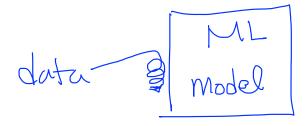
Lecture 7-2

Application & Tips: Learning and test data sets

Sung Kim <hunkim+mr@gmail.com>

Performance evaluation: is this good?



Evaluation using training set?

Size	Price	
2104	400	truling set a mode
1600	330	
2400	369	training set으로 학습하고, 다시 training set으로 모델을 평가하는 것은 모델의. 올바른 평가를 할 수 없다
1416	232	
3000	540	
1985	300	
1534	315	• 100% correct (accuracy)
1427	199	100% correct (accuracy)
1380	212	Can memorize
1494	243	

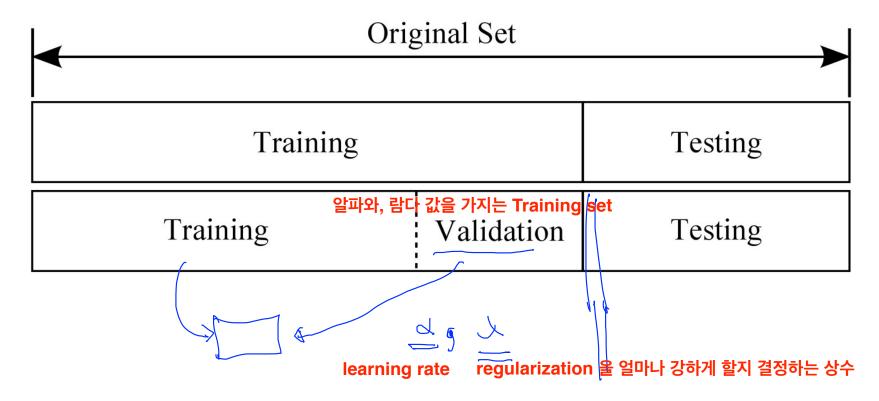
Training and test sets

좋은 방법은 데이터를 7:3 정도로 잘라서 70퍼센트(training set)로 학습 시키고,나머지 30퍼센트(test set)로 평가하는것

	Size	Price
	2104	400
	1600	330 model
Harris	2400	369
trailing (IZHAI)	1416	232
(JEJ2+44)	3000	540
	1985	300
+	1534	315
T	1427	199
fest !	1380	212
set (Like)	1494	243
しんだり		

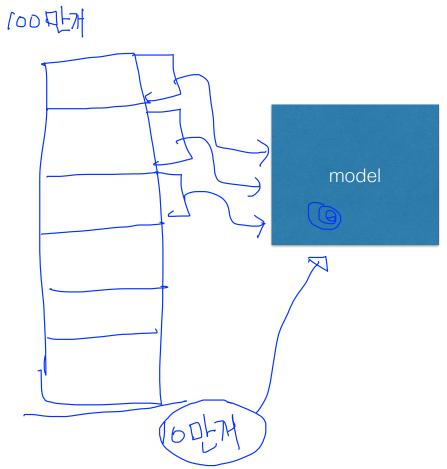
http://www.holehouse.org/mlclass/10_Advice_for_applying_machine_learning.html

Training, validation and test sets



http://www.intechopen.com/books/advances-in-data-mining-knowledge-discovery-and-applications/selecting-representative-data-sets

Online learning



http://www.intechopen.com/books/advances-in-data-mining-knowledge-discovery-and-applications/selecting-representative-data-sets

M NIST Dataset

Zip: 633

 2

train-images-idx3-ubyte.gz: training set images (9912422 bytes)

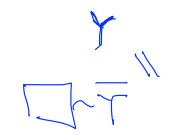
train-labels-idx1-ubyte.gz: training set labels (28881 bytes)

t10k-images-idx3-ubyte.gz: test set images (1648877 bytes)

t10k-labels-idx1-ubyte.gz: test set labels (4542 bytes)

http://yann.lecun.com/exdb/mnist/

Accuracy



- How many of your predictions are correct?
- 95% ~ 99%?
- Check out the lab video

