ J_{CV} เป็น function ของ regularization parameter λ plot ใดต่อไปนี้ที่เราคาดว่าจะได้?







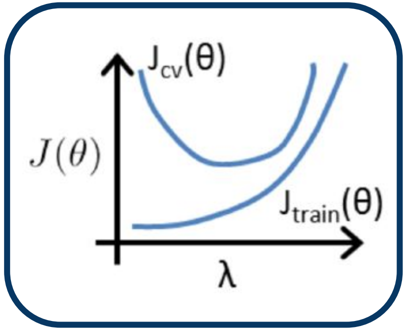


Figure source: Andrew Ng's Machine Learning course

การเลือก Regularization Parameters

Bias / variance เป็น function ของ regularization parameter λ

$$J(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2m}\Sigma_{j=1}^{n}\theta_j^2$$

$$J_{\textbf{train}}(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$J_{\textbf{cv}}(\theta) = \frac{1}{2m_{\textbf{cv}}}\Sigma_{i=1}^{m_{\textbf{cv}}}\left(h_\theta(x_{\textbf{cv}}^{(i)}) - y_{\textbf{cv}}^{(i)}\right)^2$$

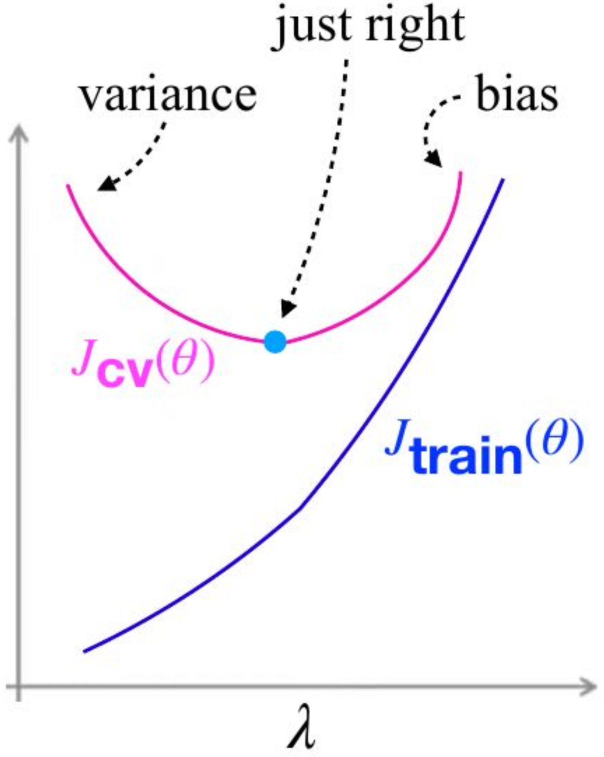


Figure source: Andrew Ng's Machine Learning course

วิธีปฏิบัติเมื่อประยุกต์ใช้ Machine Learning

Learning Curves

Krittameth Teachasrisaksakul

krittameth.teacha@gmail.com

Learning Curves : แนวคิด (idea)

- Plot $J_{\textbf{train}}(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$
- Plot $J_{\textbf{cv}}(\theta) = \frac{1}{2m_{\textbf{cv}}}\Sigma_{i=1}^{m_{\textbf{cv}}}\left(h_\theta(x_{\textbf{cv}}^{(i)}) - y_{\textbf{cv}}^{(i)}\right)^2$

ถ้าฝึก (train) algorithm โดยใช้จุดข้อมูล (data points) จำนวนน้อยมาก (เช่น 1,2, หรือ 3) จะได้ error J_{train} เป็น 0 ได้ง่าย เพราะสามารถหาเส้นโค้ง quadratic (ที่เป็นฟังก์ชันกำลังสอง) ที่ผ่านจุดเหล่านั้นเสมอ

