기계학습 (Machine Learning)

L07

- Multiclass Classification

한밭대학교

정보통신공학과

최 해 철

국립 한밭대학교

VISUAL MEDIA LAB.
HANBAT NATIONAL UNIVERSITY

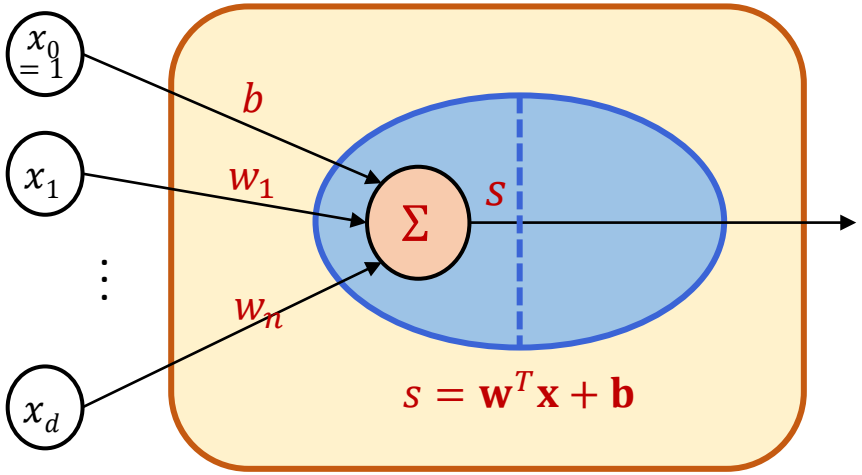
ToC

- Activation Functions
- Multiclass Classification
- Softmax Classification

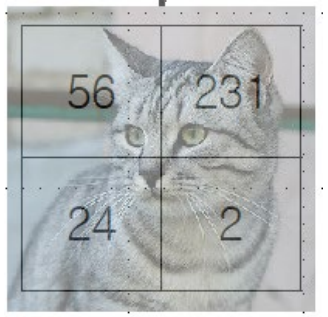
Activation Function

오일석, 기계학습, 3.3.2 활성함수

Linear Classification



$$h_{\boldsymbol{\theta}}(\mathbf{x}) = w_1 x_1 \ldots + w_d x_d + b$$
$$= \mathbf{w}^T\mathbf{x} + b = \hat{y} = o$$



