



Regularization

POSTECH CSE

Minjong Lee

minjong.lee@postech.ac.kr

Table of Contents

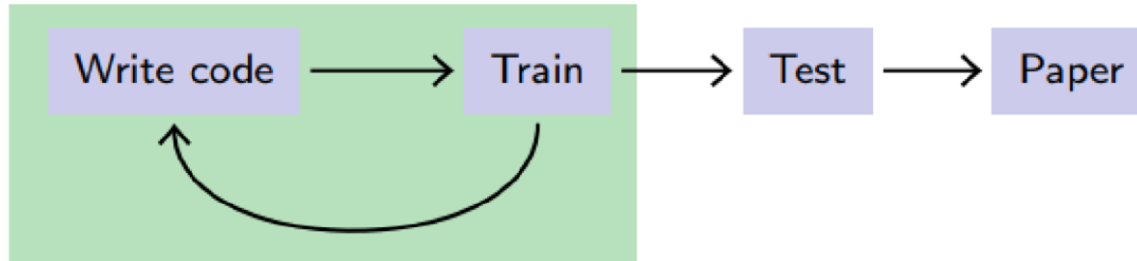
- Review the key points
 - Validation set
 - Model Ensembles
 - Regularization (Add term to loss / Dropout / batch normalization)
- Practice
 - Early stopping
 - Regularization
 - Model Ensembles

Table of Contents

- Review the key points
 - Validation set
 - Model Ensembles
 - Regularization (Add term to loss / Dropout)
- Practice
 - Early stopping
 - Regularization
 - Model Ensembles

Why Validation set??

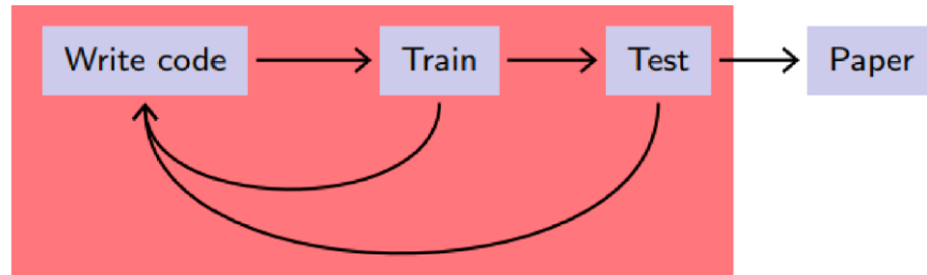
- Evaluation protocol
 - Simple training and evaluation



- Problem: 모델에서 overfitting 문제가 일어날 수 있다

Why Validation set??

- Evaluation protocol
 - Improper training and evaluation

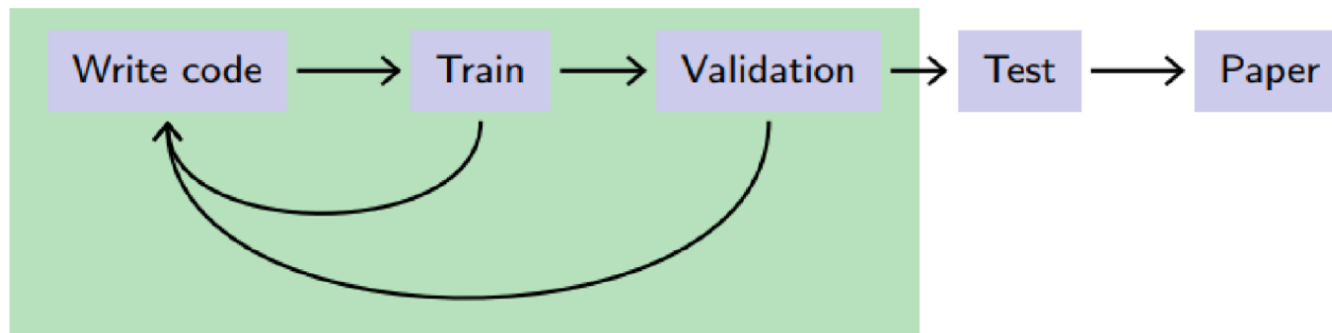


- Problem: cheating!!

