Оценяване на алгоритъм за машинно обучение

Доц. Ивайло Пенев

Кат. „Компютърни науки и технологии“

Оценяване на хипотеза

- Ако получената хипотеза е с голяма грешка (отклонява се значително от коректните стойности в обучителните примери), прилагаме следните подходи:
  - Използваме повече обучителни примери
  - Опитваме с по-малък брой променливи (feature)
  - Опитваме с допълнителни променливи
  - Опитваме с полиномни променливи (от по-висока степен)
  - Увеличаваме или намаляваме стойността на параметъра на регуляризацията λ

По-нататъшно оценяване на хипотезата

- Построената хипотеза може да бъде с малка грешка за обучителните данни, но да бъде неточна (поради претрениране - overfitting)
- Разделяме обучителните данни на две множества:
  - Обучително множество (training set)
  - Тестово множество (test set)
- Обикновено обучителното множество съдържа 70% от данните, а тестовото множество 30% от данните
- При тези множества построяването на хипотезата преминава през две стъпки:

1. Обучаваме θ и минимизираме $J_{train}(\theta)$ с обучителното множество (training set)
2. Изчисляваме грешката с тестовото множество (test set)

Изчисляване на грешката

- За линейна регресия

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

- За класификационни задачи

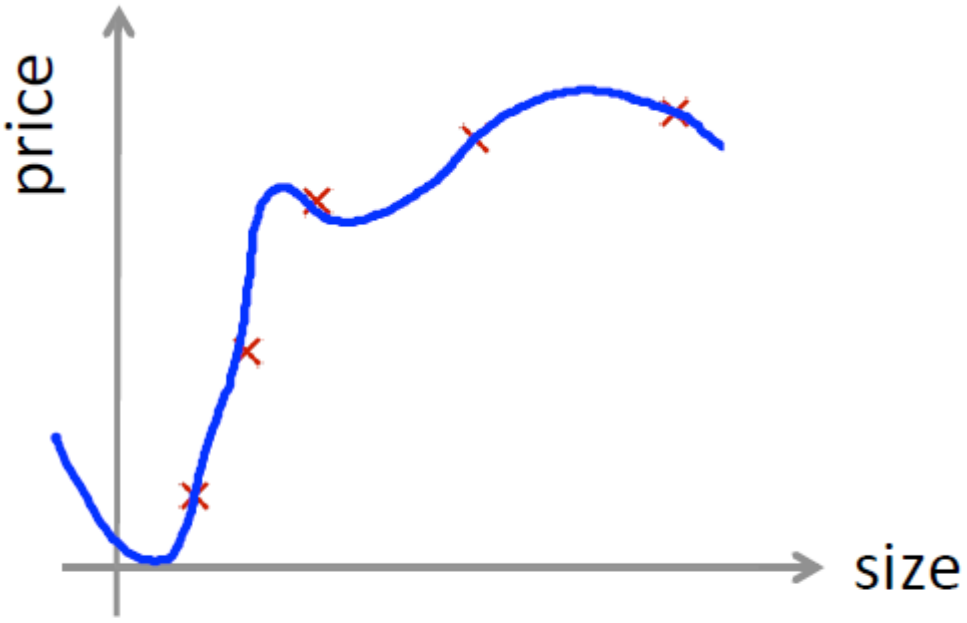
$$err(h_\theta(x), y) = 1, \text{ако } h_\theta(x) \geq 0.5 \text{ и } y = 0 \text{ или } h_\theta(x) < 0.5 \text{ и } y = 1$$
$$err(h_\theta(x), y) = 0 \text{ в противен случай}$$

Горният израз дава като резултат за грешката 1/0

- Средната грешка за тестовото множество показва каква част от тестовите данни са класифицирани грешно

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

Важен проблем в машинното обучение



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

Когато сме намерили стойностите на параметрите $\theta_0, \theta_1, \ldots, \theta_4$ за дадено множество от обучителни данни (training set), грешката за това множество е вероятно да бъде по-малка в сравнение с грешката за произволни данни.

Избор на модел

Как да изберем подходящ модел за хипотезата?

1. $h_\theta(x) = \theta_0 + \theta_1 x \longrightarrow \theta^{(1)} \rightarrow J_{test}(\theta^{(1)})$
2. $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 \longrightarrow \theta^{(2)} \rightarrow J_{test}(\theta^{(2)})$
3. $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 \longrightarrow \theta^{(3)} \rightarrow J_{test}(\theta^{(3)})$

.