Input image

0.2	-0.5	0.1	2.0

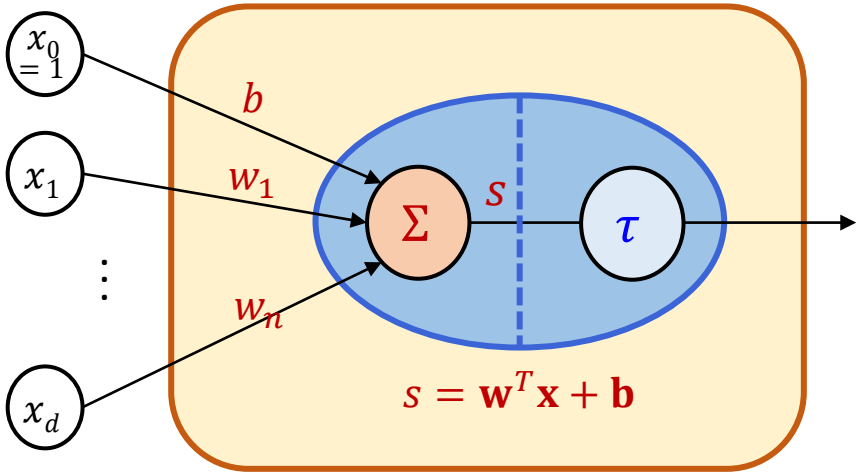
$\mathbf{w}$

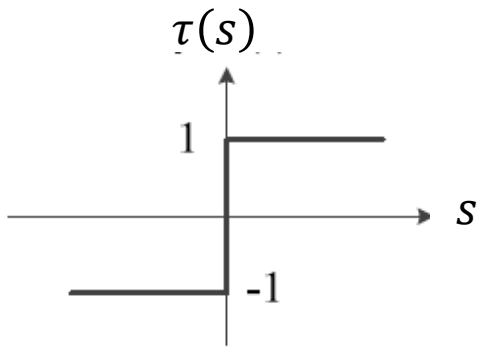
56
231
24
2

$+$ | 1.1 | b $=$ | -96.8 |

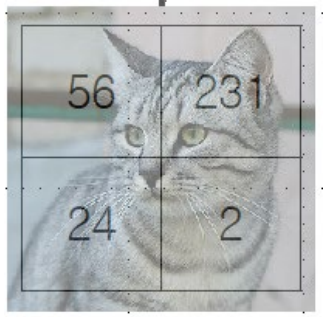
logit

Perceptron with Step Function





$$\tau(s) = \begin{cases} 1 & s \geq 0 \\ -1 & s < 0 \end{cases}$$



Input image

0.2	-0.5	0.1	2.0

$\mathbf{w}$

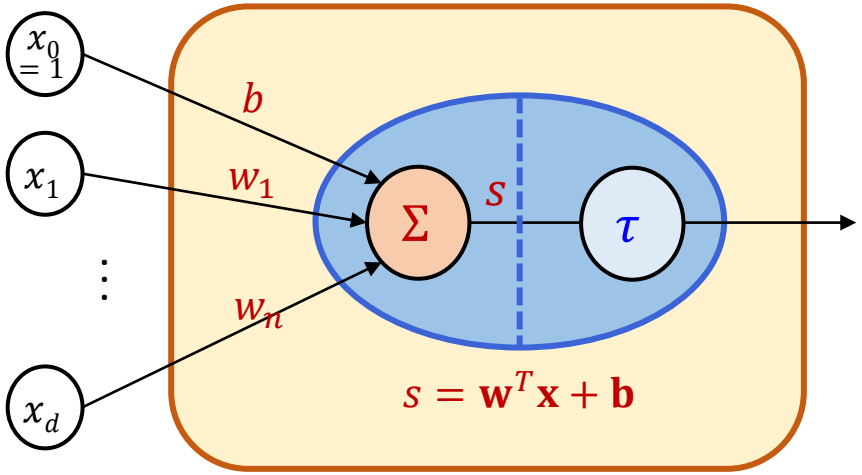
56
231
24
2

$+$ | 1.1 | $=$ | -96.8 | ⇒ | -1 |

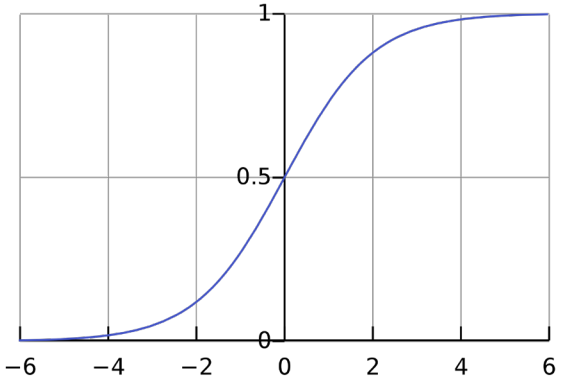
b

Logistic Regression with Logistic Sigmoid Function

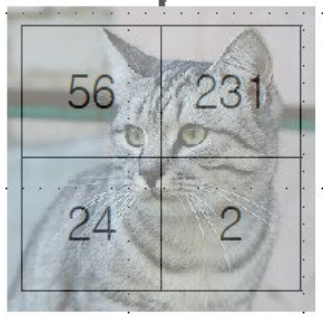
Logistic regression은 추후 학습 예정



$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \tau(w_1 x_1 \ldots + w_d x_d + b)$$
$$= \tau(\mathbf{w}^T\mathbf{x} + b) = \hat{y} = o$$



$$\tau(z) = \frac{1}{1 + e^{-z}}$$



Input image

0.2	-0.5	0.1	2.0

$\mathbf{w}$

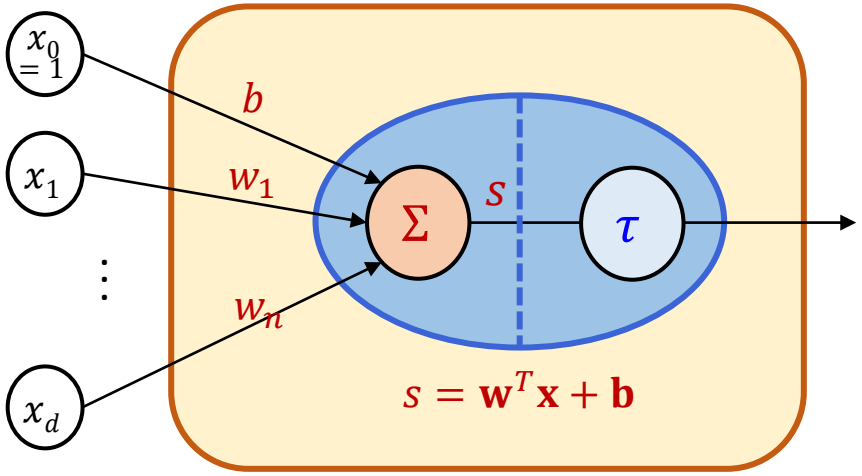
56
231
24
2

$+$ | 1.1 | $=$ | -96.8 | $\Rightarrow$ | 9.12×10^{-43} |

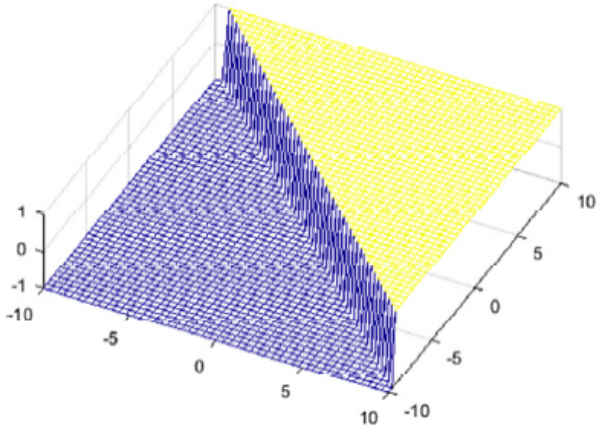
b

$y \in (0, 1)$

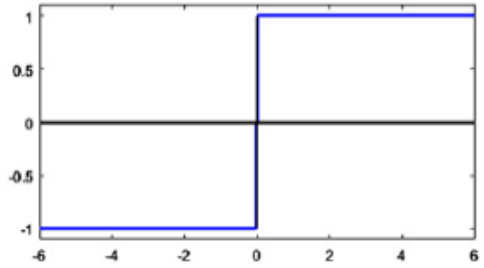
Activation Functions



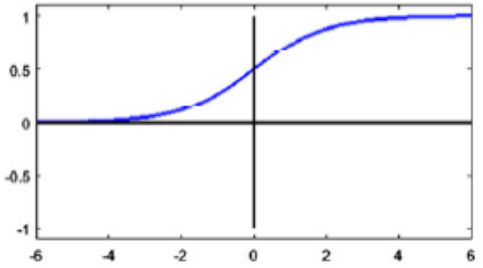
$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \tau(w_1 x_1 \ldots + w_d x_d + b)$$
$$= \tau(\mathbf{w}^T\mathbf{x} + b) = \hat{y} = o$$