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

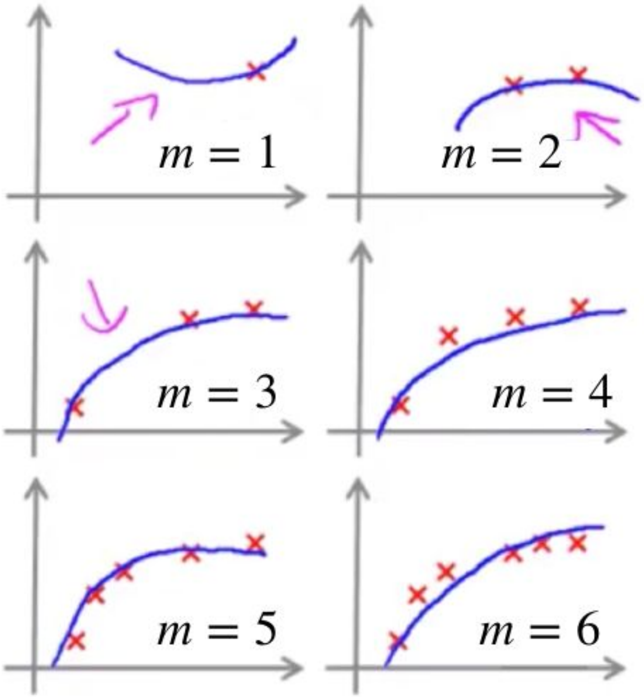
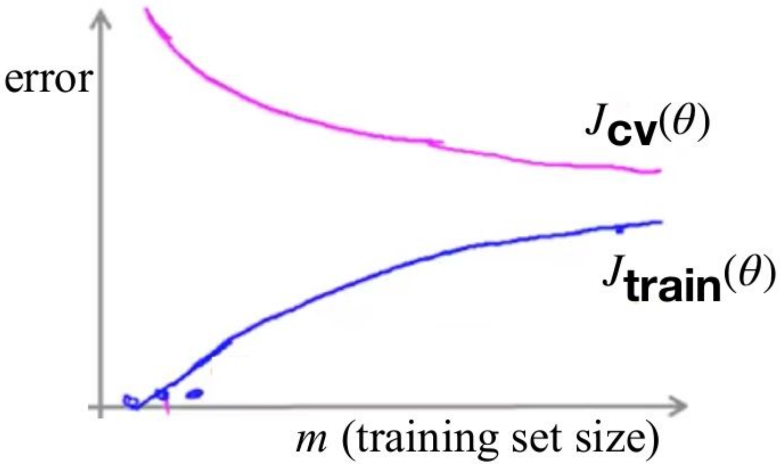


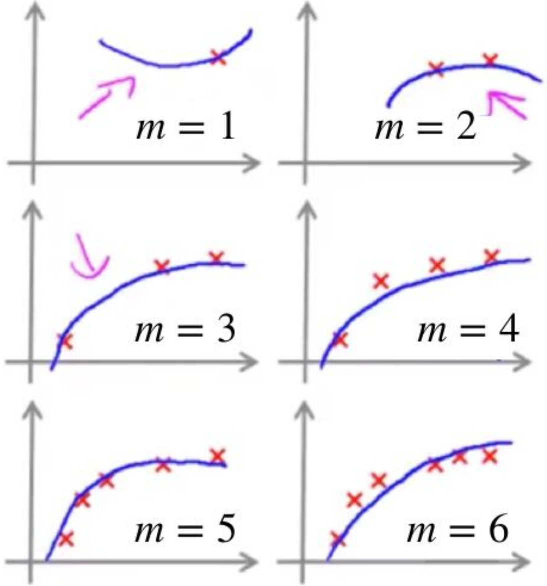
Figure source: *Andrew Ng's Machine Learning course*

Learning Curves : แนวคิด (idea)

- Plot $J_{\textbf{train}}(\theta) = \frac{1}{2m}\Sigma_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$
- Plot $J_{\textbf{cv}}(\theta) = \frac{1}{2m_{\textbf{cv}}}\Sigma_{i=1}^{m_{\textbf{cv}}}\left(h_\theta(x_{\textbf{cv}}^{(i)}) - y_{\textbf{cv}}^{(i)}\right)^2$



$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$



เมื่อ m มากขึ้น (training set ใหญ่ขึ้น) :

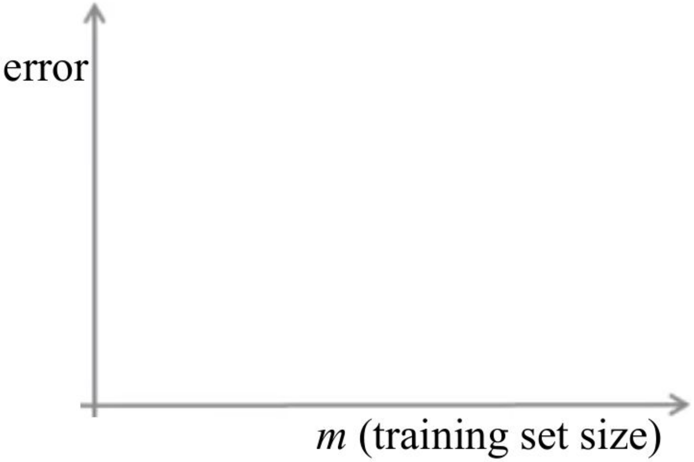
- error (ความผิดพลาด) ของ hypothesis function จะเพิ่มขึ้น
- ค่า error จะคงที่ (plateau out) หลังจากค่า m บางค่า

