

# Session 5&6

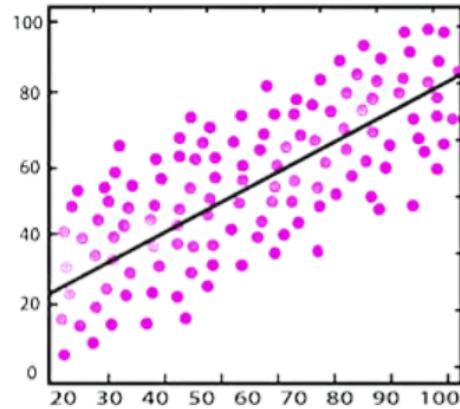
## Classification with MLP

**Deep Learning | Zahra Amini**

Telegram: @zahraamini\_ai & Instagram:@zahraamini\_ai & LinkedIn: @zahraamini-ai

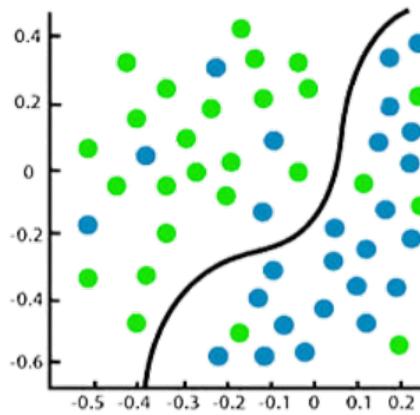
<https://zil.ink/zahraamini>

# Classification



Regression

versus



Classification

Binary  $\rightarrow$

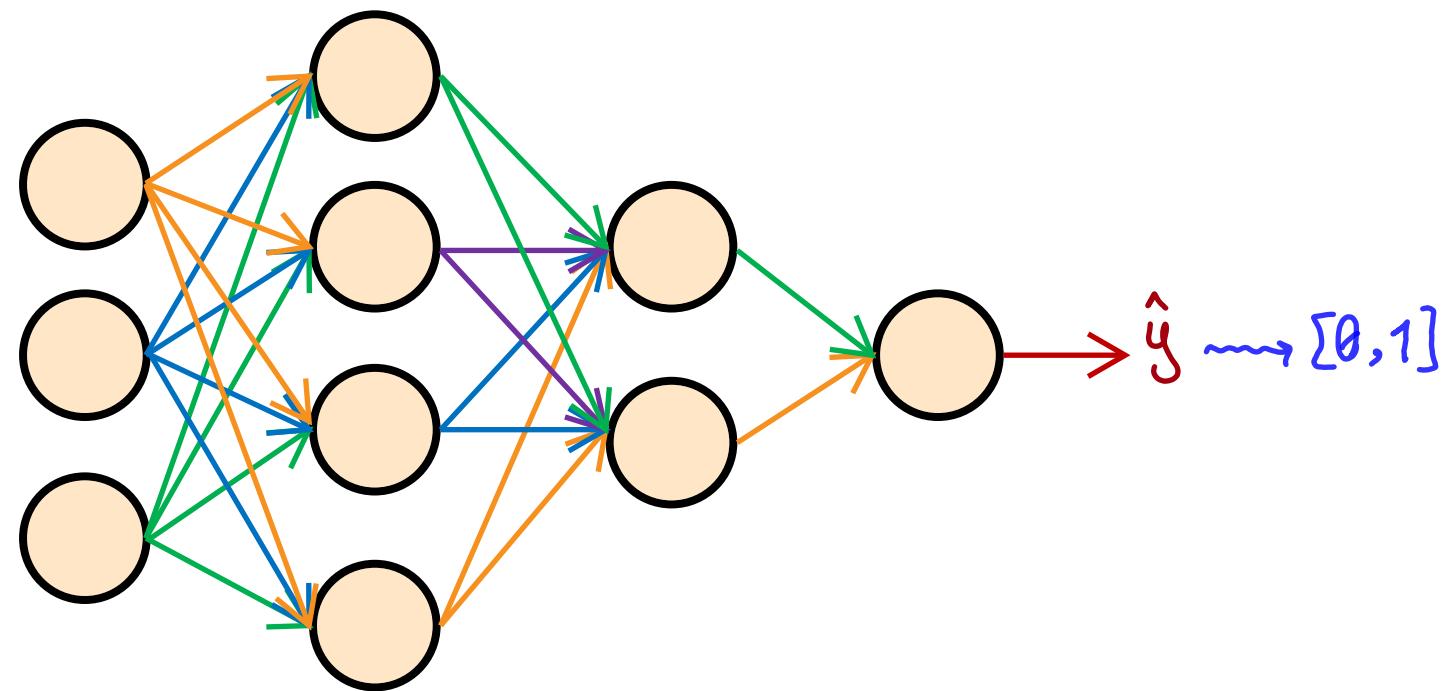
نوعی از مسائل یادگیری ماشین است که در آن هدف تقسیم داده‌ها به دو کلاس مجزا می‌باشد. این نوع از مسائل معمولاً به شکل سوالات بله یا خیر، درست یا غلط، یا دسته‌بندی به دو گروه مجزا مطرح می‌شود