.

.

10. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{10} x^{10} \longrightarrow \theta^{(10)} \rightarrow J_{test}(\theta^{(10)})$

Въвеждаме нов параметър *d* – степен на полинома в модела

Избираме модел с най-ниска грешка $J_{test}(\theta^{(d)})$

Избор на модел

- Избираме например $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 + \theta_5 x^5$
- Изчисляваме грешката за тестовите данни (test set) $J_{test}(\theta^{(5)})$
- Как да сме сигурни, че грешката не е малка само за тестовите данни (т.е., че няма претрениране на допълнителния параметър *d* за тестовите данни)?

Обучителни/валидиращи/тестови данни

Dataset:

Size	Price	
2104	400	Обучителни данни (Training set
1600	330	
2400	369	
1416	232	
3000	540	
1985	300	
1534	315	Валидиращи данни (Cross validation set - CV)
1427	199	
1380	212	Тестови данни (Test set)
1494	243	

60%

20%

20%

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

m – брой обучителни данни

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

m_{cv} – брой валидиращи данни

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

m_{test} – брой тестови дан

Изчисляване на грешки

- Грешка за обучителните данни (Training error)

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

- Грешка за валидиращите данни (Cross validation error)

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

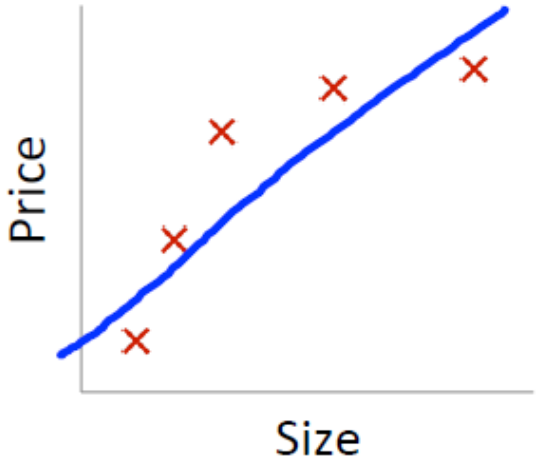
- Грешка за тестовите данни (Test error)

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

Алгоритъм за избор на модела на хипотезата

1. Оптимизираме параметрите θ, използвайки **обучителните данни (training set)** за всички полиноми (т.е. всички степени)
2. Намираме модел със степен на полинома *d* с най-малка грешка за **валидиращите данни (cross validation set)**
3. Оценяваме грешката $J_{test}(\theta^{(d)})$, където *d* – степен на полинома от стъпка 2, използвайки **тестовите данни (test set)**

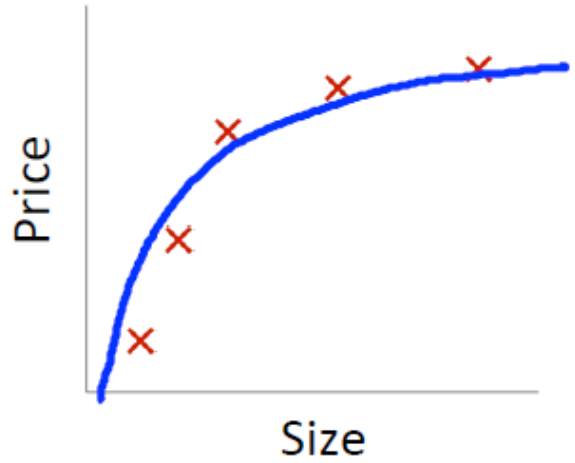
Отклонение и вариация



$\theta_0 + \theta_1 x$

High bias
(underfit)

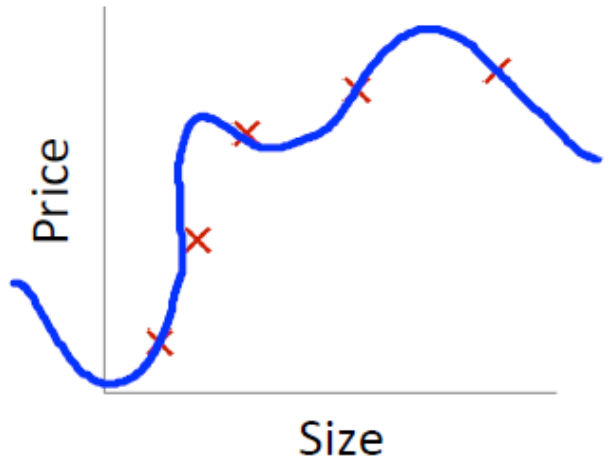
$d = 1$



$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

$d = 2$



$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance
(overfit)