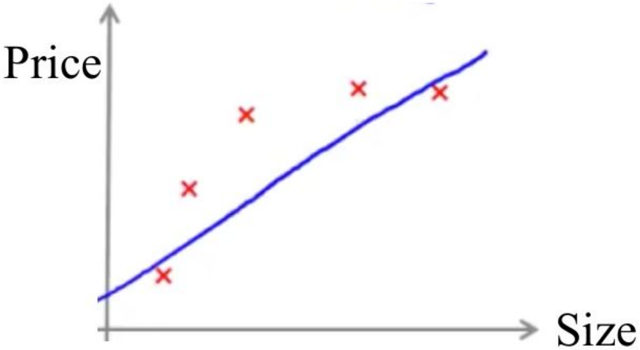
Figure source: Andrew Ng's Machine Learning course

Learning Curves

$$h_\theta(x) = \theta_0 + \theta_1 x$$

Bias สูง





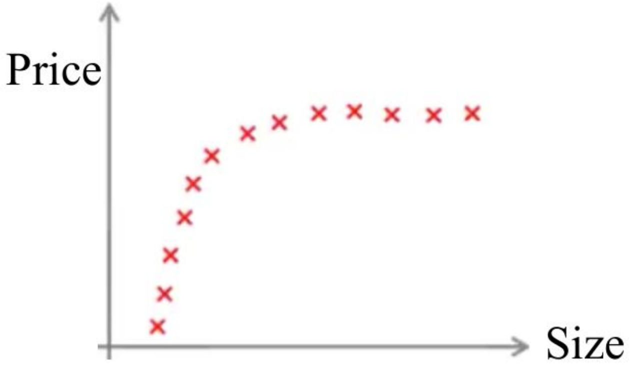
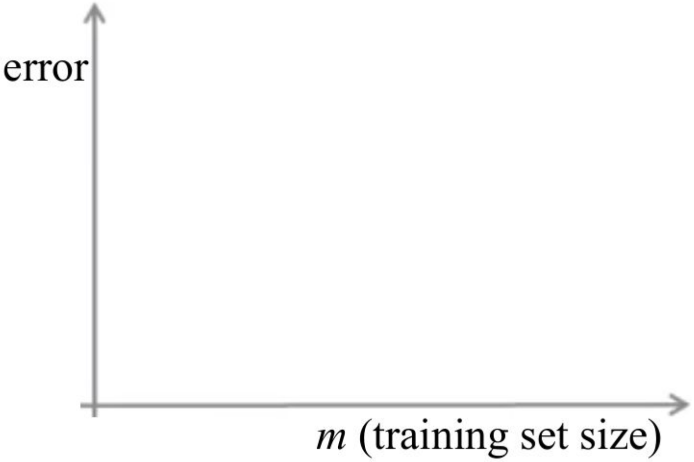