Classification

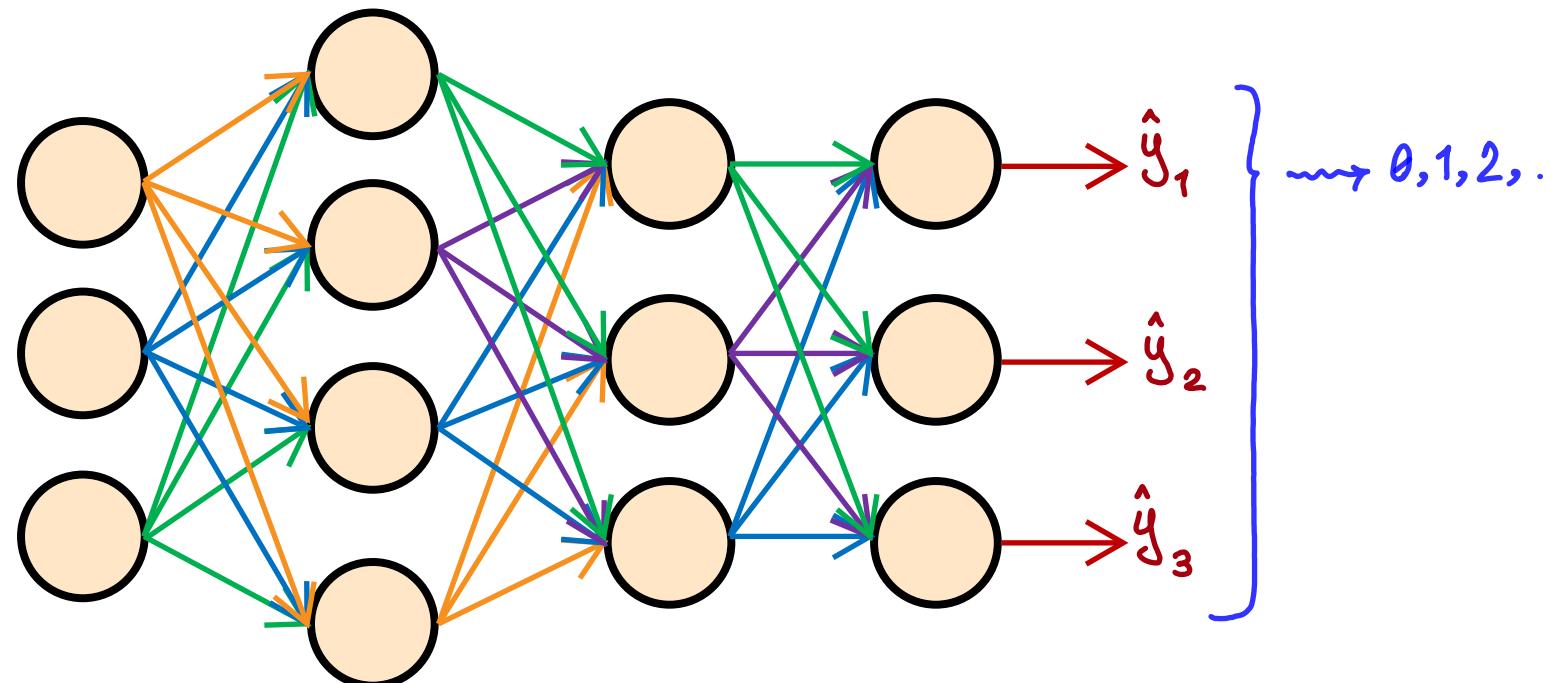
Multiclass  $\rightarrow$

نوعی از مسائل یادگیری ماشین است که در آن هدف تقسیم داده‌ها به بیش از دو کلاس مجزا می‌باشد. در این نوع مسائل، هر نمونه می‌تواند به یکی از چندین کلاس ممکن تعلق داشته باشد

# Binary



# Multiclass



# احتمال یا Probability



احتمال یک مفهوم اساسی در ریاضیات و آمار است که میزان وقوع یک رویداد خاص را در شرایط مشخص توصیف می‌کند. احتمال به صورت یک عدد بین ۰ و ۱ بیان می‌شود

احتمال ۰ به این معناست که رویداد مورد نظر هرگز رخ نمی‌دهد

احتمال ۱ به این معناست که رویداد مورد نظر همیشه رخ می‌دهد

$$P(A) = \frac{\text{تعداد دفعات رخداد } A}{\text{تعداد کل حالات ممکن}}$$

اگر  $A$  یک رویدار ماترد، احتمال وقوع  $A$   $P(A)$  تا اندازه می‌شود

مثال: فرض کنید یک تاس شش وجهی داریم و می‌خواهیم احتمال آمدن عدد ۳ را محاسبه کنیم  $\rightarrow P(3) = ?$

$$P(A) = \frac{\text{تعداد دفعات رخداد } A}{\text{تعداد کل حالات ممکن}} = \frac{1}{6} = 0.167$$

✓ تعداد حالات ممکن برابر با ۶ است (اعداد ۱ تا ۶) و تعداد حالات مطلوب یک است (عدد ۳)