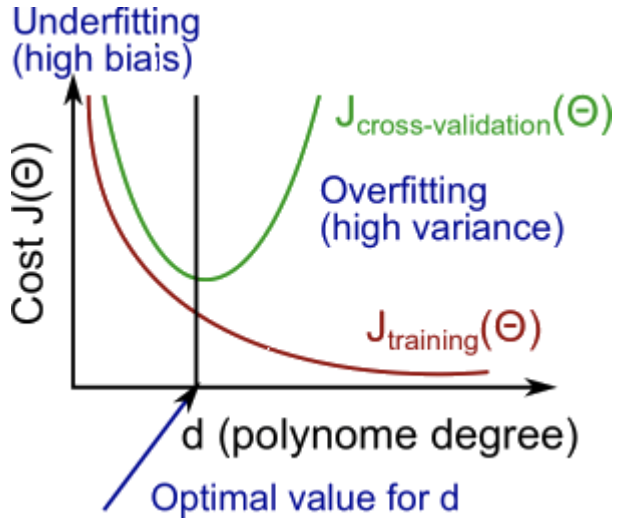
$d = 4$

Отклонение или вариация

- Грешките в получените прогнози могат да се дължат на отклонение (bias) или вариация (variance)
- Високо отклонение причинява **underfitting**, а висока вариация причинява **overfitting**. Стремежът е хипотезата да постига „златна среда“ между двете.
- Грешката при обучение **намалява** с увеличаване на степента *d* на полинома
- Грешката при валидиране **намалява** с увеличаване на степента *d* до някаква стойност и **нараства** при следващо увеличаване на *d*

Отклонение или вариация

- При голямо отклонение **(underfitting)** двете стойности $J_{train}(\theta)$ и $J_{cv}(\theta)$ са високи, $J_{cv}(\theta) \approx J_{train}(\theta)$
- При висока вариация **(overfitting)** $J_{train}(\theta)$ има ниска стойност, а $J_{cv}(\theta)$ е много по-голяма от $J_{train}(\theta)$

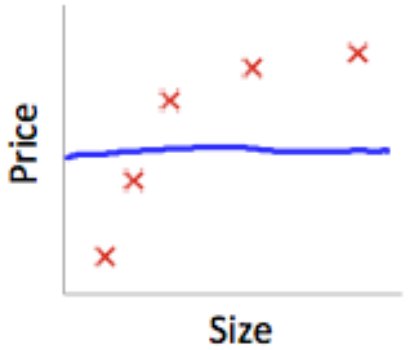


Влияние на регуляризацията върху вариация и отклонение

Linear regression with regularization

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

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

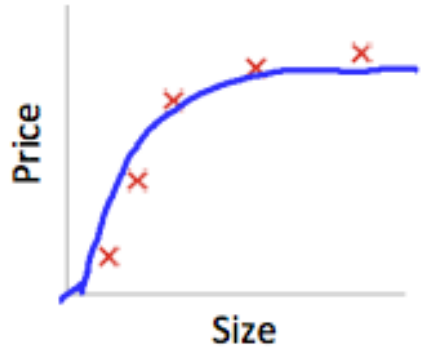


Large λ

High bias (underfit)

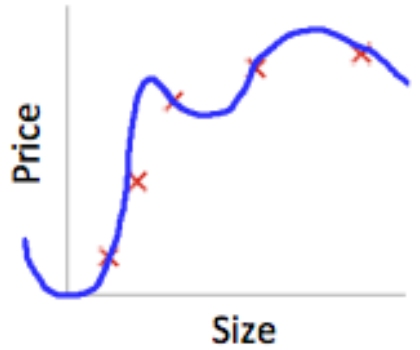
$\lambda = 10000.\ \theta_1 \approx 0, \theta_2 \approx 0, \ldots$

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



Intermediate λ

"Just right"



Small λ

High variance (overfit)

$\lambda = 0$

Andrew Ng

Влияние на регуляризацията върху вариация и отклонение

С нарастване на стойността на параметъра λ кривата на хипотезата става по-плоска. От друга страна, когато λ приближава 0, се получава overfitting. Как да изберем правилна стойност на λ? За избор на модел на хипотезата и на стойността λ следваме стъпки:

