

8) Hyperparameter Tuning

Öncelik sırasına göre: (Genel olarak)

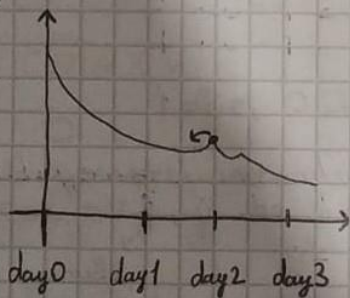
- 1) α
- 2) $\beta \sim 0.9$ Default
- 3) Hidden Units number
- 4) Mini Batch size
- 5) # layers
- 6) Learning rate decay
- 7) $\beta_1, \beta_2, \epsilon$
0.9 0.999 10^{-8} → Defaults

2 farklı şekilde tune edebiliriz:

GPU ve CPU 'müz yeterli değil; yani computational power'ımız

az ise;

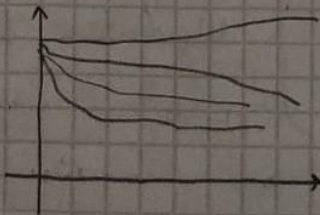
a)



⇒ Babysitting one model
(Gün gün deneyerek)

Computational power var ise;

b)



⇒ Training many models
in parallel.

9) Fitting Batch Norm into a Neural Network (normalization)

For Single Batch

$$\mu = \frac{1}{m} \sum_i z^{(i)}$$

(mini batch size)

$$\sigma^2 = \frac{1}{m} \sum_i (z_i - \mu)^2$$

* We are calculating mean and variance for each mini-batch when applying "Batch Normalization"

$$\underline{z_{norm}^{(i)}} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}} \rightarrow \text{Mean} = 0; \text{sd} = 1.$$

* Fakat; We do not want the hidden units to always have mean 0 and variance 1.

* It makes sense for hidden units to have a different distribution.

$$\underline{\tilde{z}^{(i)}} = \gamma \underline{z_{norm}^{(i)}} + \beta$$

* Datanın dağılımı layer'lar arasında çokça değişir buna "internal covariate shift" denir. Batch Norm bu problemi çözer.

(Training speed artar)

*! Batch normalization applies the normalization process not just to the input layer, but hidden unit values, z .

Tek fark;

X (input features) \rightarrow mean = 0, sd = 1.

$z^{(i)} \rightarrow$ mean = 0, sd = 1 olmasını isteyebiliriz. Bu yüzden mean and sd. γ, β ile kontrol edilir.

(Burun momentum veya adam'daki β ile hiçbir ilgisi yoktur)

$$w^{[l]} = w^{[l]} - \alpha dw^{[l]}$$

$$\beta^{[l]} = \beta^{[l]} - \alpha d\beta^{[l]}$$

$$\gamma^{[l]} = \gamma^{[l]} - \alpha d\gamma^{[l]}$$

} Update Parameters

$\rightarrow X$ ve z 'deki bu normalizasyonlar Eğitimi çok hızlandırır!!

(Cost function'a smoother ve balanced yaptığımız için)

Not \rightarrow Calculate μ, σ^2 on the entire training set or a weighted average across the mini batches and **Use at Test time.**

The idea of our data distribution changing over is called "Covariate shift".

→ Training (X) kümesinin dağılımı ile Y output ürettin. Fakat zamanla "Distribution of X " değişirse yani yeni gelen test setlerin dağılımı X gibi değilse modeli yeniden Train etmek gerekir.

→ Bu problemi Batch norm çözer. Her iterasyonda W, b değişecek ve bu durumda Z' ler de değişecek. Batch norm uyguladığı durumda, the mean and variance will remain the same. Z' lerin gerçek değeri değişse bile ; mean and variance aynı kalacak.