## خواص احتمال



1 غیر منفی بودن: احتمال یک رویداد همیشه یک عدد غیر منفی است

$$0 \leq P(A) \leq 1$$

2 مجموع احتمالات: مجموع احتمالات تمامی رویدادهای ممکن در یک فضای نمونه برابر با ۱ است

$$\sum_{i=1}^n P(A_i) = 1$$

3 احتمال متمم یک رویداد: احتمال رخدادن یک رویداد (متمم آن) برابر است با ۱ منهای احتمال

رخداد آن رویداد

$$P(\neg A) = 1 - P(A)$$

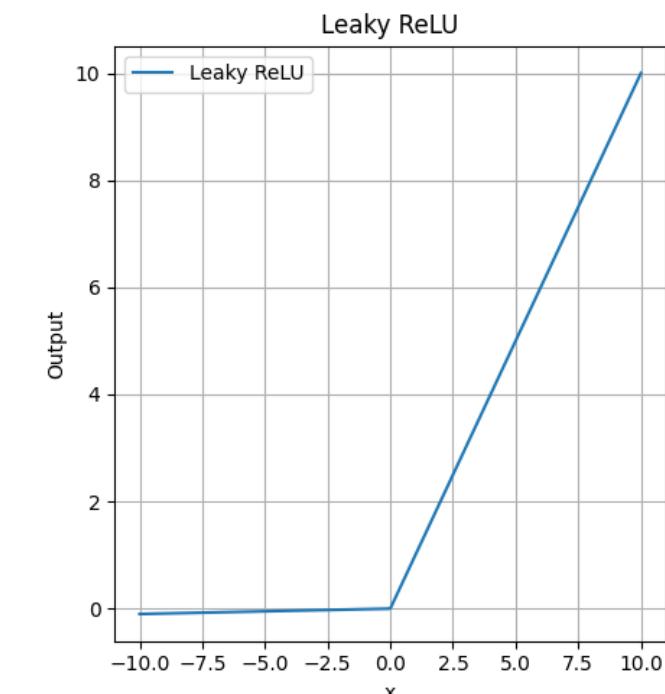
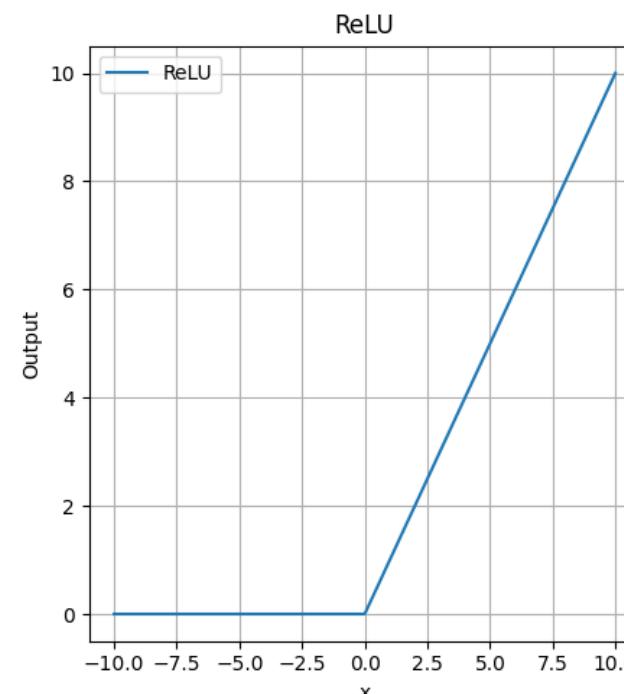
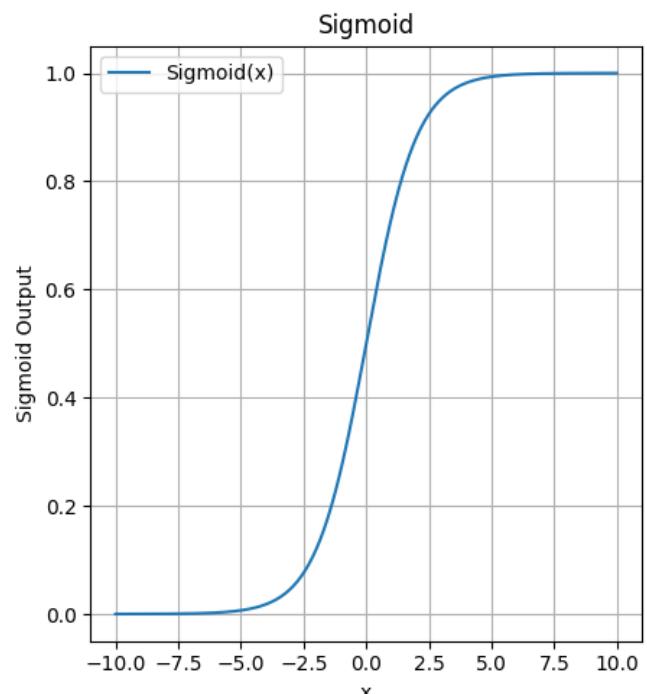
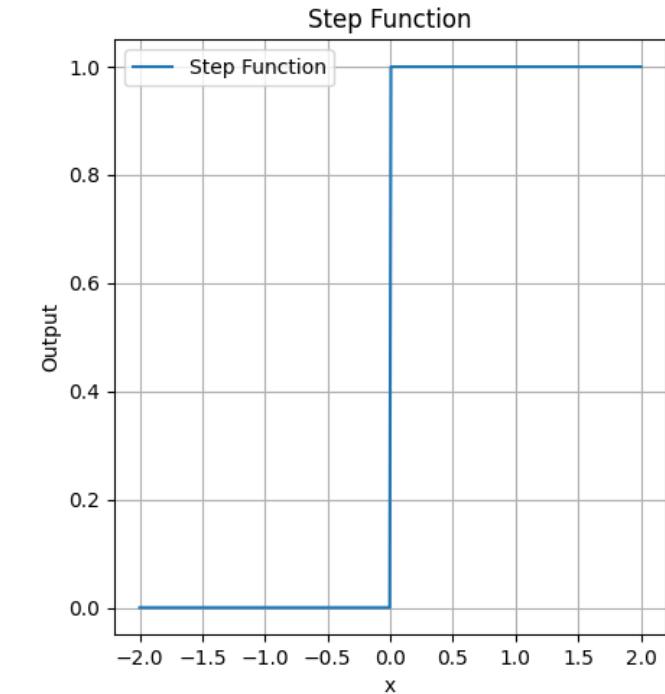
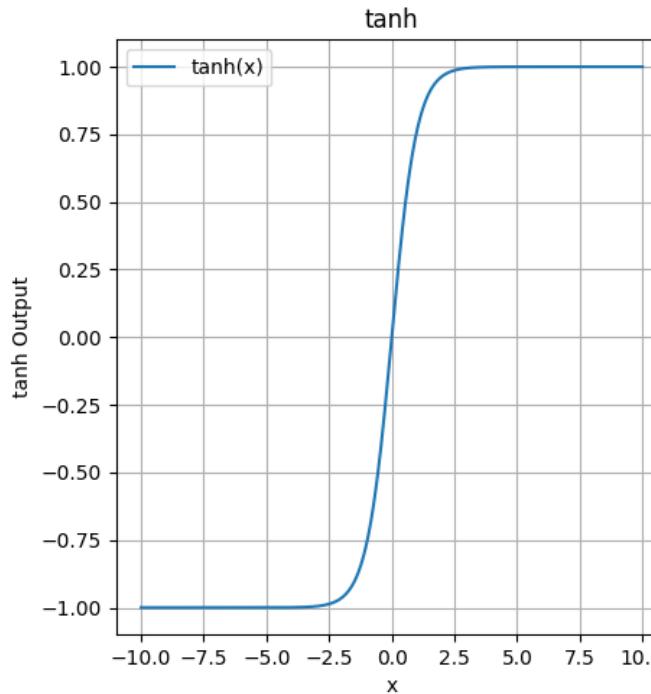
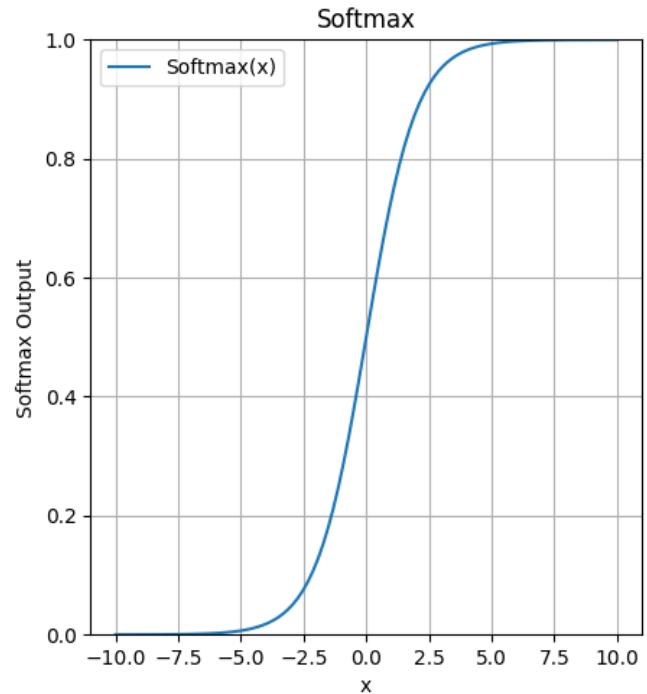
فرض کنید که می‌خواهیم احتمال روآمدن شیر یا خط در پرتاب یک سکه را محاسبه کنیم. یک سکه معمولی دو رو