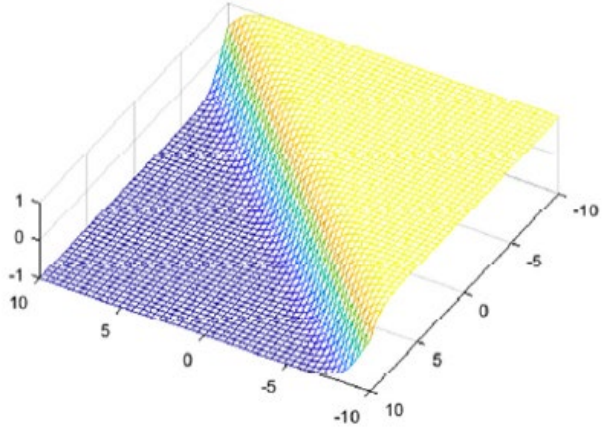
(a) 계단함수의 딱딱한 공간 분할



(a) 계단 함수



(b) 로지스틱 시그모이드



(b) 로지스틱 시그모이드의 부드러운 공간 분할

그림 3-13 퍼셉트론의 공간 분할 유형

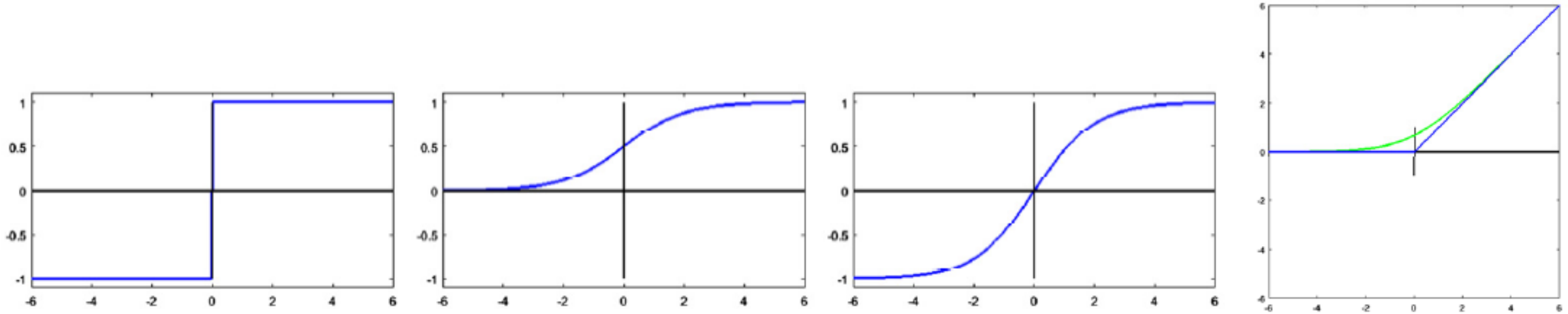
Perceptron

Logistic regression

Activation Functions

◆ 딱딱한hard 공간 분할과 _____ 공간 분할

- 계단함수는 딱딱한 의사결정(영역을 점으로 변환). 나머지 활성함수는 부드러운 의사결정(영역을 영역으로 변환)



(a) 계단 함수 (b) 로지스틱 시그모이드 (c) 하이퍼볼릭 탄젠트 시그모이드 (d) softplus와 rectifier

그림 3-12 신경망이 사용하는 활성함수

Activation Function

◆ 신경망이 사용하는 다양한 활성함수

- 로지스틱 시그모이드와 하이퍼볼릭 탄젠트는 s가 커질수록 계단함수에 가까워짐
- 모두 1차 도함수 계산이 빠름 (특히 ReLU는 비교 연산 한 번)
- 퍼셉트론은 계단함수, 다층 퍼셉트론은 로지스틱 시그모이드와 하이퍼볼릭 탄젠트, 딥러닝은 ReLU를 사용

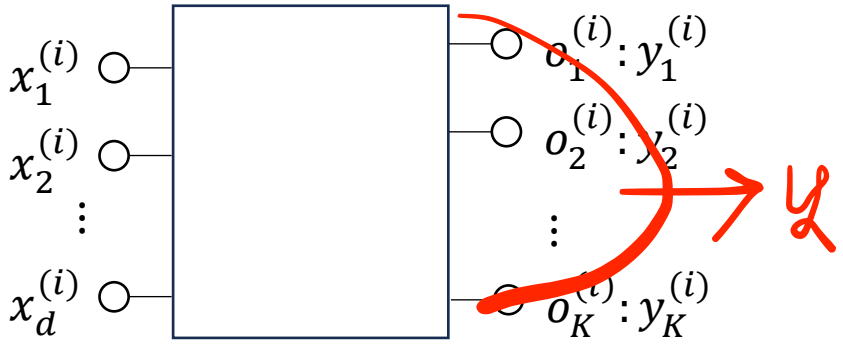
표 3-1 활성함수로 사용되는 여러 함수

함수 이름	함수	1차 도함수	범위
계단	$\tau(s)=\begin{cases}1 & s\geq 0\\ -1 & s<0\end{cases}$	$\tau'(s)=\begin{cases}0 & s\neq 0\\ \text{불가} & s=0\end{cases}$	-1과 1
로지스틱 시그모이드	$\tau(s)=\frac{1}{1+e^{-as}}$	$\tau'(s)=a\tau(s)(1-\tau(s))$	(0,1)
하이퍼볼릭 탄젠트	$\tau(s)=\frac{2}{1+e^{-as}}-1$	$\tau'(s)=\frac{a}{2}(1-\tau(s)^2)$	(-1,1)
소프트플러스	$\tau(s)=\log_e(1+e^s)$	$\tau'(s)=\frac{1}{1+e^{-s}}$	$(0,\infty)$
렉티파이어(ReLU)	$\tau(s)=\max(0,s)$	$\tau'(s)=\begin{cases}0 & s<0\\ 1 & s>0\\ \text{불가} & s=0\end{cases}$	$[0,\infty)$

Multiclass Classification

Binary Classification vs. Multiclass Classification

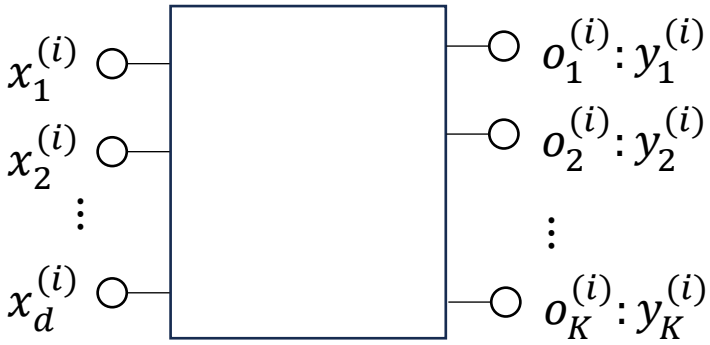
Binary Classification



훈련 집합: $\{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})\}$

$\mathbf{x}^{(i)} =$

Multiclass Classification



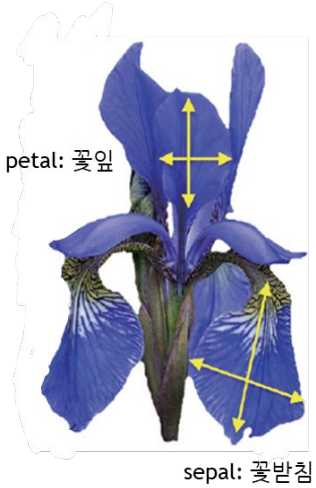
훈련 집합: $\{(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), (\mathbf{x}^{(2)}, \mathbf{y}^{(2)}), \ldots, (\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}$

$$\mathbf{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \ldots \\ x_d^{(i)} \end{bmatrix} \qquad \mathbf{y}^{(i)} =$$