Figure source: Andrew Ng's Machine Learning course

Learning Curves

Bias สูง



$$h_\theta(x) = \theta_0 + \theta_1 x$$

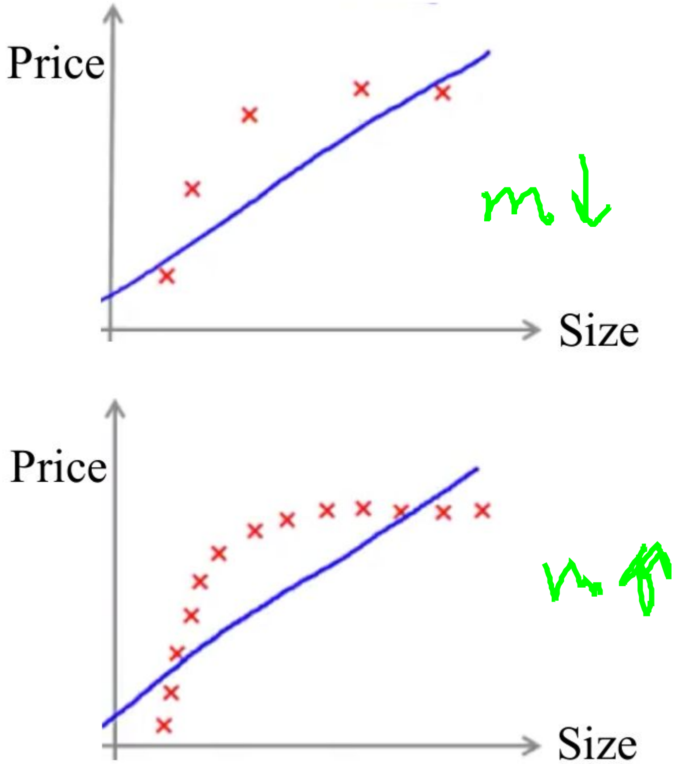
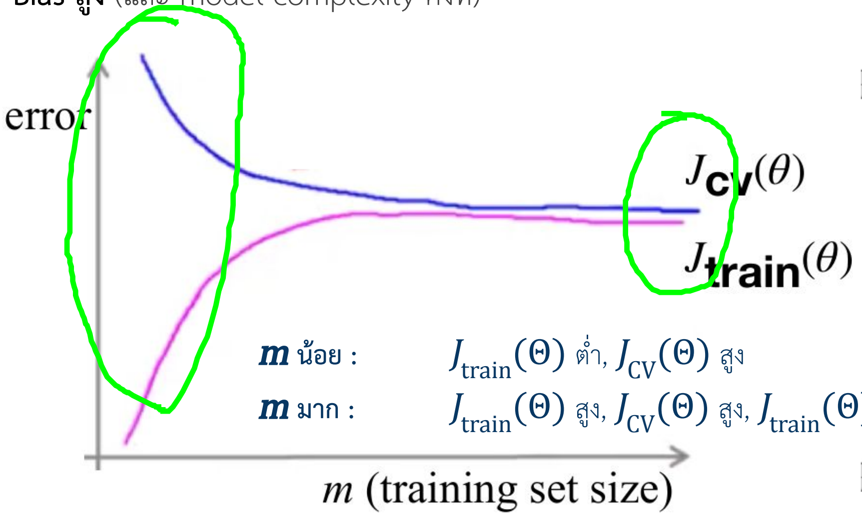


Figure source: Andrew Ng's Machine Learning course

Learning Curves

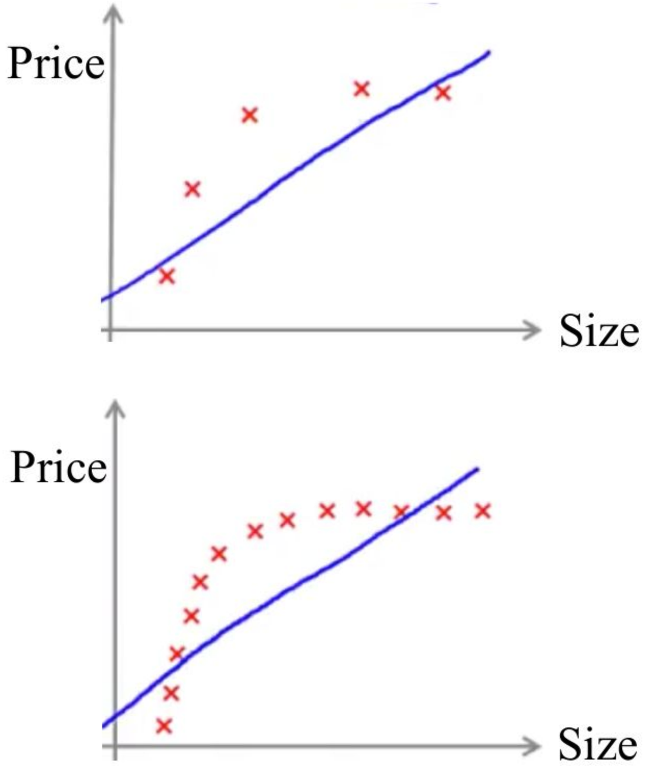
Bias สูง (และ model complexity คงที่)



$\boldsymbol{m}$ น้อย : $J_{\text{train}}(\Theta)$ ต่ำ, $J_{\text{CV}}(\Theta)$ สูง

$\boldsymbol{m}$ มาก : $J_{\text{train}}(\Theta)$ สูง, $J_{\text{CV}}(\Theta)$ สูง, $J_{\text{train}}(\Theta) \approx J_{\text{CV}}(\Theta)$

$$h_\theta(x) = \theta_0 + \theta_1 x$$

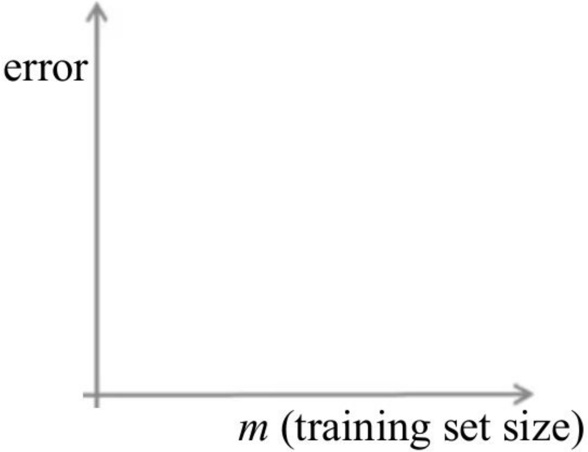


ถ้า learning algorithm ทำงานแย่ลงเพราะ bias สูง $\longrightarrow$ ใช้ข้อมูล training มากขึ้น อย่างเดียว จะไม่ช่วยมาก !