دارد: یک رو شیر و یک رو خط

$$P(\text{شیر}) = \frac{1}{2} = 0.5$$

$$P(\text{خط}) = 1 - P(\text{شیر}) = 0.5$$

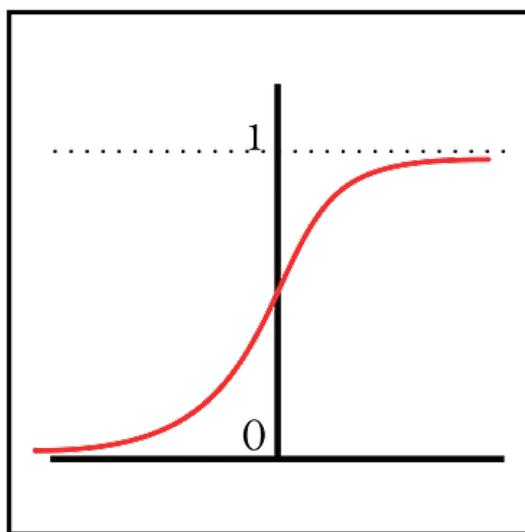
# Activation Function



**Input**

$X \in \mathbb{R}$  →

Sigmoid / SoftMax



**Output**

→  $P(Y=k | X) \in [0,1]$

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

Sigmoid

2 classes

$$\text{out} = P(Y=\text{class1}|X)$$

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^k e^{z_j}}$$

SoftMax

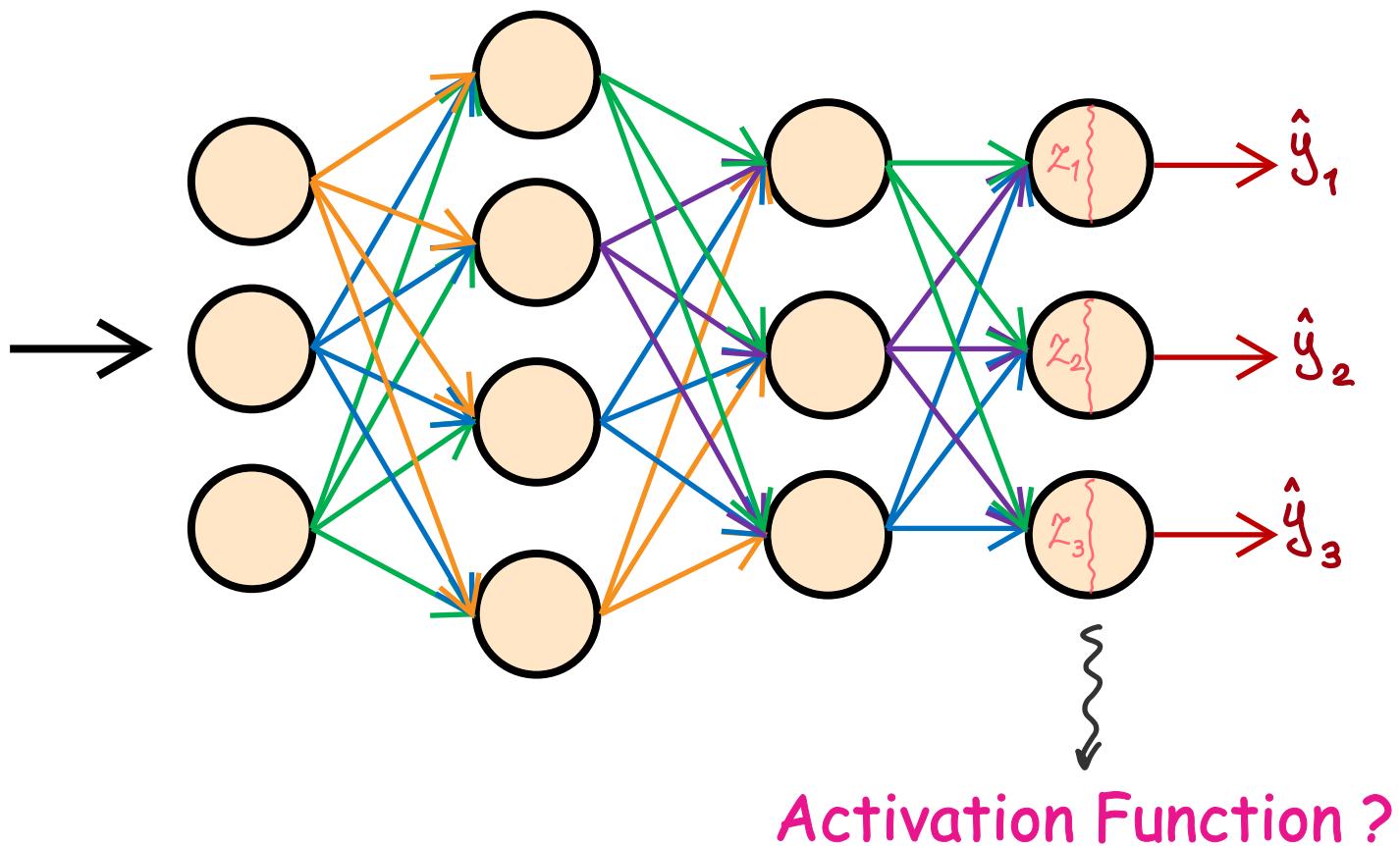
k>2 classes

$$\text{out} = \begin{bmatrix} P(Y=\text{class1}|X) \\ P(Y=\text{class2}|X) \\ P(Y=\text{class3}|X) \\ \vdots \\ P(Y=\text{classk}|X) \end{bmatrix}$$

$$\sum_{i=1}^2 P_i = 1 \quad \xrightarrow{P_1 + P_2 = 1}$$

$$\sum_{i=1}^k P_i = 1 \quad \xrightarrow{P_1 + P_2 + P_3 + \dots + P_k = 1}$$

# Multiclass Classification





در نظر بگیرید که ورودی‌های تابع سیگموید برای سه کلاس گربه، سگ و خرگوش به این صورت است

$$\left. \begin{array}{l} \text{گربه} \rightarrow z_1 = 2 \\ \text{سگ} \rightarrow z_2 = 1 \\ \text{خرگوش} \rightarrow z_3 = 0.1 \end{array} \right\} \text{Sigmoid} \quad \sigma(z) = \frac{1}{1+e^{-z}}$$

