AIL721: Deep Learning

Instructor: James Arambam

ScAI | Yardi School of Artificial Intelligence
Indian Institute of Technology Delhi

Class Announcements

- Project Team Size: **1~3**

- Please form the team by **20th Jan** or will be assigned **randomly.**

- **New students:** Please email me for the **piazza link** and **access code**.

Class Announcements

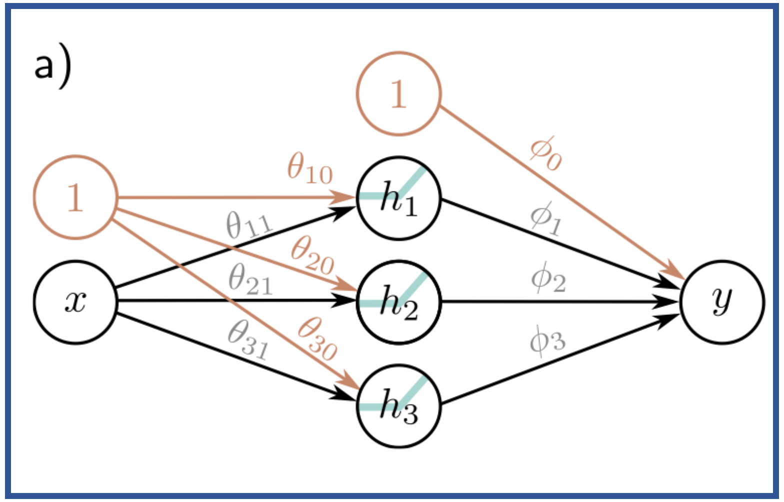
- Google Cloud Compute Credit – **Approved!**
  - 50 USD per student.
- Class on Thursday
  - **Tuesday's Schedule**
  - **5-6 PM**

Class Announcements

- IIT Delhi HPC Compute Credit – **Approved!**
  - 800 INR per student.
  - More details over email.
  - First time users: Start setting up HPC or read about it.

Deep Neural Network

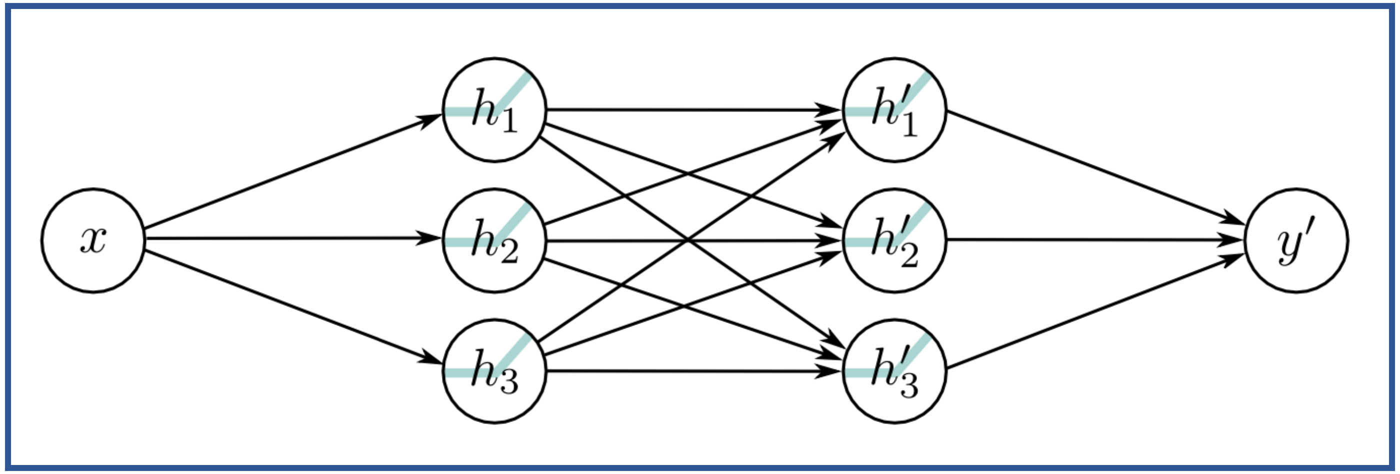
- Shallow Network:



How many parameters?