1. Създаваме списък със стойности λ (напр. $\lambda = \{0,0.01,0.02,0.04,0.08,0.16,0.32,0.64,1.28,2.56,5.12,10.24\}$)
2. Създаваме модели с различна степен на полинома
3. За всяко λ от списъка за всички модели обучаваме параметрите θ
4. Изчисляваме грешката с валидиращите данни за $J_{cv}(\theta)$, използвайки обучените θ (получени с λ), без регуляризация (т.е. при $\lambda = 0$)
5. Избираме модела с най-малка грешка за валидиращите данни
6. Прилагаме избраните модели и λ върху тестовите данни за изчисляване на $J_{test}(\theta)$, за да проверим каква е грешката при непознати данни

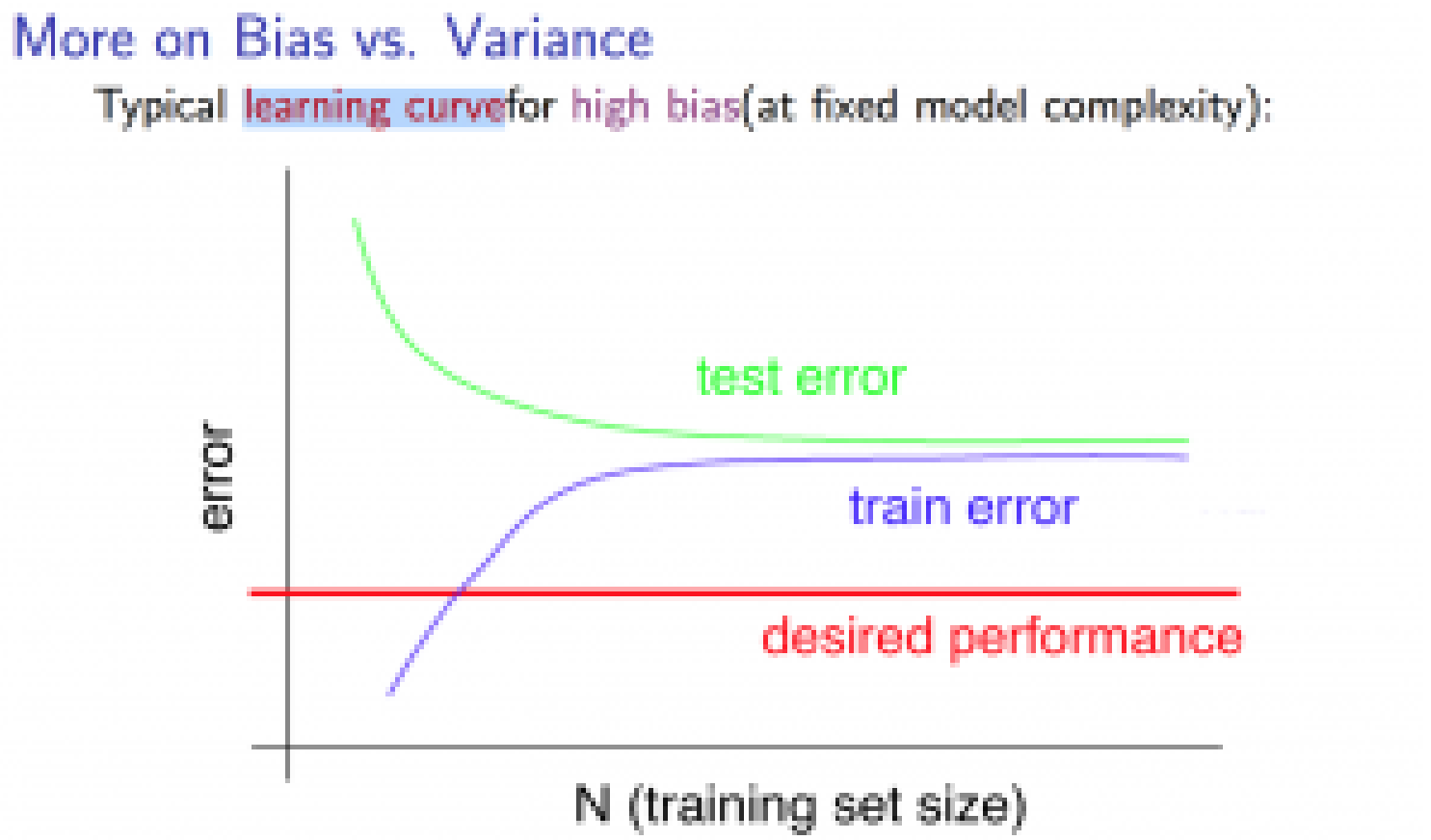
Криви на обучението

- Обучение на алгоритъм с малко данни (напр. 1,2,3) лесно ще доведе до грешка 0, защото ще се намери крива, която да преминава през всички точки
- С увеличаване на обучителните данни грешката ще нараства
- След някаква стойност *m* на броя на обучителните данни грешката няма да нараства повече