$\sum \quad \quad \quad \text{--}$

$$\sigma(z_1) = 0.88 \rightarrow \% 88$$

$$\sigma(z_2) = 0.73 \rightarrow \% 73$$

$$\sigma(z_3) = 0.52 \rightarrow \% 52$$

$$\sigma(z_1) + \sigma(z_2) + \sigma(z_3) = 2.13$$

$$\left. \begin{array}{l} \text{گربه} \rightarrow z_1 = 2 \\ \text{سگ} \rightarrow z_2 = 1 \\ \text{خرگوش} \rightarrow z_3 = 0.1 \end{array} \right\} \text{Softmax} \quad \sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^3 e^{z_j}}$$

$\sum \quad \quad \quad \text{--}$

$$\sigma(z_1) = \frac{e^2}{e^2 + e^1 + e^{0.1}} = \frac{7.39}{11.22} = 0.66 \rightarrow \% 66$$

$$\sigma(z_2) = \frac{e^1}{e^2 + e^1 + e^{0.1}} = \frac{2.72}{11.22} = 0.24 \rightarrow \% 24$$

$$\sigma(z_3) = \frac{e^{0.1}}{e^2 + e^1 + e^{0.1}} = \frac{1.11}{11.22} = 0.1 \rightarrow \% 10$$

$$\sigma(z_1) + \sigma(z_2) + \sigma(z_3) = 1$$

$$\left. \begin{array}{l} e^{z_1} = e^2 = 7.39 \\ e^{z_2} = e^1 = 2.72 \\ e^{z_3} = e^{0.1} = 1.11 \end{array} \right\} \sum_{j=1}^3 = e^{z_1} + e^{z_2} + e^{z_3} = 11.22$$

## 1 Binary Cross-Entropy Loss

$$L(y, \hat{y}) = -\frac{1}{m} \sum_{j=1}^m (y \log(\hat{y}) + (1-y) \log(1-\hat{y}))$$

✓ if  $y = 1 \rightarrow L(y, \hat{y}) = -(\log(\hat{y}))$   
 if  $y = 0 \rightarrow L(y, \hat{y}) = -(\log(1 - \hat{y}))$

## 2 Cross-Entropy Loss (Log Loss)

$$L(y, \hat{y}) = -\frac{1}{m} \sum_{j=1}^m \sum_{i=1}^C y_i \log(\hat{y}_i)$$

مناسب برای مدل‌هایی که خروجی احتمال دارند (شبکه‌های عصبی) ✓



فرض کنید یک مسئله طبقه‌بندی باینری داریم. در این مثال، یک نمونه داده با

$$P(1) = 0.8$$
$$P(0) = 0.2$$

و خروجی مدل به شرح زیر است

$$y = 1 \quad y = [0.8]$$

$$L(y, \hat{y}) = -\frac{1}{m} \sum_{j=1}^m (y \log(\hat{y}) + (1-y) \log(1-\hat{y}))$$

$$\rightarrow L(y, \hat{y}) = -(1 \cdot \log(0.8) + (1-1) \cdot \log(1-0.8)) = -\log(0.8) \approx 0.2231$$



فرض کنید یک مسئله طبقه‌بندی چندکلاسه داریم، یک نمونه داده داریم با  $y = A$

$$P(A) = 0.7$$

$$P(B) = 0.2$$

$$P(C) = 0.1$$

و خروجی مدل به شرح زیر است

$$y = A \rightsquigarrow \hat{y} = \begin{bmatrix} 1, 0, 0 \end{bmatrix} \quad \hat{y} = \begin{bmatrix} 0.7, 0.2, 0.1 \end{bmatrix}$$

$$L(y, \hat{y}) = -\frac{1}{m} \sum_{j=1}^m \sum_{i=1}^C y_j \log(\hat{y}_i)$$

$$\rightarrow L(y, \hat{y}) = - ( 1 \cdot \log(0.7) + (0) \cdot \log(0.2) + (0) \cdot \log(0.1) ) = -\log(0.7) \approx 0.3567$$