3D+1

- Deep Network:



How many parameters?

3D+(K-1)D^2+(K-1)D+1

Story So Far

Story So Far

- Brief history of neural networks.
- Basic mathematical model of neural network.
- Shallow neural network.
  - Neural network equation in *normal* form and matrix form?
  - Importance of activation functions – ReLU.
  - Visualization of a neural network.
- Deep neural network.
  - Why deep?

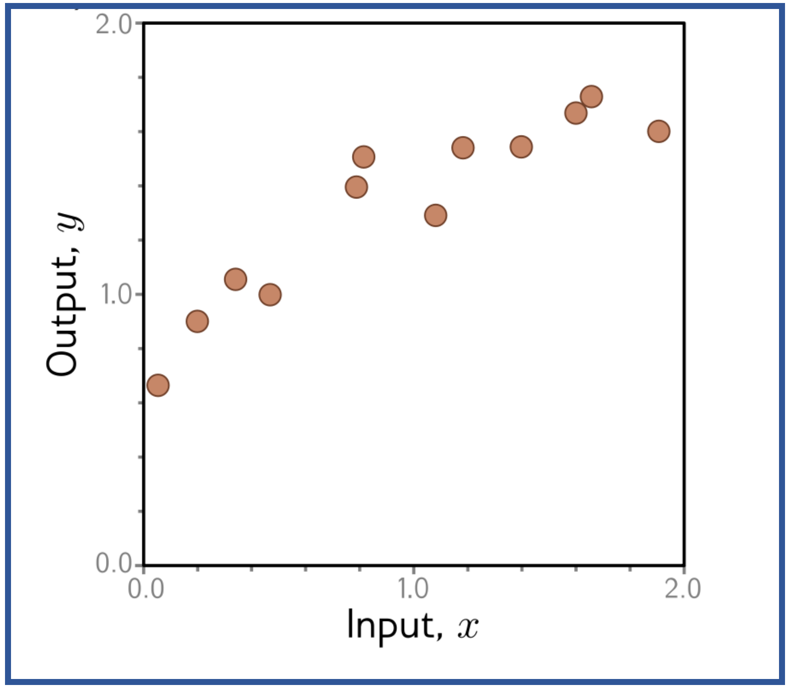
ScAI

What’s the next step?

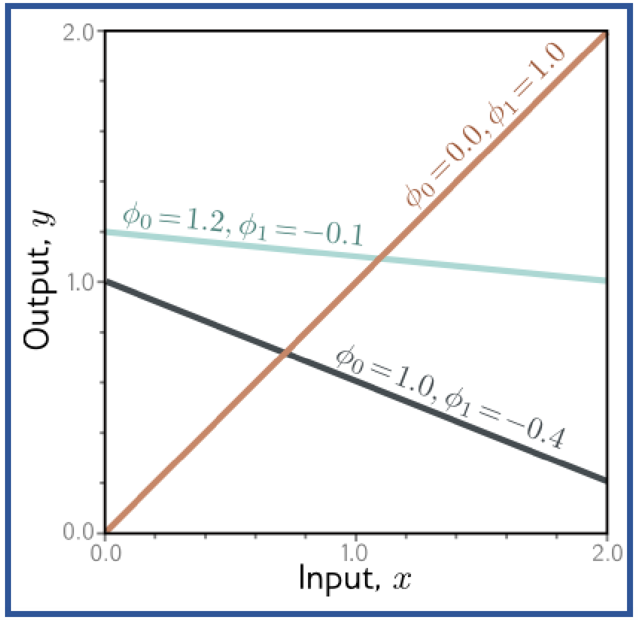
Optimization

Recap

❑ Regression



$$f(x;\boldsymbol{\phi}) = \phi_0 + \phi_1 \cdot x$$



How to find $\boldsymbol{\phi}$ that **best fits** the **given data**?

