Multiclass Classification

1. Ordinal (Numerical, 서열) Variables vs Nominal (Categorical, 명목) Variables
   값의 순서를 정할 수 있음 / 값의 순서를 정할 수 없음
   EX) IRIS Dataset
   input: length, width → ordinal
   output: species → Nominal

2. One hot, One Cold
   Multiclass에서 N차원의 출력 벡터가 있다고 했을때 해당하는 class는 1, 나머지는 0으로 표현한 벡터
   One hot: 0 1
   One Cold

   EX) $\vec{y} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ 개 / 고양이 / 늑대 / 사슴

3. One Versus Rest Method
   이진분류기 C개를 이용해서 class k와 C-1개 class (Rest)를 분류
   이진 분류기 K의 값을 $h_K(x)$라고 했을때 $\hat{K} = \arg\max_K h_K(x)$
   하지만, 훈련 집합이 불균형

4. One Versus One Method
   각각의 클래스 간의 이진분류 후
   분류된 클래스는 1표 투표, 최종적으로 최다득표한 클래스로 분류
   하지만, 이진 분류기가 $_cC_2$개 필요하므로
   계산복잡도가 C^2으로 OVR Method는 C인 것에 비해 복잡
   대신 훈련집합 불균형 문제는 해결

Softmax Classifier

1. Softmax Function
   $\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$

2. Binary CE
   $-\frac{1}{n}\sum_{i=1}^{n}(y^{[i]}\log a^{[i]} + (1-y^{[i]})\log(1-a^{[i]}))$ n=샘플수, K=class수

3. Multinomial CE and Gradient
   $\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} -y_k^{[i]}\log(a_k^{[i]})$, $\frac{\partial J(\vec{w})}{\partial \vec{w}_k} = \frac{1}{n}\sum_{i=1}^{N}(a_k^{[i]} - y_k^{[i]})\vec{x}^{[i]}$

Underfitting & Overfitting

1. Underfitting (과소적합)
   모델의 용량이 너무 작아서 (파라미터 수가 적어서) 오차가 클 수밖에 없는 현상
   → 용량이 더 큰 모델을 쓴다.

2. Overfitting (과잉적합)

모델의 용량이 너무 커서(파라미터 수가 많아서) 잡음까지 수용 → 테스트 집합에 대한 정확도 ↓

→ 모델이 커도 데이터가 충분히 많으면 됨

Data Augmentation을 통해 데이터 확보

검증 집합을 통해서 최적의 모델 선정

과잉적합된 모델의 가중치는 절댓값이 매우 큼

Weight Penalty 등의 규제 기법을 사용

3. 규제

명시적 규제: 목적함수와 신경망 구조를 '직접' 수정, ex) Weight Penalty, Dropout

암시적 규제: '간접적'으로 영향을 미침 ex) 조기 멈춤, 데이터 증대, 잡음 추가, 앙상블

가중치 감쇠: 목적함수에 규제항 추가

규제항은 큰 가중치에 벌칙을 주어서 가중치를 작게 유지한다.

주로 L_2 Norm, L_1 Norm을 쓴다.

모델의 구조적용량(모델 용량, Layer수, Node수, 등등)은 크게 유지하고

수치적용량(가중치 크기)를 제한

$J_R(\theta) = J(\theta) + \lambda\|\theta\|_2^2$ L_2 Norm → L_1 Norm $\cdots + \lambda\|\theta\|_1$

$\nabla J_R(\theta) = \nabla J(\theta) + 2\lambda\theta$ $\cdots + \lambda\,\mathrm{sign}(\theta)$

↳ θ의 성분의 부호만 가져온다. ex) $[1, -1, \cdots]^T$

업데이트: $\theta = \theta - \rho\nabla J_R(\theta)$

$= \theta - \rho\nabla J(\theta) - 2\rho\lambda\theta$ $= \rho\nabla J(\theta) - \rho\lambda\,\mathrm{sign}(\theta)$

$= (1 - 2\rho\lambda)\theta - \rho\nabla J(\theta)$

즉, θ를 $2\rho\lambda$ 비율로 줄인 후 업데이트 / 고정값 $\rho\lambda$만큼 빼고 업데이트

↳ 희소성 효과: 0이 되는 가중치가 많이 발생

λ는 검증 집합으로 성능을 검증하여 Error가 최소가 되는 λ를 고른다.

가중치 감쇠 과정은 bias에는 영향이 미치지 않도록 한다.