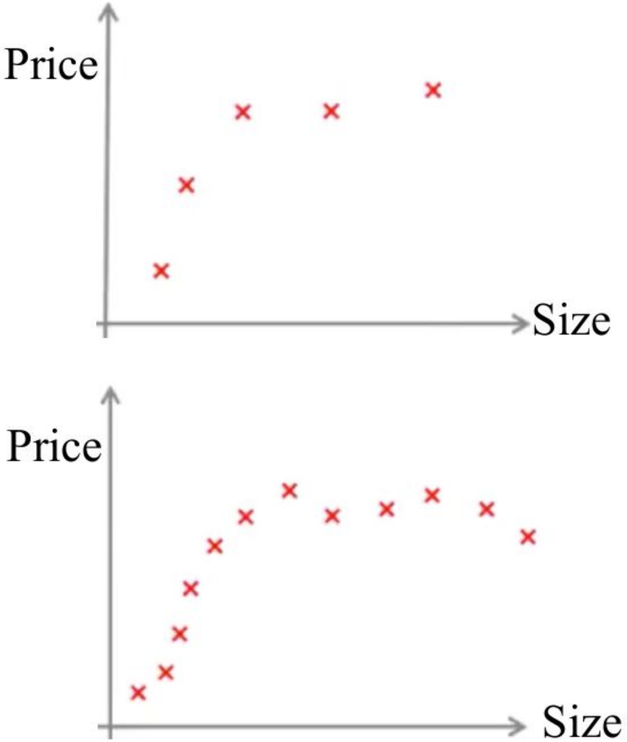
Figure source: *Andrew Ng's Machine Learning course*

Learning Curves

Variance สูง

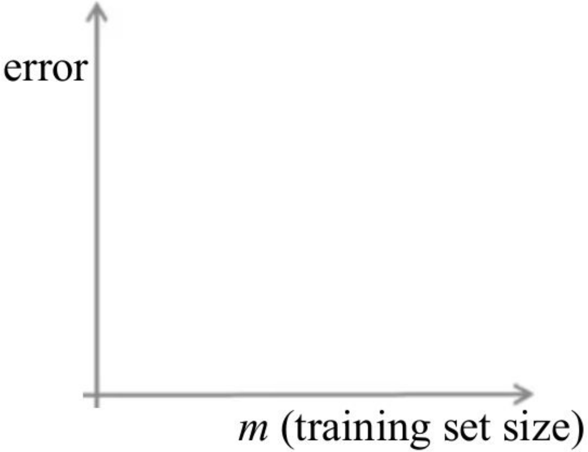


$$h_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_{100} x^{100}$$
(and small λ)



Learning Curves

Variance สูง



$$h_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_{100} x^{100}$$

(and small λ)

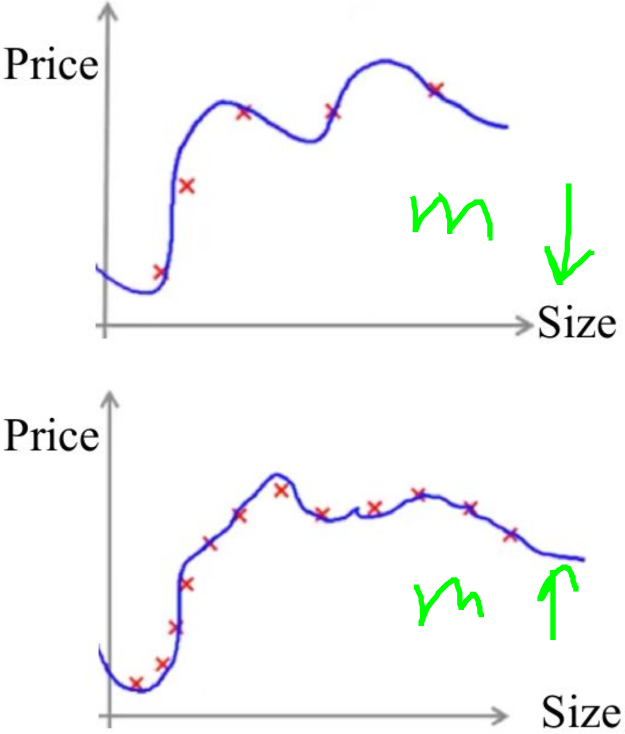
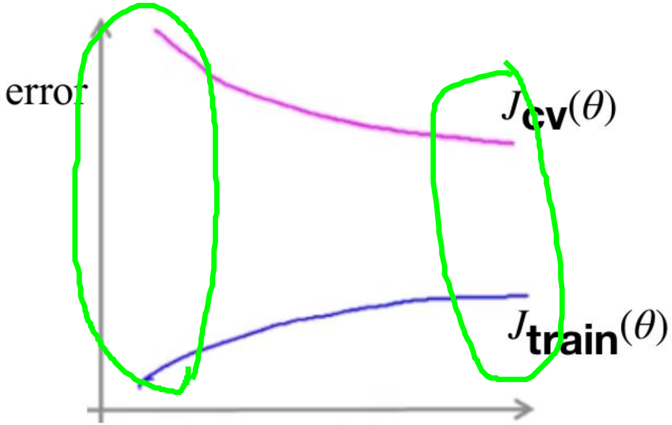