Optimization

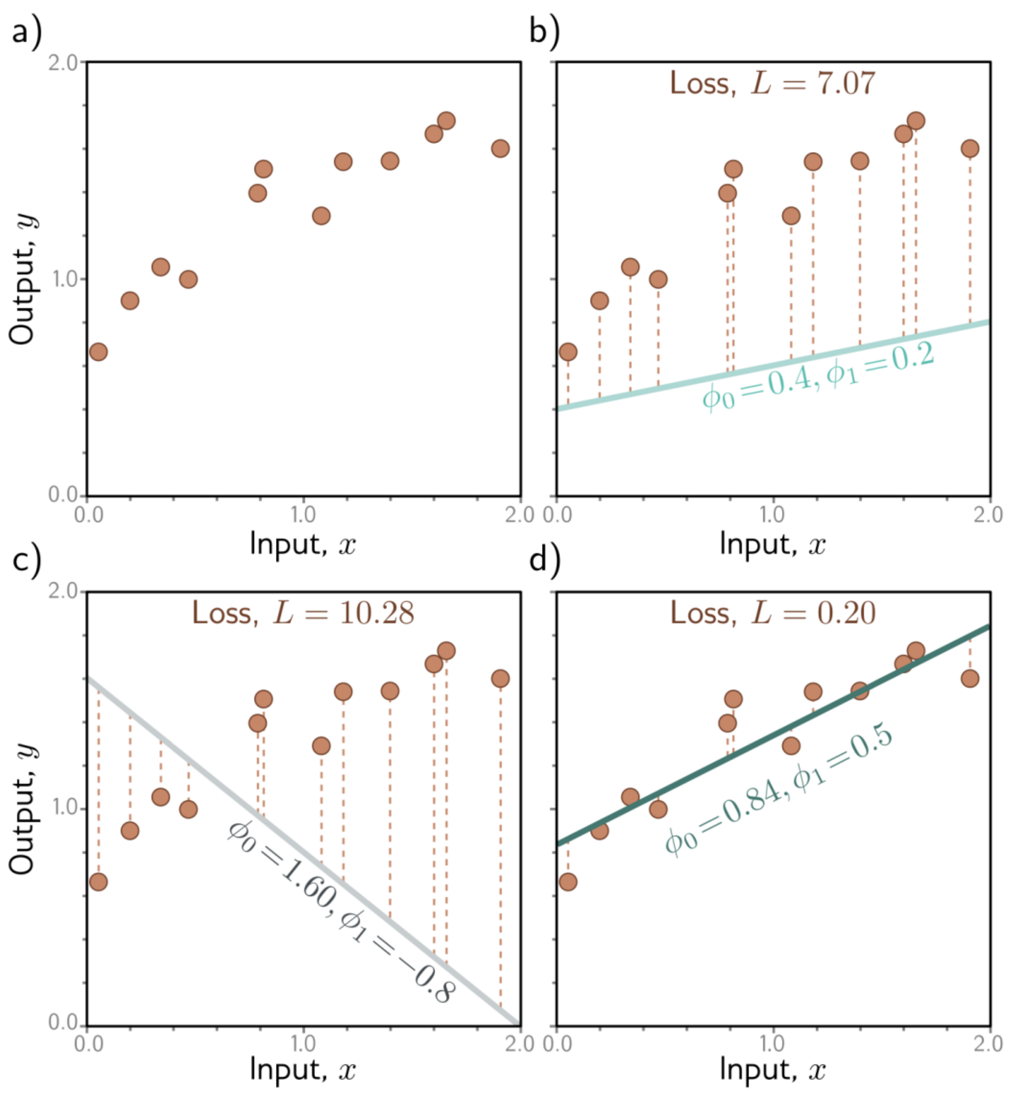
$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\Big[L\left[\boldsymbol{\phi}\right]\Big]$$

[Image from Prince 2023]

Loss Function

❑ Least Square Error Loss:

$$
\begin{aligned}
L[\phi] &= \sum_{i=1}^{I} \left(\mathrm{f}[x_i, \phi] - y_i\right)^2 \\
&= \sum_{i=1}^{I} \left(\phi_0 + \phi_1 x_i - y_i\right)^2
\end{aligned}
$$