Multiclass Classification

◆ Multiclass (Multinomial) Classification

- 세 개 이상의 클래스로 분류하는 문제
  - 출력이 범주형(categorical)
- 라벨은 더 이상 binary가 아니며, 다중 명목형(Multinomial)이다.: e.g. $y \in$ _____.
  - Nominal data: 범주 간에 순서나 순위가 없음. 범주 간 동등 관계 (L05에서 다룸)



	color	size	price	classlabel
0	green	2	10.1	class1
1	red	3	13.5	class2
2	blue	5	15.3	class1

Review L05) One-Hot Encoding

Categorical Data -> Nominal Data (Class Labels)

- `.get_dummies()` : 범주형 데이터를 가지고 있는 열(칼럼)을 원-핫(one-hot) 인코딩하는 데에 사용

	color	size	price	classlabel
0	green	2	10.1	0
1	red	3	13.5	1
2	blue	5	15.3	0

```
pd.get_dummies(df)          #pd.get_dummies(df, columns=['color'])
```

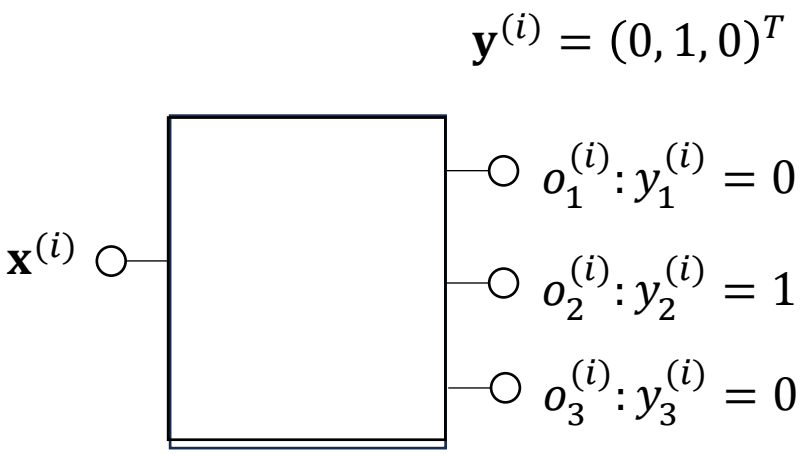
	size	price	classlabel	color_blue	color_green	color_red
0	2	10.1	0	0	1	0
1	3	13.5	1	0	0	1
2	5	15.3	0	1	0	0

HANBAT NATIONAL UNIVERSITY
VISUAL MEDIA LAB. HANBAT NATIONAL UNIVERSITY

One-Hot Encoding

◆ One-Hot Encoding in ML

- k 번째 class의 target vector를 k 번째 자리는 1, 나머지는 0이 되도록 설정
- cross entropy 계산에 적합해 짐 (추후 cross entropy 학습 예정)
  - Target $\boldsymbol{y}$의 원소들의 합이 1이 되므로 각 원소를 그 class의 정답 확률 로 볼 수 있다.



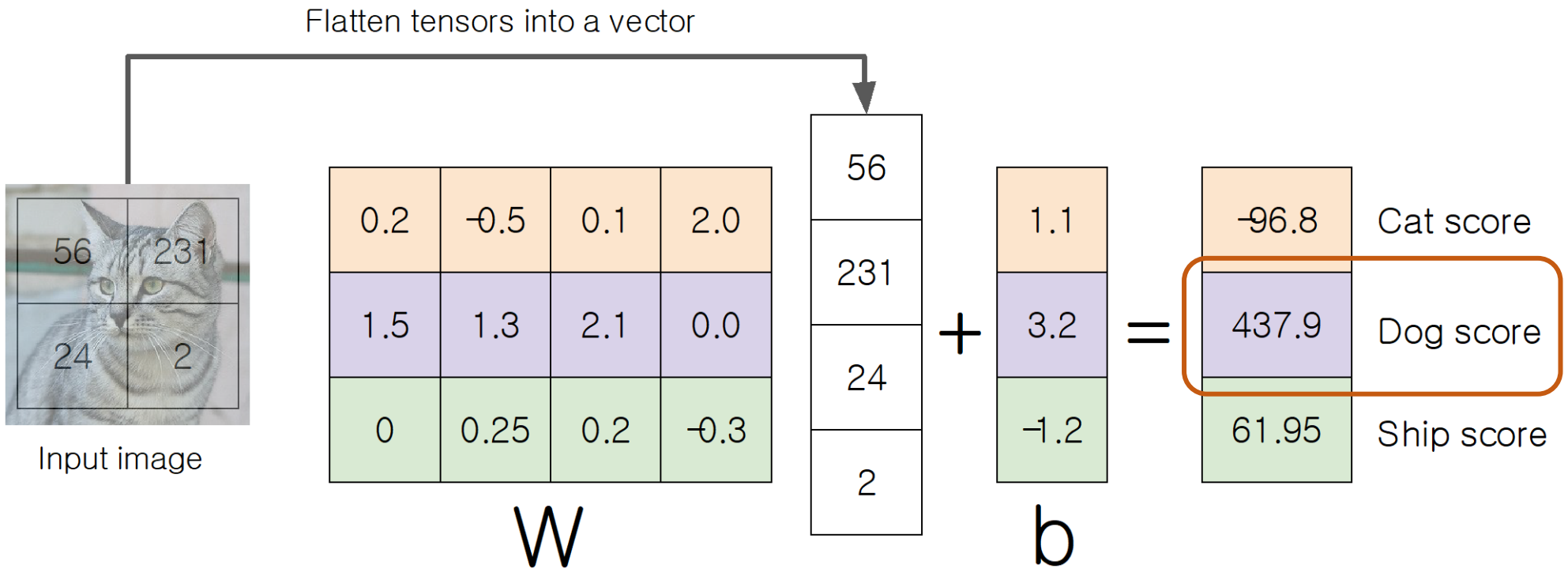
Index	Job
1	Police
2	Doctor
3	Student
4	Teacher
5	Driver

One hot encoded data
[1 0 0 0 0]
[0 1 0 0 0]
[0 0 1 0 0]
[0 0 0 1 0]
[0 0 0 0 1]

