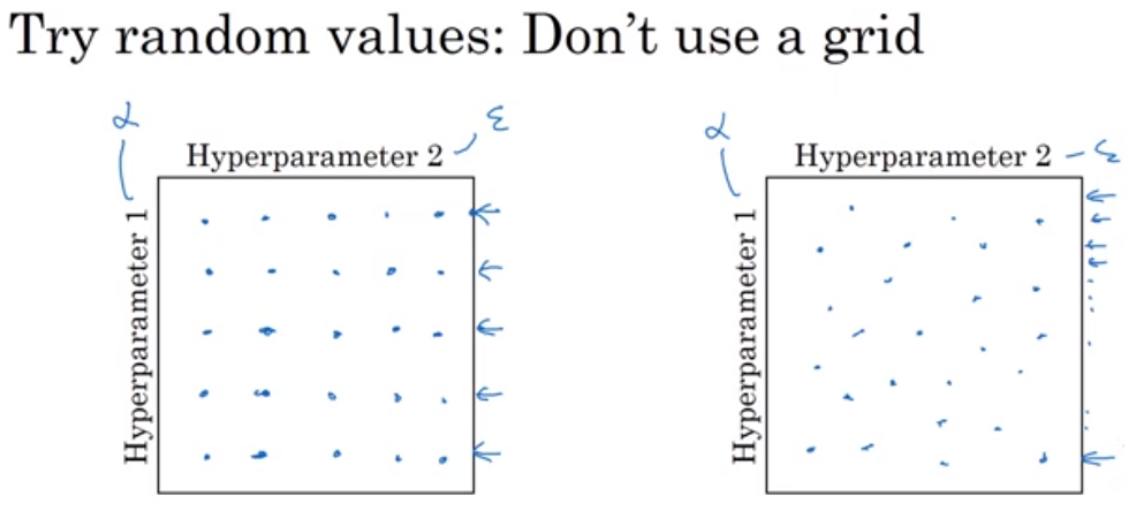
[딥러닝 2단계] 5. 하이퍼파라미터 튜닝

하이퍼파라미터 종류 (중요도순)

- 학습률(α)
  - 가장 중요
- 모멘텀(Momentum) 알고리즘의 β
  - 기본값 0.9
- 은닉 유닛의 수
- 미니배치 크기
- 은닉층의 갯수
- 학습률 감쇠(learning rate decay) 정도
- 아담(Adam) 알고리즘의 $\beta 1$, $\beta 2$, ε
  - 보통 기본값을 사용

튜닝 프로세스

1. 무작위 접근 방식



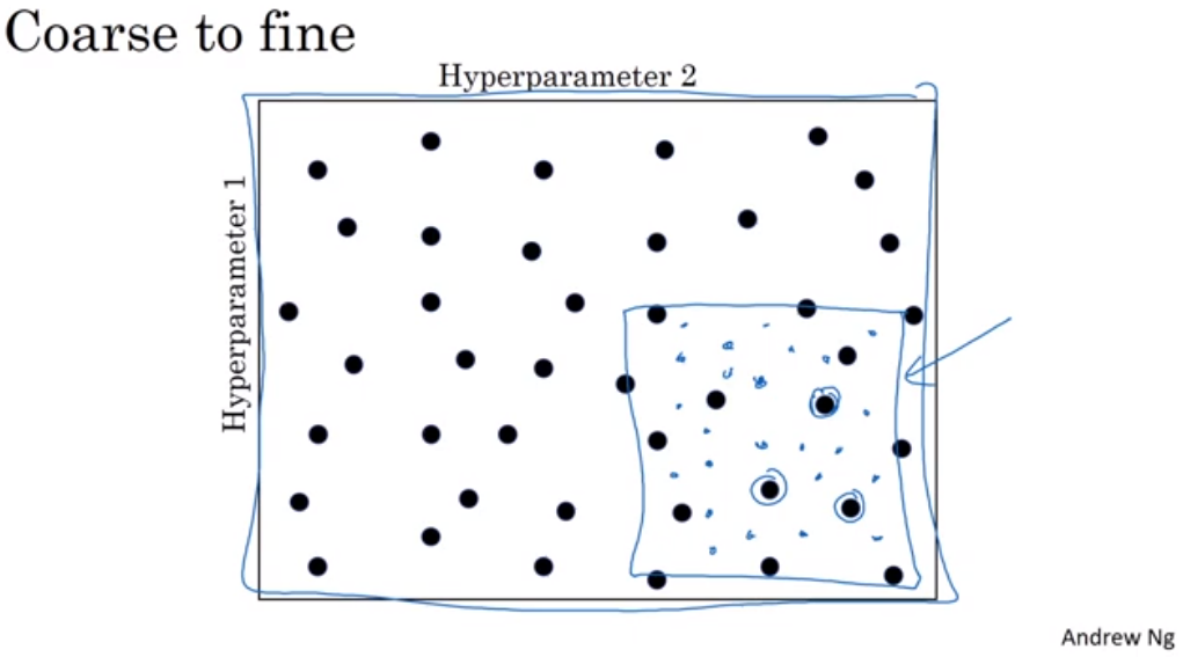
과거에는 왼쪽 방법을 사용했다. 이 방법은 데이터의 수가 적을 때 쓰기 좋다.