Validation set

- Evaluation protocol

- Hyper parameter 조정이나, early stopping을 위해서 validation set을 사용

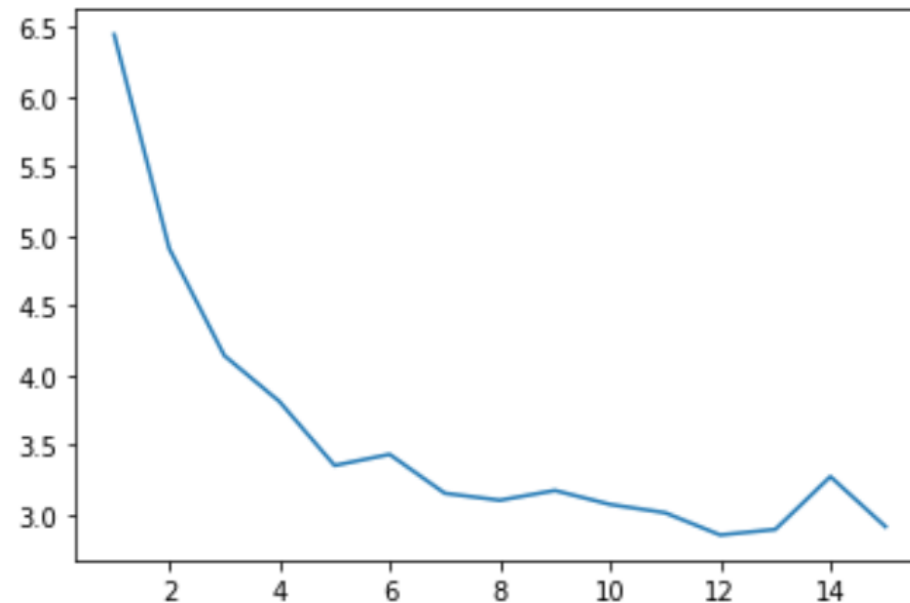


- Typically, train:validation = 4:1

Validation set

- Early stopping

- 미리 지정한 epoch까지 다 가지 않더라도, overfitting이 일어나게 되면 학습을 중단



best epoch : 12

Table of Contents

- Review the key points
 - Validation set
 - **Model Ensembles**
 - Regularization (Add term to loss / Dropout)
- Practice
 - Early stopping
 - Regularization
 - Model Ensembles

Model Ensembles

- Train multiple independent models
 - At test time, average their results
 - Take average of predicted probability distributions, then choose argmax
-
- 여러 개의 다른 모델을 따로 학습시켜서 그 모델들이 test data의 output값에 기여하도록 하겠다는 아이디어
 - 다른 모델들이 한가지 test set에서 모두 똑같은 결과를 보이지는 않을 거라는 아이디어

Table of Contents

- Review the key points
 - Validation set
 - Model Ensembles
 - Regularization (Add term to loss / Dropout)
- Practice
 - Early stopping
 - Regularization
 - Model Ensembles

Add term to loss (weight decay)