Multiclass Classification with Multiple Binary Classifiers

◆ One-versus-Rest (One-versus-All) Method

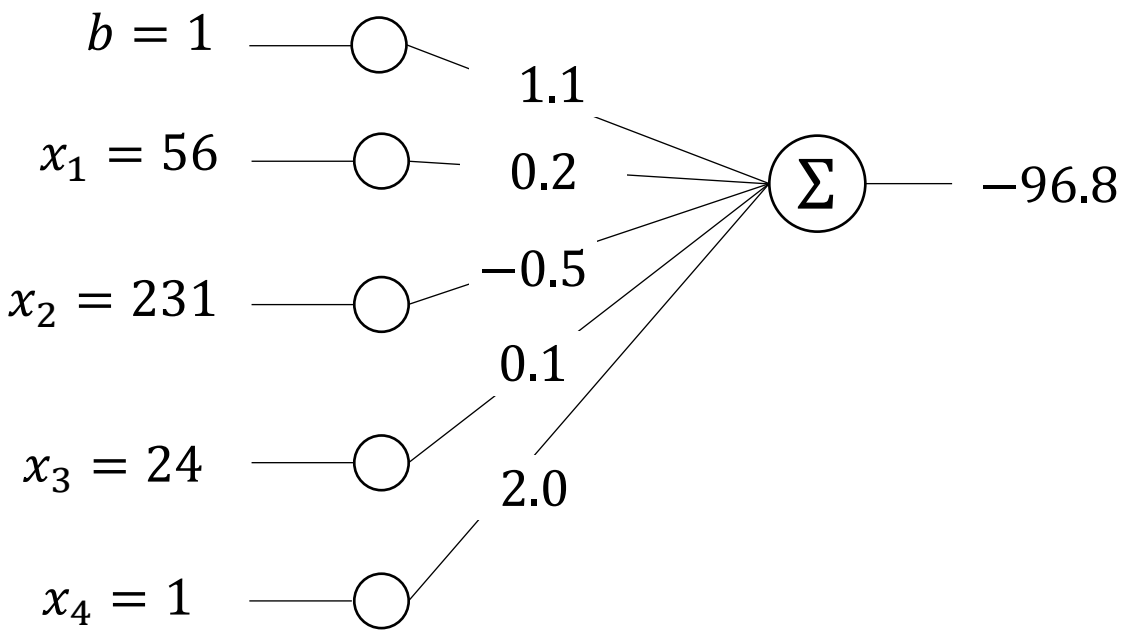
- 이진 분류기 c 개를 독립적으로 사용하여 class k와 나머지 $c-1$ 개 class를 분류
- Class k에 대한 이진 분류기를 h_k라 하면, $h_k(\mathbf{x})$가 가장 큰 값을 갖는 k로 분류함



Multiclass Classification with Multiple Binary Classifiers

◆ One-versus-Rest (One-versus-All) Method

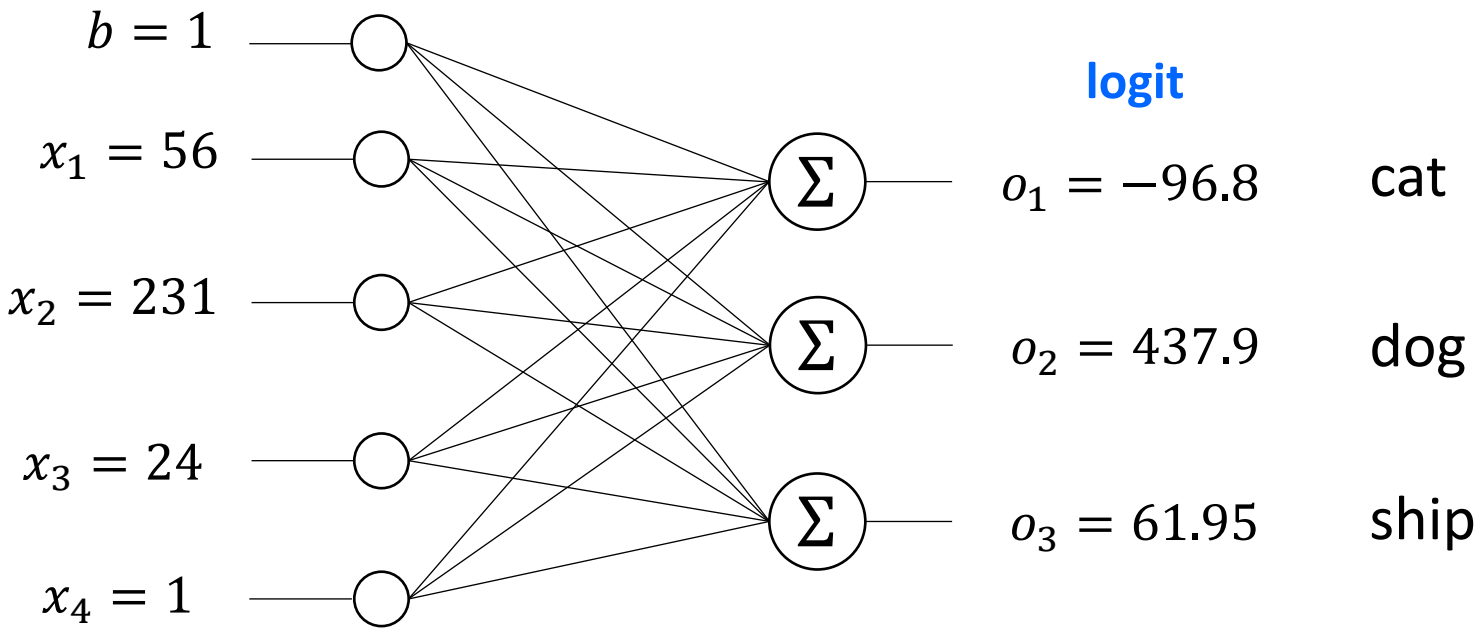
- 이진 분류기 c 개를 독립적으로 사용하여 class k와 나머지 $c-1$ 개 class를 분류 ($1:\boldsymbol{c-1}$)
- Class k에 대한 이진 분류기를 h_k라 하면, $h_k(\mathbf{x})$가 가장 큰 값을 갖는 k로 분류함



Multiclass Classification with Multiple Binary Classifiers

◆ One-versus-Rest (One-versus-All) Method

- 이진 분류기 c 개를 독립적으로 사용하여 class k와 나머지 $c-1$ 개 class를 분류 ($1:\boldsymbol{c-1}$)
- Class k에 대한 이진 분류기를 h_k라 하면, $h_k(\mathbf{x})$가 가장 큰 값을 갖는 k로 분류함