현재 딥러닝에서는 오른쪽과 같이 **무작위로 선택된 지점의 값**을 쓴다. 어떤 값이 좋을지 미리 알 수 없기 때문이다.

예를 들어 α와 ε을 튜닝한다고 했을 때, 왼쪽의 방법을 쓰면 5가지의 알파에 대해 훈련하게 되지만, 오른쪽 방법을 쓰면 25가지의 알파 값에 대해 훈련할 수 있다.

어떤 하이퍼파라미터가 가장 핵심적이든 그 하이퍼파라미터의 여러 값에 대해 훈련할 수 있다.

2. 정밀화 접근 방식



전체 공간에서 탐색한 후, 성능이 좋은 구역이 있다면 그 구역 안에서 정밀하게 탐색하는 방법이다.

적절한 척도 선택하기

'무작위'가 모든 값들 중 공평하게 뽑는다는 뜻은 아니다. 적절한 척도를 선택하는 것이 중요하다.

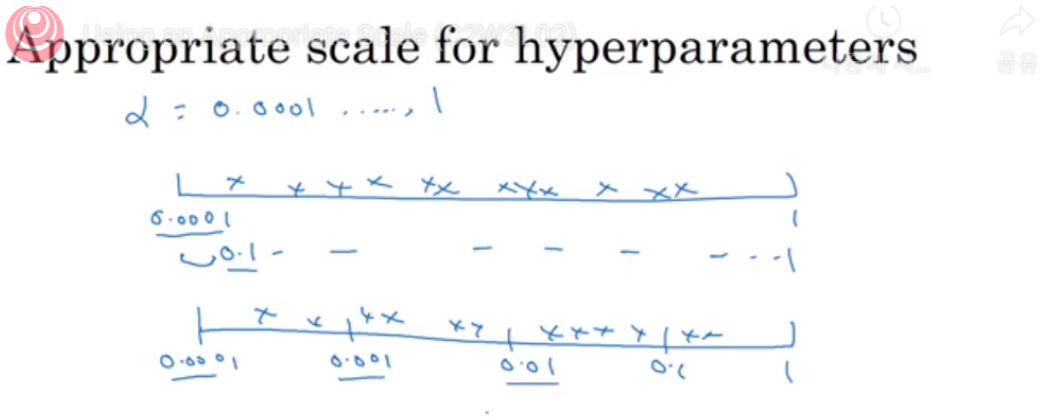
무작위로 뽑는 게 좋은 하이퍼파라미터

은닉 유닛의 수, 은닉층의 수

→ 무작위, Grid search 모두 가능

척도가 필요한 하이퍼파라미터

1. 학습률 알파

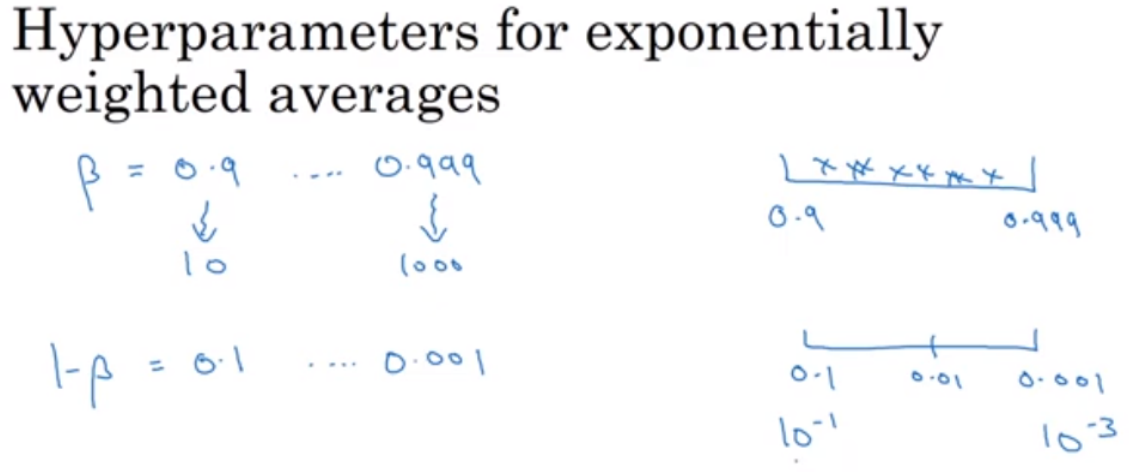


예) 알파를 0.0001~1 사이의 값으로 설정할 때, 일반적인 선형 척도를 이용하면 오직 10%의 값만 0.0001~0.1 사이에 존재한다.