$$L(W) = \frac{1}{N} \sum_{i=1}^N \ell(y_i, f_W(x_i)) + \lambda R(W)$$

- In common use

- L2 regularization

$$R(W) = \sum_{i,j} W_{ij}^2$$

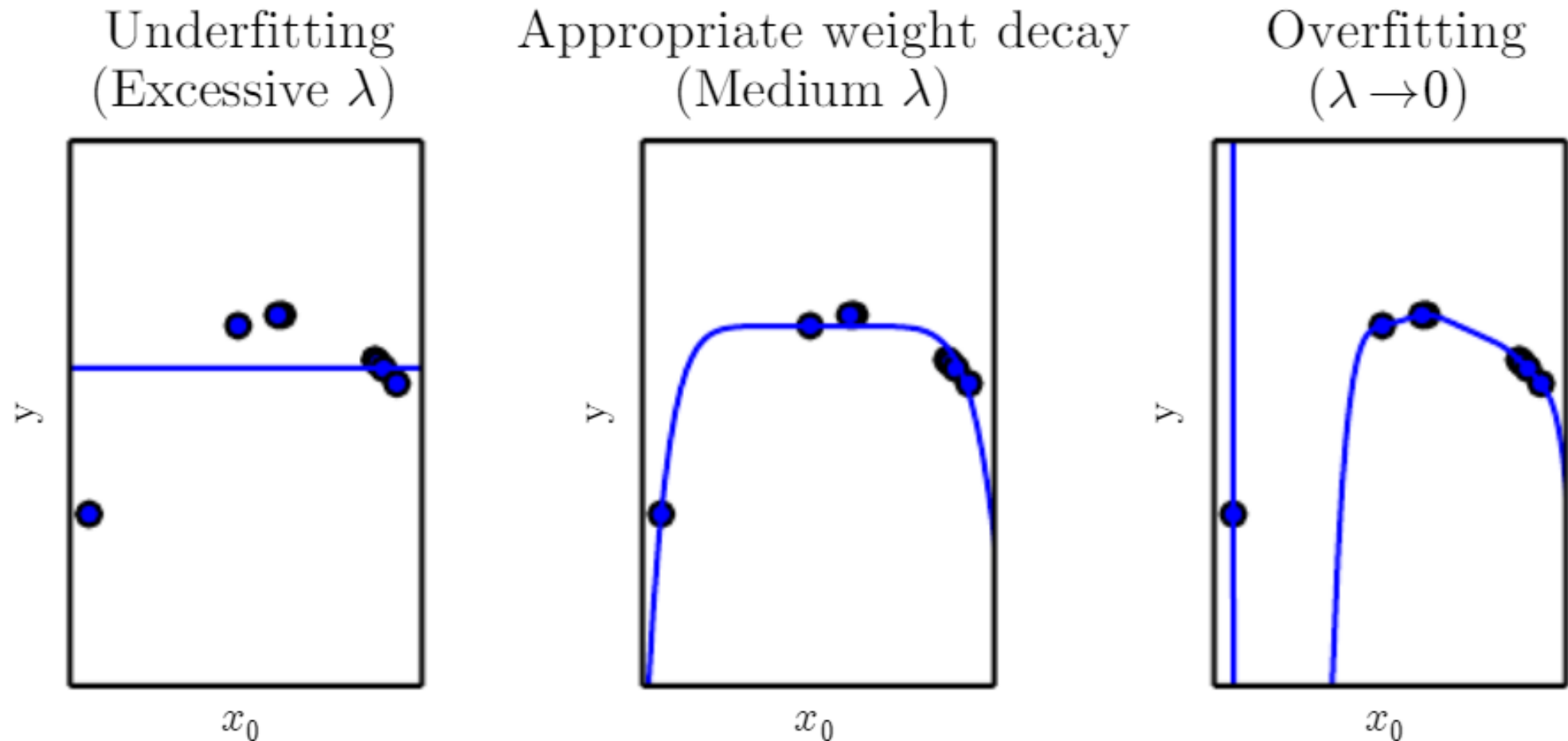
- L1 regularization

$$R(W) = \sum_{i,j} |W_{ij}|$$

- Elastic net (L1 + L2)

$$R(W) = \sum_{i,j} \beta W_{ij}^2 + |W_{ij}|$$

Add term to loss (weight decay)



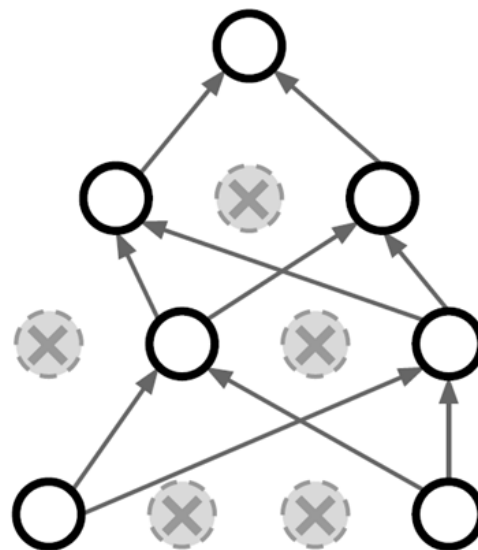
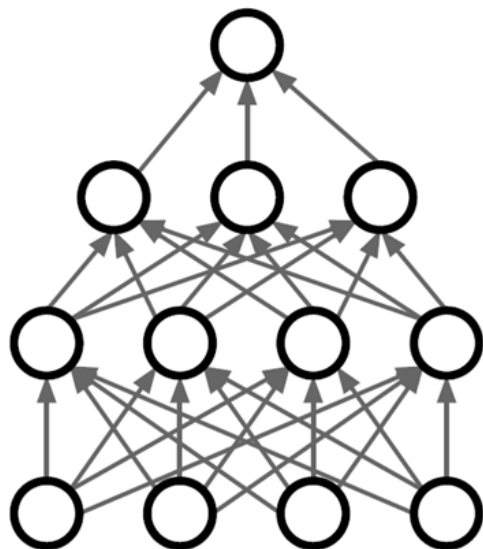
Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016). <http://www.deeplearningbook.org>