Figure source: Andrew Ng's Machine Learning course

Learning Curves

Variance สูง (และ model complexity คงที่)



***m* น้อย :** $J_{train}(\Theta)$ ต่ำ, $J_{CV}(\Theta)$ สูง

เมื่อเพิ่ม *m* : $J_{train}(\Theta)$ เพิ่ม เรื่อยๆ, $J_{CV}(\Theta)$ ลด เรื่อยๆ

***m* มาก :** $J_{train}(\Theta)$ กับ $J_{CV}(\Theta)$ ยังคงต่างกันมาก

$$h_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_{100} x^{100}$$
(and small λ)

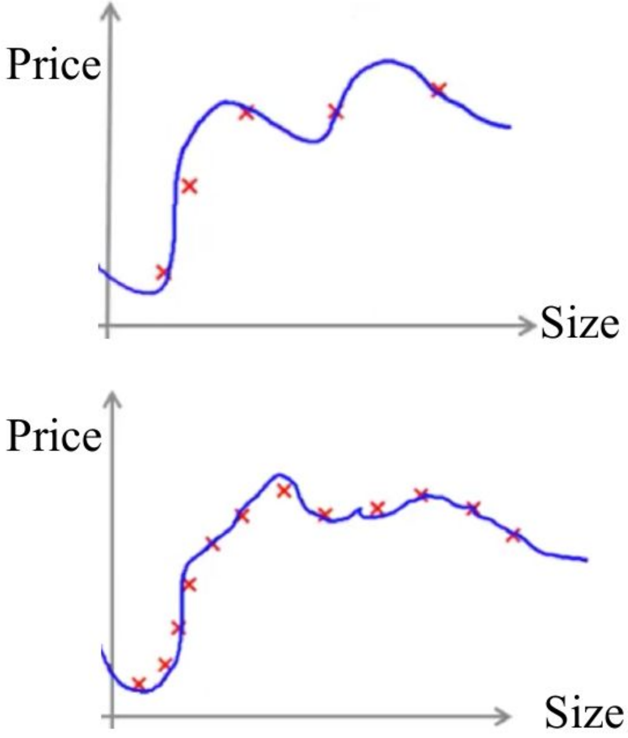


Figure source: Andrew Ng's Machine Learning course

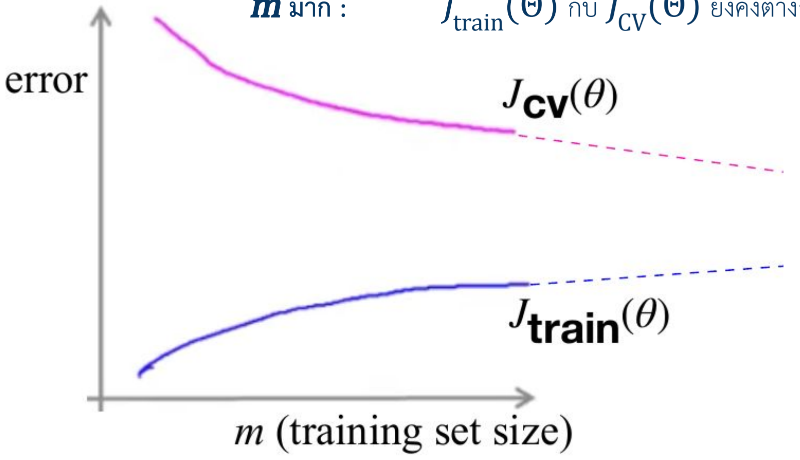
Learning Curves

Variance สูง

$\boldsymbol{m}$ น้อย : $J_{\text{train}}(\Theta)$ ต่ำ, $J_{\text{CV}}(\Theta)$ สูง