# آیا می توانیم از هر تابع فعال سازی در لایه های پنهان مدل استفاده کنیم؟

نه 😐، استفاده از توابع فعال ساز سیگموید، سافت مکس و تانژانت هایپربولیک در

لایه های پنهان شبکه های عصبی برای مدل مشکلات اساسی ایجاد می کنند

## 1 Vanishing Gradient یا ناپدید شدن گرادیان

تابع سیگموید مقدار خروجی را به محدوده (۰, ۱) محدود می کند. در طول فرآیند بکپراپگیشن، گرادیانت ها ممکن است به دلیل مشتقات کوچک سیگموید کاهش یابند، که باعث می شود آپدیت وزن ها بسیار کوچک شود و یادگیری مدل کند یا حتی متوقف شود

## 2 Non-zero-centered Output غیراشباع شدن

خروجی سیگموید همیشه مثبت است، که باعث می شود گرادیانت ها غیرمتمرکز باشند. این موضوع می تواند منجر به پایداری و همگرایی کمتر مدل شود

## 3 نرخ یادگیری پایین

با استفاده از تابع سیگموید، تغییرات کوچک در ورودی ها ممکن است تغییرات بزرگی در خروجی ایجاد نکنند، که باعث کاهش نرخ یادگیری می شود

## 4 محاسبات غیرکارآمد

نیاز به محاسبات بیشتری دارد و از لحاظ ReLU تابع سیگموید شامل محاسبات نمایی است که نسبت به توابع فعال سازی دیگر مثل محاسباتی کارآمد نیست

✓ به دلیل این مشکلات، توابع فعال ساز دیگری مانند ReLU, Leaky ReLU, ELU

به طور گسترده تری در لایه های پنهان شبکه های عصبی مدرن استفاده می شوند

تفسیر	Gradient Norm (Layer 20)	Gradient Norm (Layer 1)	Accuracy After Training	Loss After Training	Activation Function
گرادیان‌ها در لایه‌های اولیه بسیار کوچک هستند و به مرور تا لایه‌های بالاتر افزایش می‌یابند. مشکل ونیشینگ گرادیانت باعث کاهش دقت نهایی می‌شود. <span style="color: red;">X</span>	1.318084	0.000908	0.622807	0.662834	Sigmoid
گرادیان‌ها به خوبی توزیع شده و مقادیر مناسبی دارند. ReLU مشکل ونیشینگ گرادیانت را کاهش داده و دقت نهایی بسیار بالاست. <span style="color: green;">✓</span>	0.407361	3.004350	0.964912	0.177435	ReLU
گرادیان‌ها به خوبی توزیع شده و مقادیر مناسبی دارند. Leaky ReLU نیز مشکل ونیشینگ گرادیانت را کاهش داده و دقت نهایی بسیار بالاست. <span style="color: green;">✓</span>	0.456306	2.161442	0.973684	0.176373	Leaky ReLU

گرادیان‌ها در لایه‌های اولیه بسیار کوچک هستند و به مرور تا لایه‌های بالاتر افزایش می‌یابند. این نشان‌دهنده مشکل ونیشینگ گرادیانت است، که در آن گرادیان‌ها در لایه‌های ابتدایی به شدت کاهش می‌یابند و یادگیری مدل در این لایه‌ها دشوار می‌شود X

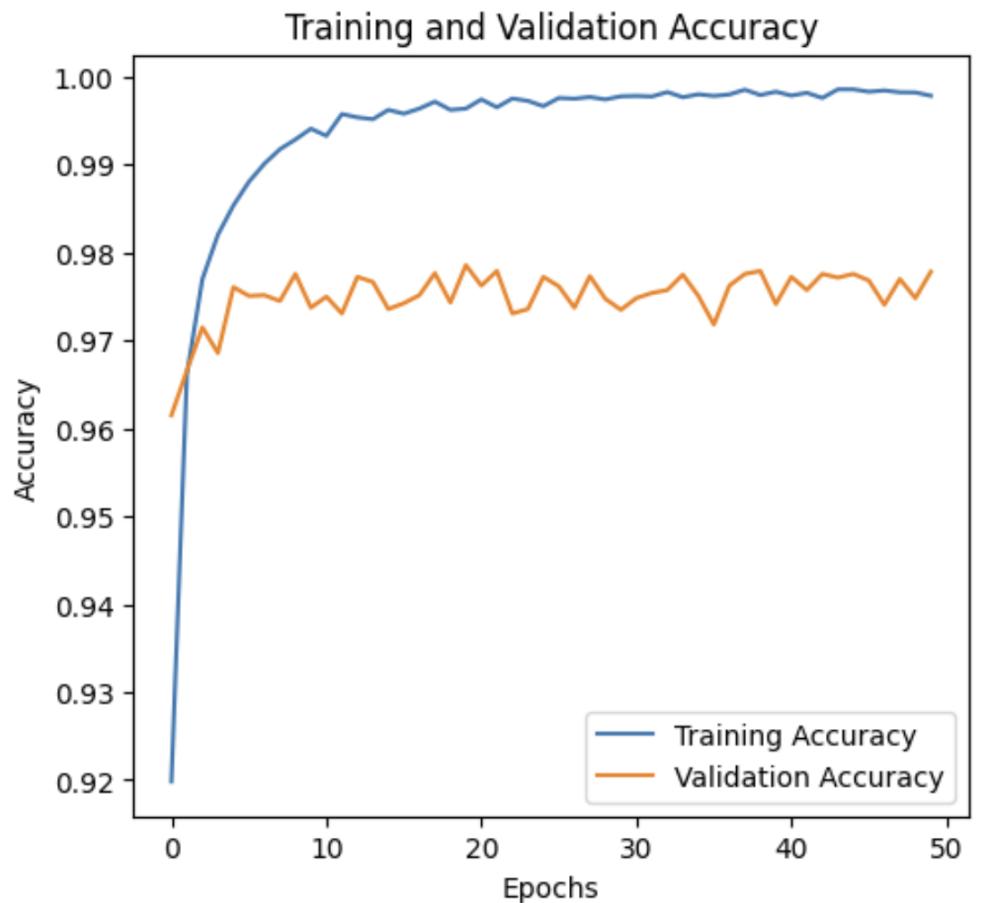
گرادیان‌ها در تمامی لایه‌ها به خوبی توزیع شده‌اند و مقادیر مناسبی دارند. این نشان‌دهنده این است که ReLU مشکل ونیشینگ گرادیانت را کاهش می‌دهد و یادگیری در تمامی لایه‌ها به خوبی صورت می‌گیرد ✓