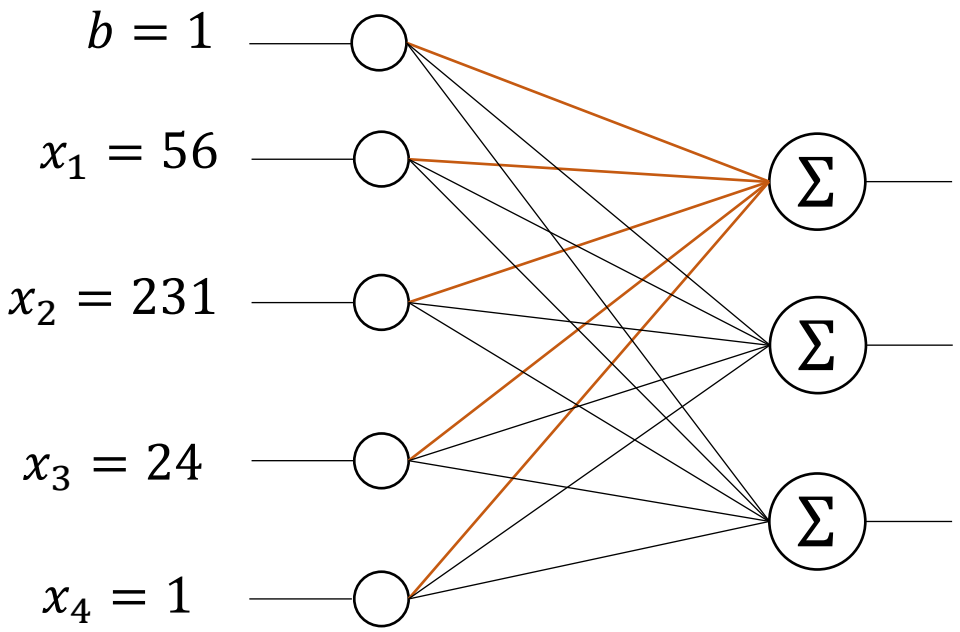


$\hat{k} =$

Multiclass Classification with Multiple Binary Classifiers

◆ One-versus-Rest (One-versus-All) Method

- 필요한 이진 분류기의 개수: c 개
- 각 이진 분류기에 대해 훈련집합의 불균형 을 일으킴 (class k 샘플수 ≪ 나머지 샘플수)



logit

$o_1 = -96.8$ cat

$o_2 = 437.9$ dog

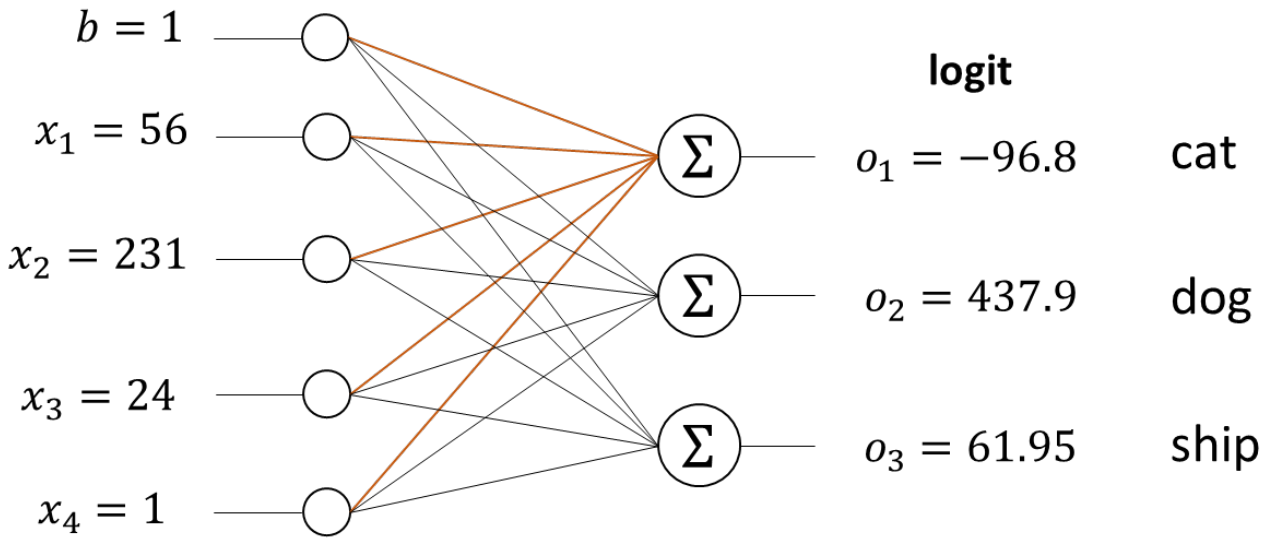
$o_3 = 61.95$ ship

$$\hat{k} = \arg\max_k h_k(\mathbf{x})$$

Multiclass Classification with Multiple Binary Classifiers

◆ One-versus-One Method

- 이진 분류기 $\mathrm{C}(c,2)$개를 독립적으로 사용하여 class k 와 class l 을 분류 1 : 1
  - $\mathrm{C}(c,2) = \dfrac{c!}{(c-2)!\,2!} = c(c-1)/2$
- 가장 많은 이진 분류기가 선택(투표)한 class를 최종 결과로 결정