เมื่อเพิ่ม $\boldsymbol{m}$: $J_{\text{train}}(\Theta)$ เพิ่ม เรื่อยๆ, $J_{\text{CV}}(\Theta)$ ลด เรื่อยๆ

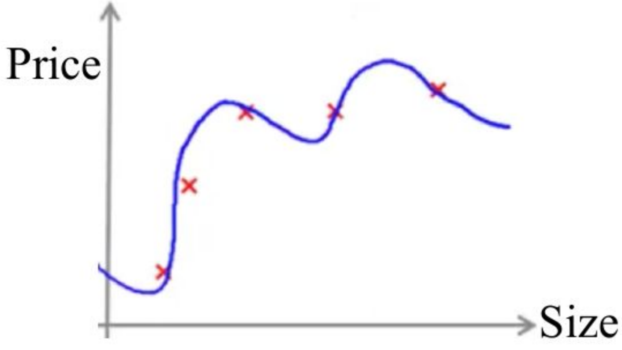
$\boldsymbol{m}$ มาก : $J_{\text{train}}(\Theta)$ กับ $J_{\text{CV}}(\Theta)$ ยังคงต่างกันมาก



ถ้า learning algorithm ทำงานแย่ลงเพราะ variance สูง ⟶ ใช้ข้อมูล training มาก ขึ้นอย่างเดียว น่าจะช่วยได้ !

$$h_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_{100} x^{100}$$

(and small λ)



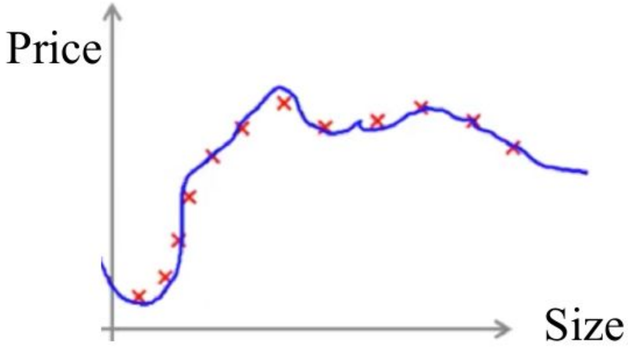


Figure source: Andrew Ng's Machine Learning course

Question

ในสถานการณ์ใดต่อไปนี้ที่ใช้ข้อมูล training data มากขึ้น น่าจะช่วย performance ของ learning algorithm ได้อย่างมาก ?

(i) Algorithm เกิดปัญหาจาก bias สูง

(ii) Algorithm เกิดปัญหาจาก variance สูง

(iii) $J_{\text{CV}}(\theta)$ (cross validation error) มากกว่า $J_{\text{train}}(\theta)$ (training error) มากๆ

(iv) $J_{\text{CV}}(\theta)$ (cross validation error) ประมาณเท่าๆกับ $J_{\text{train}}(\theta)$ (training error)

Question

ในสถานการณ์ใดต่อไปนี้ที่ใช้ข้อมูล training data มากขึ้น น่าจะช่วย performance ของ learning algorithm ได้อย่างมาก ?

(i) Algorithm เกิดปัญหาจาก bias สูง

(ii) Algorithm เกิดปัญหาจาก variance สูง

(iii) $J_{\text{CV}}(\theta)$ (cross validation error) มากกว่า $J_{\text{train}}(\theta)$ (training error) มากๆ

(iv) $J_{\text{CV}}(\theta)$ (cross validation error) ประมาณเท่าๆกับ $J_{\text{train}}(\theta)$ (training error)

วิธีปฏิบัติเมื่อประยุกต์ใช้ Machine Learning

Deciding what to try next

(ตัดสินใจว่าจะลองทำอะไรต่อไป)

Krittameth Teachasrisaksakul

krittameth.teacha@gmail.com

Motivating Example

Debugging (แก้ปัญหา,จุดบกพร่องในโปรแกรม) learning algorithm

สมมติ เรา implement regularized linear regression เพื่อทำนายราคาบ้าน แต่เมื่อเราทดสอบ hypothesis กับข้อมูลบ้านชุดใหม่ แล้วพบว่า hypothesis มี error สูงมากเมื่อทำนายค่า ⟶ เราควรลองทำอะไรต่อไป?