اصطلاح	تعریف
بایاس (Bias)	میزان خطای مدل به دلیل ساده‌سازی بیش از حد. مدل‌های با بایاس بالا نمی‌توانند الگوهای پیچیده داده‌ها را به خوبی یاد بگیرند.
واریانس (Variance)	میزان حساسیت مدل به تغییرات کوچک در داده‌های آموزشی. مدل‌های با واریانس بالا به جزئیات و نویزهای داده‌ها بسیار حساس هستند.
تعمیم‌دهی (Generalization)	توانایی مدل در عملکرد خوب بر روی داده‌های جدید و نادیده. مدل‌های با تعیین‌دهی بالا الگوهای یادگرفته شده را به داده‌های جدید به درستی اعمال می‌کنند.

	<b>Underfitting</b>	<b>Just right</b>	<b>Overfitting</b>
<b>Symptoms</b>	<ul style="list-style-type: none"> <li>• High training error</li> <li>• Training error close to test error</li> <li>• High bias</li> </ul>	<ul style="list-style-type: none"> <li>• Training error slightly lower than test error</li> </ul>	<ul style="list-style-type: none"> <li>• Very low training error</li> <li>• Training error much lower than test error</li> <li>• High variance</li> </ul>
<b>Regression illustration</b>			
<b>Classification illustration</b>			
<b>Deep learning illustration</b>			
<b>Possible remedies</b>	<ul style="list-style-type: none"> <li>• Complexify model</li> <li>• Add more features</li> <li>• Train longer</li> </ul>		<ul style="list-style-type: none"> <li>• Perform regularization</li> <li>• Get more data</li> </ul>

تعمیم‌دهی (Generalization)	واریانس (Variance)	بایاس (Bias)	تعریف	اصطلاح
ضعیف	پایین	بالا	مدل به اندازه کافی پیچیدگی ندارد تا الگوهای موجود در داده‌ها را به درستی یاد بگیرد. عملکرد ضعیف در آموزش و تست.	آندرفیت
ضعیف	بالا	پایین	مدل به قدری پیچیده است که نویزها و جزئیات غیرضروری را نیز یاد بگیرد. عملکرد عالی در آموزش و ضعیف در تست و جدید.	اورفیت
بالا	متعادل	متعادل	مدل به اندازه کافی پیچیده است تا الگوهای موجود در داده‌ها را به درستی یاد بگیرد بدون اینکه به نویزها توجه کند. عملکرد خوب در آموزش و تست و می‌تواند به خوبی تعمیم دهد.	گودفیت

# تحليل نتائج



تعیین‌دهی (Generalization)	واریانس (Variance)	بایاس (Bias)	Validation Accuracy	Training Accuracy	Validation Loss	Training Loss	معیار
-	-	-	افزایش اولیه، سپس ثابت و نوسانی	افزایش سریع و پیوسته	کاهش اولیه، سپس افزایش	کاهش سریع و پیوسته	رفتار در دوره‌های اولیه
-	-	-	ثابت با نوسان کمی	نزدیک به %100	افزایش قابل توجه پس از دوره‌های اولیه	بسیار کم (نزدیک به صفر)	رفتار نهایی
ضعیف	واریانس بالا (مدل حساس به نویز)	بایاس پایین (مدل پیچیده)	مشکل در تعیین‌دهی	یادگیری دقیق از داده‌های آموزشی	اورفیت (Overfitting)	یادگیری خوب از داده‌های آموزشی	نشان‌دهنده
-	-	بایاس پایین	-	عالی	-	عالی	عملکرد در آموزش
ضعیف	واریانس بالا	-	متوسط (ثابت و نوسانی)	-	ضعیف (افزایش Loss)	-	عملکرد در تست
بهبود تعیین‌دهی با داده‌های بیشتر Cross- Validation	کاهش واریانس با Regularization	تعادل بایاس- واریانس	استفاده از Cross- Validation	-	استفاده از Regularization، افزایش داده‌های آموزشی، استفاده از Dropout	-	پیشنهاد بهبود