Regularization

- Weight decay에서는 linear function이 작은 weight을 갖는게 좋다는 것을 loss function에 추가적인 term을 추가하여 explicit하게 표현함
- 이러한 선호도를 implicit또는 explicit하게 표현하는 방식으로는 많은 방법이 있고 이러한 방식을 regularization이라고 말함
- Regularization의 목적은 generalization error를 줄이는 것임

Drop out

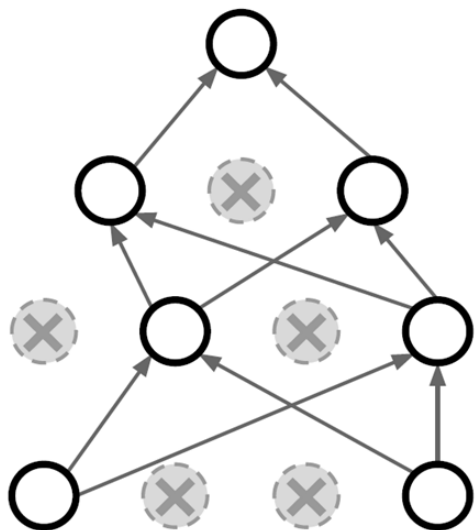
- 미리 정한 확률로 특정 뉴런을 **학습**에서 제외할지 정함
 - 보통 확률을 0.5를 사용함



Srivastava et al, "Dropout: A simple way to prevent neural networks from overfitting", JMLR 2014

How to work?

- Feature의 co-adaptation을 방지하여 네트워크에 불필요한 representation을 제외한다
 - Co-adaptation이란 한 뉴런이 다른 뉴런에 의존적으로 학습되는 현상

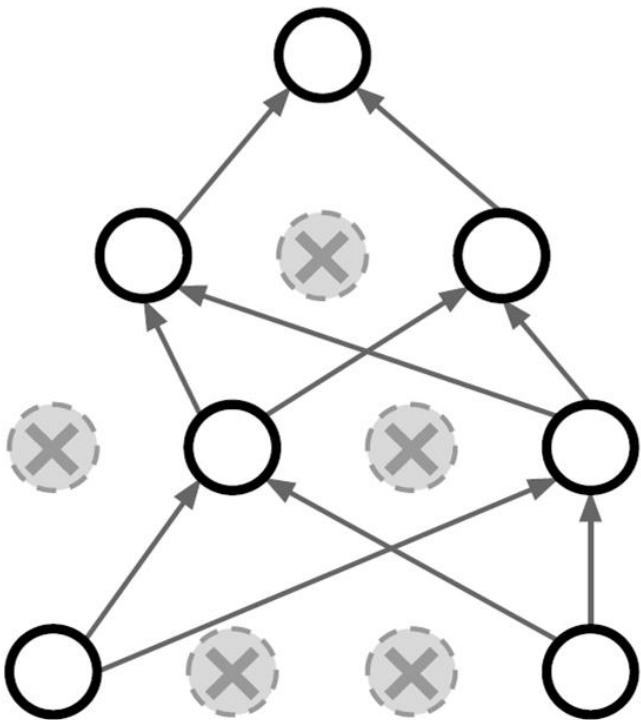


Forces the network to have a redundant representation;
Prevents co-adaptation of features



Another interpretation

- Drop out은 확률적으로 특정 뉴런을 사용하는지 결정
 - 각각의 step마다 다른 network를 학습시키는 것으로 볼 수 있고, 이를 model ensemble관점에서 볼 수 있음

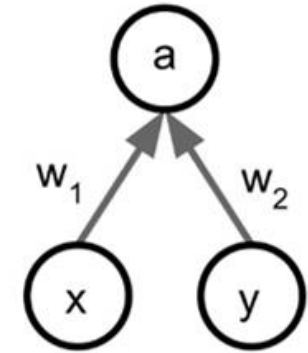


Dropout in test time

- Test time에는 dropout 확률을 곱해줘야 함
 - Test time에서의 output은 train time에서의 output의 기댓값