$h_{(k,l)}(\mathbf{x})$: Class k와 l을 비교하는 이진 분류기

$h_{(1,2)}(\mathbf{x}) = \mathbf{2}$

$h_{(1,3)}(\mathbf{x}) = \mathbf{3}$

$h_{(2,3)}(\mathbf{x}) = \mathbf{2}$

voting → **2**

Multiclass Classification with Multiple Binary Classifiers

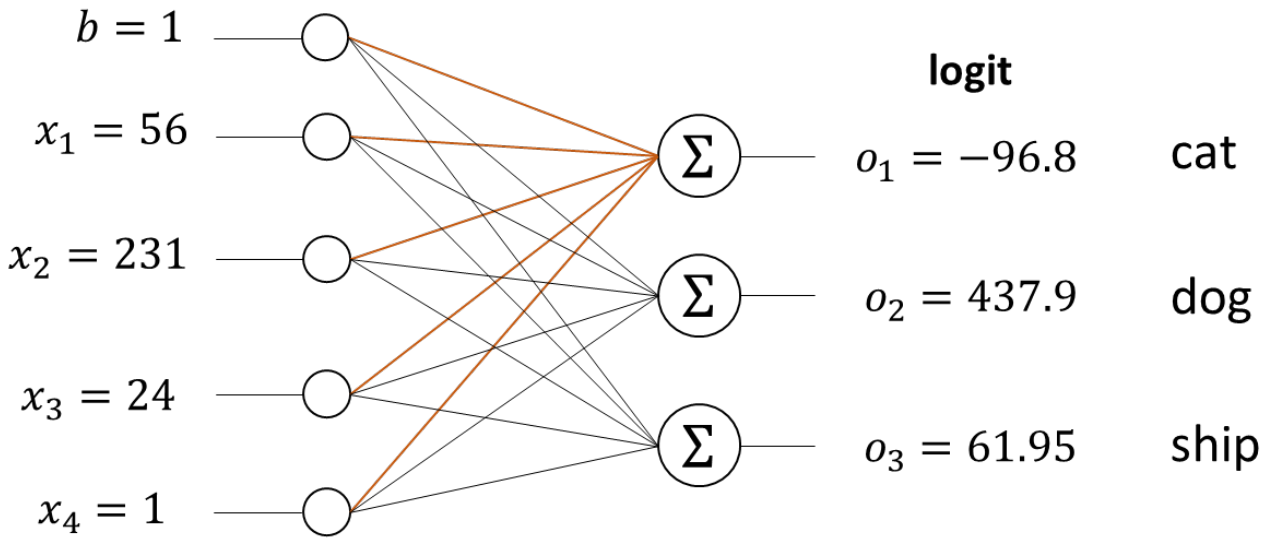
◆ One-versus-One Method

- **이진 분류기 $C(c,2)$개**를 **독립적으로** 사용하여 **class k 와 class l** 을 분류 (1 : 1)
  - $$C(c,2) = \frac{c!}{(c-2)!\,2!} = c(c-1)/2$$
- 가장 많은 이진 분류기가 선택(투표)한 class를 최종 결과로 결정
  - Class k와 l 비교하는 이진 분류기를 $h_{(k,l)}(\mathbf{x})$라 하자.
  - $h_{(k,l)}(\mathbf{x})$가 class k(또는 l)를 출력하면, class k(또는 l)에 한 표를 추가.
  - C(c, 2)개 이진 분류기에 대해 가장 많은 표를 획득한 class를 최종 결과로 결정
    - ✓ 최대 표의 개수: c -1
    - ✓ 비유) 야구나 축구 리그에서 가장 승리를 많이 한 팀이 우승

Multiclass Classification with Multiple Binary Classifiers

◆ One-versus-One Method

- 훈련집합의 불균형을 일으키지 않음: class k 샘플수 ≈ class l 샘플수
- 사용되는 이진 분류기의 개수: $c(c-1)/2$ → c^2에 비례: 높은 training/testing 복잡도



$h_{(k,l)}(\mathbf{x})$: Class k와 l을 비교하는 이진 분류기

$h_{(1,2)}(\mathbf{x}) = \mathbf{2}$

$h_{(1,3)}(\mathbf{x}) = \mathbf{3}$

$h_{(2,3)}(\mathbf{x}) = \mathbf{2}$

voting → **2**

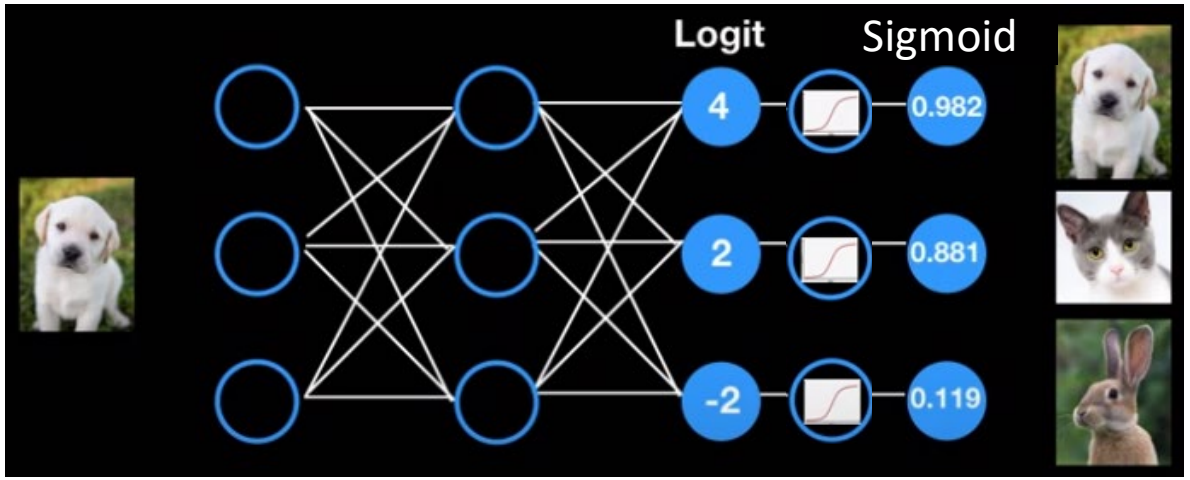
Softmax Classification

Motivation

◆ (-∞, ∞)의 출력값(logit)을 다중 클래스 분류기를 위한 확률로 변환할 수 없을까?

- *What we want* in the output layer
  : conditional probabilities P(y | x)