- เก็บตัวอย่างข้อมูล training เพิ่ม
- ลองใช้ชุด features ที่เล็กลง
- ลองเพิ่ม features ใหม่
- ลองเพิ่ม polynomial features (เช่น $x_1^2, x_2^2, x_1 x_2$ เป็นต้น)
- ลองลดค่า λ
- ลองเพิ่มค่า λ

Motivating Example

Debugging (แก้ปัญหา,จุดบกพร่องในโปรแกรม) learning algorithm

สมมติ เรา implement regularized linear regression เพื่อทำนายราคาบ้าน แต่เมื่อเราทดสอบ hypothesis กับข้อมูลบ้านชุดใหม่ แล้วพบว่า hypothesis มี error สูงมากเมื่อทำนายค่า ⟶ เราควรลองทำอะไรต่อไป?

- เก็บตัวอย่างข้อมูล training เพิ่ม ⟶ แก้ปัญหา variance สูง
- ลองใช้ชุด features ที่เล็กลง ⟶ แก้ปัญหา variance สูง
- ลองเพิ่ม features ใหม่ ⟶ แก้ปัญหา bias สูง
- ลองเพิ่ม polynomial features (เช่น x_1^2, x_2^2, x_1x_2 เป็นต้น) ⟶ แก้ปัญหา bias สูง
- ลองลดค่า λ ias สูง
- ลองเพิ่มค่า λ ⟶ แก้ปัญหา variance สูง

Neural Networks

Neural network ขนาดเล็ก

(มี parameter จำนวนน้อยกว่า;

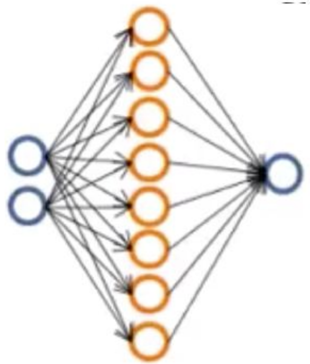
มีแนวโน้มจะ underfitting มากกว่า)

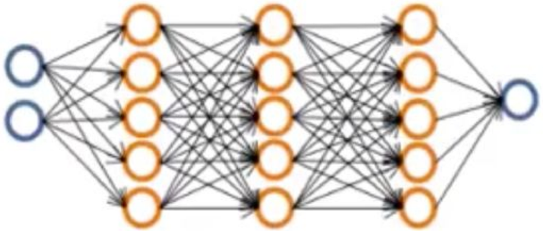
Computationally cheaper

Neural network ใหญ่

(มี parameter จำนวนมากกว่า;

มีแนวโน้มจะ overfitting มากกว่า)





Computationally more expensive

ใช้ regularization (เพิ่ม λ) เพื่อจัดการปัญหา overfitting

เพื่อเลือกจำนวน hidden layers;

คำนวณ $J_{\text{CV}}(\theta)$ หรือ $J_{\text{test}}(\theta)$

Figure source: Andrew Ng’s Machine Learning course

สรุป

(model complexity = ความซับซ้อนของ model)

Model Complexity	Order ของ Polynomial	Bias	Variance	Model fit (เข้ากับ) ข้อมูล training และข้อมูล test
ต่ำ	ต่ำ	สูง	ต่ำ	model เข้ากับข้อมูล training และ ข้อมูล test ได้แย่ (fit poorly)
สูง	สูง	ต่ำ	สูง	model เข้ากับข้อมูล training ได้ดีมากๆ และเข้ากับข้อมูล test ได้แย่มากๆ

Question

สมมติ เรา fit neural network ที่มี hidden layer 1 ชั้น เราพบว่า cross validation error $J_{\text{CV}}(\theta)$ มากกว่า training error $J_{\text{train}}(\theta)$ มาก การเพิ่มจำนวน hidden units น่าจะช่วยได้หรือไม่?

(i) ได้ เพราะมันเพิ่มจำนวน parameters และทำให้ network สามารถใช้เป็นตัวแทน (represent) function ที่ซับซ้อนมากขึ้นได้

(ii) ได้ เพราะมันเกิดปัญหาจาก bias สูง

(iii) ไม่ได้ เพราะมันเกิดปัญหาจาก bias สูง ดังนั้นการเพิ่มจำนวน hidden units น่าจะไม่ช่วย

(iv) ไม่ได้ เพราะมันเกิดปัญหาจาก variance สูง ดังนั้นการเพิ่มจำนวน hidden units น่าจะไม่ช่วย

References

1. Andrew Ng, Machine Learning, Coursera.
2. Teeradaj Racharak, AI Practical Development Bootcamp.
3. What is Machine Learning?, https://www.digitalskill.org/contents/5

Figure Source: Teeradaj Racharak; AI Practical Develop- ment Bootcamp [L 1.2]

T

C

T

C

T

C

T

C

C