Голямо отклонение

- При малко обучителните данни - $J_{train}(\theta)$ е малка, $J_{cv}(\theta)$ е голяма
- При много обучителни данни - $J_{train}(\theta)$ и $J_{cv}(\theta)$ са големи, $J_{train}(\theta) \approx J_{cv}(\theta)$
- **Извод - ако обучаващият алгоритъм е с голямо отклонение, добавянето на нови данни НЯМА да доведе до подобрение на точността на обучението**

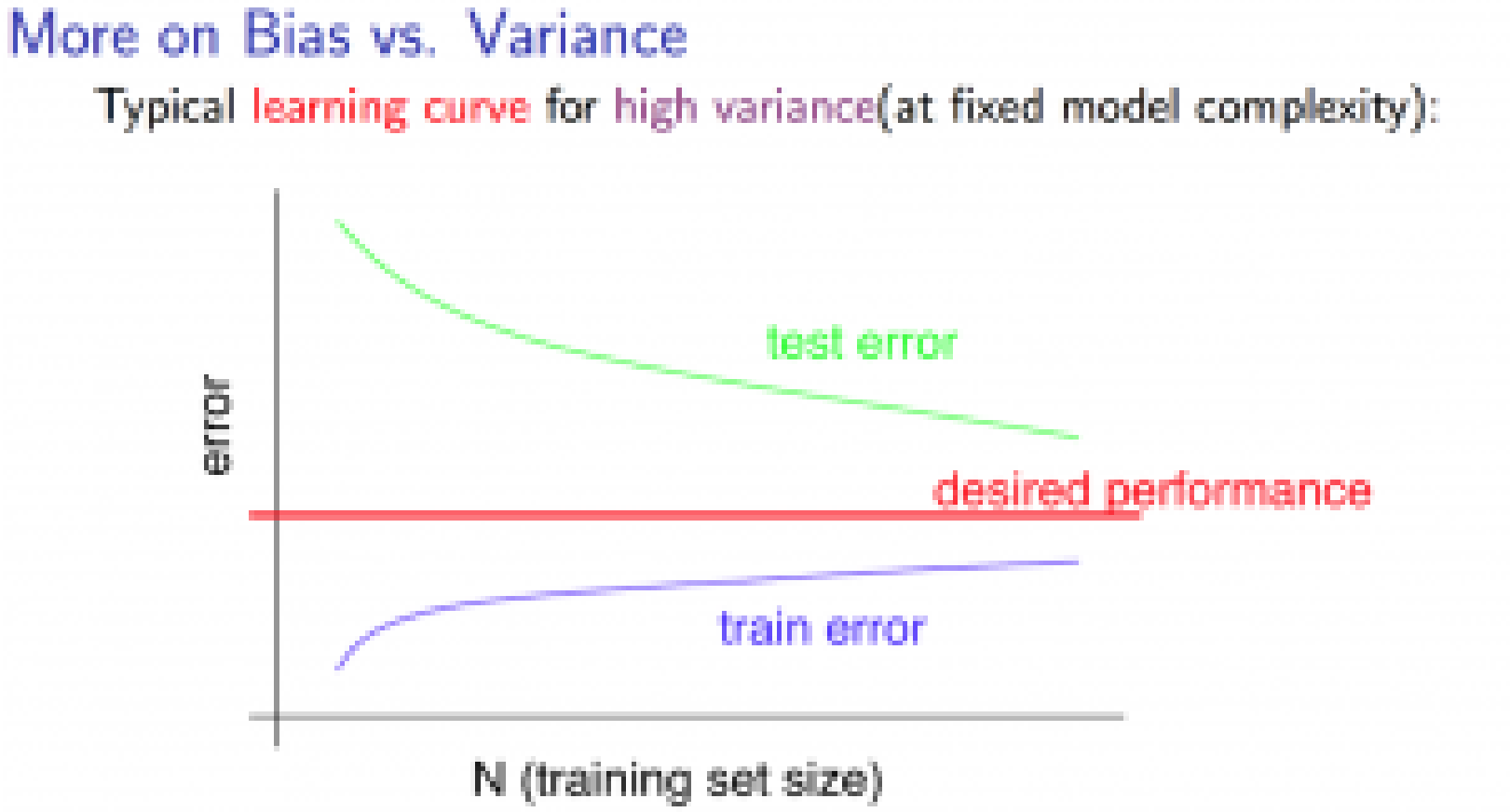
Крива на обучението при голямо отклонение