- *Sigmoid* activations in the output layer
  : do Not sum up to 1

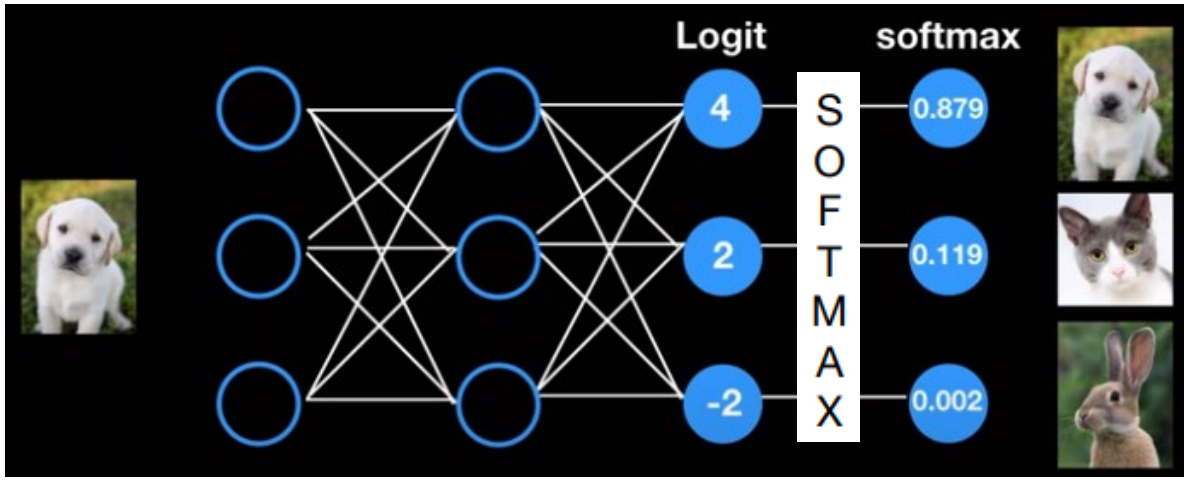


$\mathbf{x}$ $\mathbf{s}$ $\tau(\mathbf{s})$

Motivation

◆ (-∞, ∞)의 출력값(logit)을 다중 클래스 분류기를 위한 확률로 변환할 수 없을까?

- *Softmax* activations in the output layer
  - *do* sum up to 1
  - suits well to Cross-Entropy Loss
    - ✓ 출력벡터의 원소들의 합이 1이 되므로 각 원소를 그 class의 확률 추정치로 간주
    - ✓ *Cross-Entropy Loss* (L08에서 학습 예정)



$\mathbf{x}$ $\mathbf{s}$ $\tau(\mathbf{s})$

Softmax Function

◆ Softmax Function

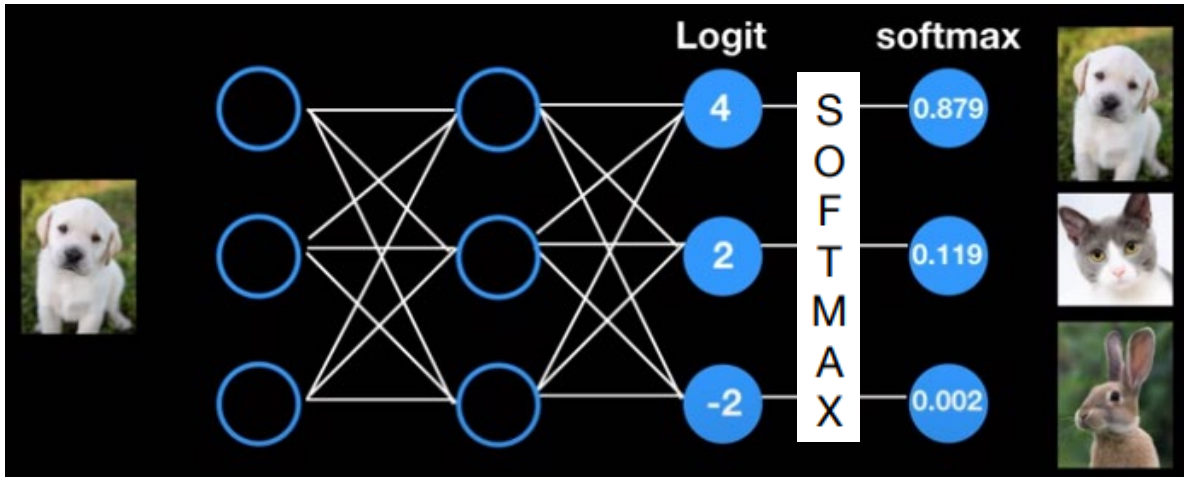
● K개의 실수 값을 갖는 입력 벡터 $\mathbf{s}$에 대해, K개의 확률을 갖는 확률 분포로 정규화

$$\tau: \mathbb{R}^K \rightarrow (0,1)^K$$

$$\tau(s_k) = \frac{e^{s_k}}{\sum_{k=1}^{K} e^{s_k}}$$

$P(y = k|\mathbf{x} = \mathbf{x}^{(i)})$

for $i = 1,2,\ldots,K$ and $\mathbf{s} = (s_1, s_2, \ldots, s_K) \in \mathbb{R}^K$



$\mathbf{x}$ $\mathbf{s}$ $\tau(\mathbf{s})$

출력벡터의 원소들의 합이 이 1인 확률: $\sum_{k=1}^{K} \tau(s_k) = 1$

문헌에 따라 τ대신 σ를 사용

Softmax Classifier 예제

Want to interpret raw classifier scores as **probabilities**

cat	**3.2**
car	5.1
frog	-1.7

s_k

Softmax Classifier 예제

Want to interpret raw classifier scores as **probabilities**

$$o = f(x_i; W)$$

$$P(Y = k|X = x_i) = \frac{e^{s_k}}{\sum_j e^{s_j}}$$

Softmax Function

Probabilities must be >= 0

cat	**3.2**	exp →	**24.5**
car	5.1		164.0
frog	-1.7		0.18

unnormalized probabilities

s_k e^{s_k}

Softmax Classifier 예제

Want to interpret raw classifier scores as **probabilities**

$$o = f(x_i; W)$$

$$P(Y = k|X = x_i) = \frac{e^{s_k}}{\sum_j e^{s_j}}$$

Softmax Function

Probabilities must be >= 0

Probabilities must sum to 1

		exp		normalize	
cat	**3.2**	→	**24.5**	→	**0.13**
car	5.1	→	164.0	→	0.87
frog	-1.7	→	0.18	→	0.00
			unnormalized probabilities		probabilities
	s_k		e^{s_k}		$\frac{e^{s_k}}{\sum_{k=1}^{K} e^{s_k}}$

Softmax Classifier 예제

Want to interpret raw classifier scores as **probabilities**

$$o = f(x_i; W)$$

$$P(Y = k|X = x_i) = \frac{e^{s_k}}{\sum_j e^{s_j}}$$

Softmax Function

Probabilities must be >= 0

Probabilities must sum to 1

	Unnormalized log-probabilities / logits	exp	unnormalized probabilities	normalize	probabilities
cat	**3.2**		**24.5**		**0.13**
car	5.1		164.0		0.87
frog	-1.7		0.18		0.00

s_k

e^{s_k}

$$\frac{e^{s_k}}{\sum_{k=1}^{K} e^{s_k}}$$

감사합니다.