- Test에서 a 값의 기대값:

$$E[a] = w_1x + w_2y$$



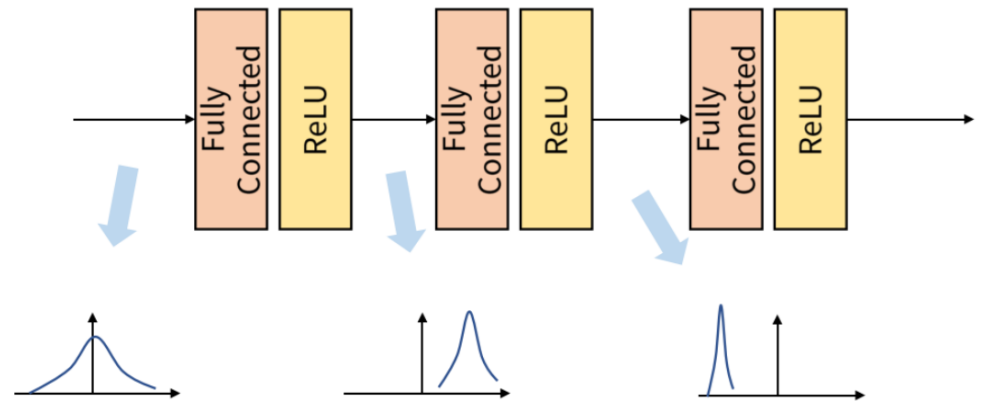
- 학습에서 a 값의 기대값 :

$$E[a] = \frac{1}{4}(w_1x + w_2y) + \frac{1}{4}(w_1x + 0y) + \frac{1}{4}(0x + 0y) + \frac{1}{4}(0x + w_2y) = \frac{1}{2}(w_1x + w_2y)$$

Batch normalization

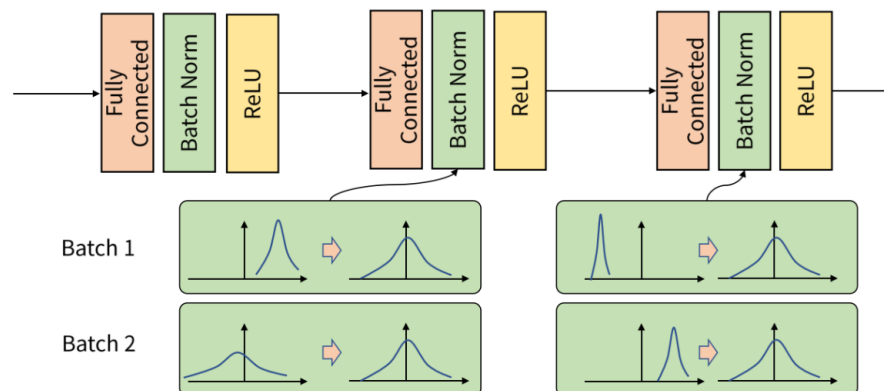
- Internal covariate shift현상의 해결

- 학습 중간에 weight가 업데이트 되면서 학습데이터의 분포가 계속 바뀌는 현상 □ 이로 인해 batch간 데이터 분포의 차이가 발생
- Batch단위로 학습할 경우 하나의 데이터 셋을 다 사용하기 전에 update가 일어나기 때문



Batch normalization

- 이러한 현상을 해결하기 위해 batch 및 layer마다 정규화를 진행



- 학습단계에서는 learning parameter를 이용하여 평균 및 분산을 구하고, 배치 정규화를 진행
- Test 단계에서는 학습에서 사용했던 평균 및 분산을 사용

Table of Contents

- Review the key points
 - Validation set
 - Model Ensembles
 - Regularization (Add term to loss / Batch normalization / Dropout)
- Practice
 - Early stopping
 - Regularization

Early stopping

- Early stopping을 구현해 best epoch 찾아보기

Let n be the number of steps between evaluations.

Let p be the “patience,” the number of times to observe worsening validation set error before giving up.

Let θ_o be the initial parameters.

$\theta \leftarrow \theta_o$

$i \leftarrow 0$

$j \leftarrow 0$

$v \leftarrow \infty$

$\theta^* \leftarrow \theta$

$i^* \leftarrow i$

while $j < p$ **do**

 Update θ by running the training algorithm for n steps.

$i \leftarrow i + n$

$v' \leftarrow \text{ValidationSetError}(\theta)$

if $v' < v$ **then**

$j \leftarrow 0$

$\theta^* \leftarrow \theta$

$i^* \leftarrow i$

$v \leftarrow v'$

else

$j \leftarrow j + 1$

end if

end while

Best parameters are θ^* , best number of training steps is i^* .

Regularization

- Neural Network에 batch normalization 적용
- Neural Network에 dropout 적용
- Loss function에 weight decay 추가(optional)

Thank You :)

saemi@postech.ac.kr