Голяма вариация

- При малко обучителни данни - $J_{train}(\theta)$ е малка, $J_{cv}(\theta)$ е голяма
- При много обучителни данни - $J_{train}(\theta)$ нараства с увеличване на данните, а $J_{cv}(\theta)$ намалява. $J_{train}(\theta) < J_{cv}(\theta)$, но разликата между двете остава значителна
- **Извод - ако обучаващият алгоритъм е с голяма вариация, добавянето на нови данни МОЖЕ да доведе до подобрение на точността на обучението**

Крива на обучението при голяма вариация



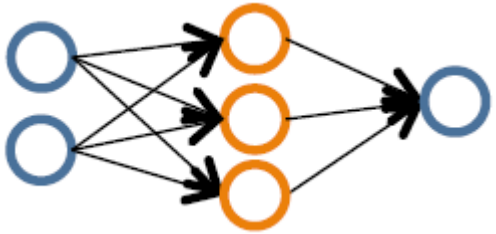
Препоръки за прилагане на обучаващ алгоритъм

Да предположим, че сме реализирали регуляризирана линейна регресия за предсказване на цени на къщи. При тестване на построената хипотеза установяваме, че тя дава голяма грешка в прогнозите. Какво можем да направим, за да подобрим точността?

- Да използваме повече обучаващи примери (training examples) – справяме се с голяма вариация
- Да използваме по-малко променливи (features) – справяме се с голяма вариация
- Да използваме допълнителни променливи – справяме се с голямо отклонение
- Да използваме допълнителни полиномни променливи ($x_1^2, x_2^2, x_1 x_2$ и др.) – справяме се с голямо отклонение
- Да опитаме намаляване на λ - справяме се с голямо отклонение
- Да опитаме увеличаване на λ - справяме се с голяма вариация

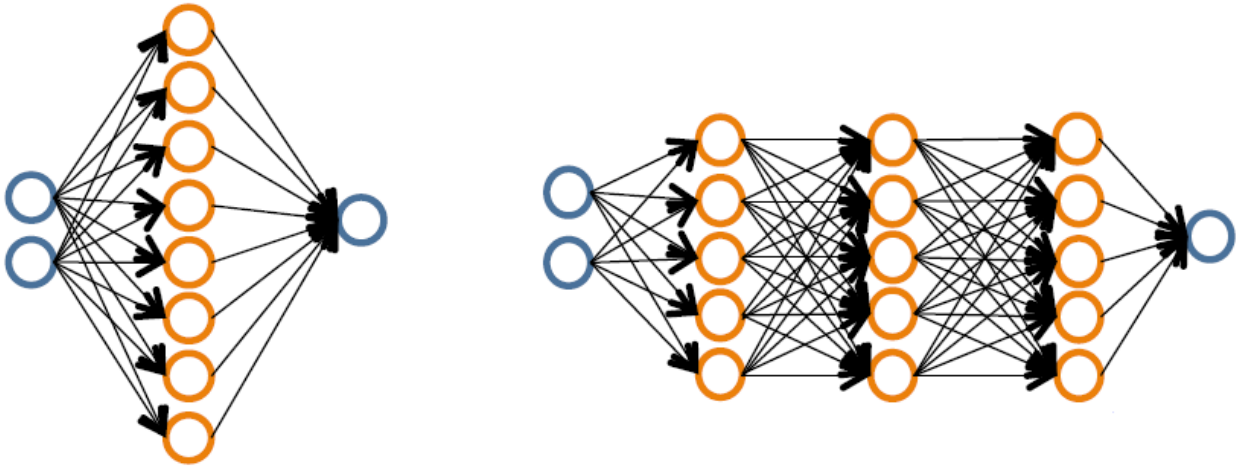
Препоръки при използване на невронни мрежи

„Малка“ невронна мрежа (малко параметри, вероятен underfitting)



Необходими са малко изчисления

„Голяма“ невронна мрежа (повече параметри, вероятен overfitting)



Необходими са повече изчисления

Можем да използваме регуляризация, за да се справим с overfitting