→ **로그 척도**를 이용해 균일한 비율이 나오도록 한다.

```
r = -4 * np.random.rand() # 지수를 랜덤하게 구함
a = math.pow(10, r)
```

2. 지수 가중 이동 평균에서 사용되는 β



베타는 0.9와 0.999 사이의 값인데, 앞과 같이 균일한 범위에서 탐색하기 위해 로그 척도를 이용한다. 다만, 1-베타를 튜닝해 0.1~0.001 사이의 값을 찾는다.

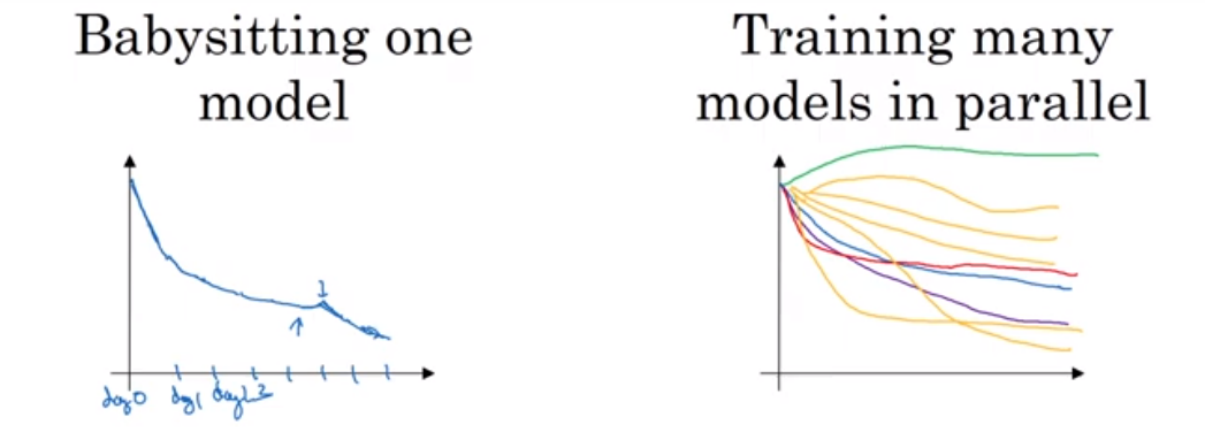
▼ 선형 척도에서 샘플을 뽑는 것이 안 좋은 이유

만약 β가 1에 가깝다면
β가 아주 조금만 바뀌어도 결과가 아주 많이 바뀌게 됩니다
예를 들어 β가 0.9에서 0.9005로 바뀌었다면
결과에 거의 영향을 주지 않습니다
하지만 β가 0.999에서 0.9995로 바뀌었다면
알고리즘의 결과에 큰 영향을 줄 겁니다
이 경우는 대략 10개의 값을 평균내는 것이지만
여기에서는 마지막 1000개 값의 지수가중평균을 내는 것에서
마지막 2000개 값의 평균을 내는 것으로 바뀌었으니까요
왜냐하면 1/(1-β)라는 식이
β가 1에 가까워질수록 작은 변화에도 민감하게 반응하기 때문입니다
따라서 β가 1보다 가까운 곳에서 더 조밀하게 샘플을 뽑습니다
반대로 1-β는 0이 가까운 곳이 되겠지요
따라서 가능한 결과 공간을 탐색할 때
더 효율적으로 샘플을 추출할 수 있는 것입니다

척도를 완벽하게 설정하지 않아도, 정밀화 접근 방식을 사용하면 좋은 하이퍼파라미터를 찾을 수 있다.

실습

하이퍼파라미터를 찾는 과정은 NLP, CV, logistics 등 딥러닝의 분야에 개별적으로 적용된다. 즉, 과거에 찾은 파라미터가 다른 분야에서 잘 작동하지는 않을 수 있다.



1. 모델 돌보기 (Babysitting one model), 판다 접근

컴퓨터의 자원이 많이 필요하지 않거나, 적은 숫자의 모델을 한 번에 학습시킬 수 있을 때 사용

하나의 모델로 매일 성능을 지켜보면서, 학습 속도를 조금씩 바꾸는 방식이다.

온라인 광고, CV 앱과 같이 많은 데이터를 필요로 해서 모델의 크기가 크면 판다 접근을 주로 사용한다.

2. 병렬적으로 여러 모델 훈련, 캐비어 접근

충분한 컴퓨터 자원을 가지고 있다면 사용, 다양한 하이퍼파라미터 테스트 가능