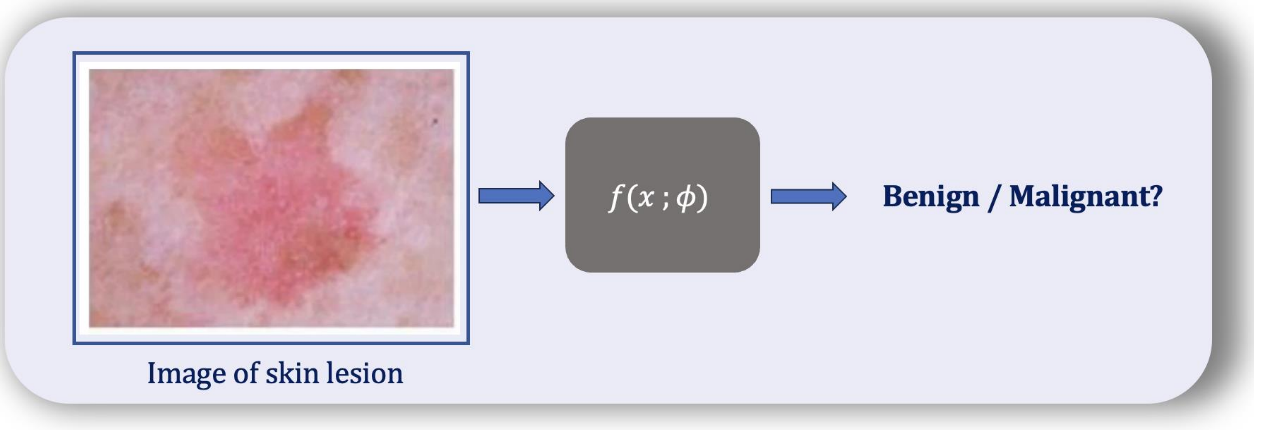
[Image from Prince 2023]

Loss Function

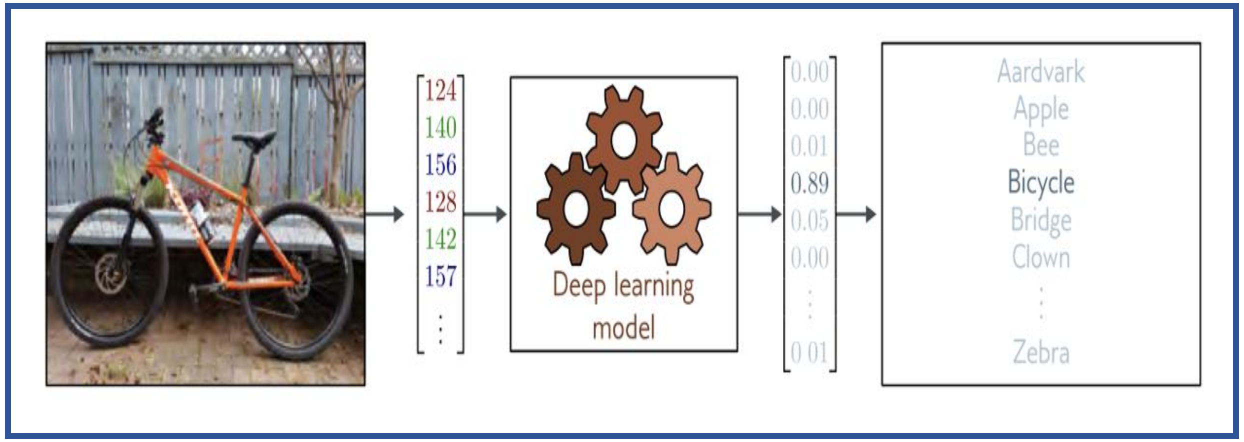
❑ Least Square Error Loss:

$$L[\phi] = \sum_{i=1}^{I} \left(\mathrm{f}[x_i, \phi] - y_i\right)^2$$

$$= \sum_{i=1}^{I} \left(\phi_0 + \phi_1 x_i - y_i\right)^2$$



Binary Classification?



Multiclass Classification?

[Image from Prince 2023]

Loss Function

❑ Least Square Error Loss:

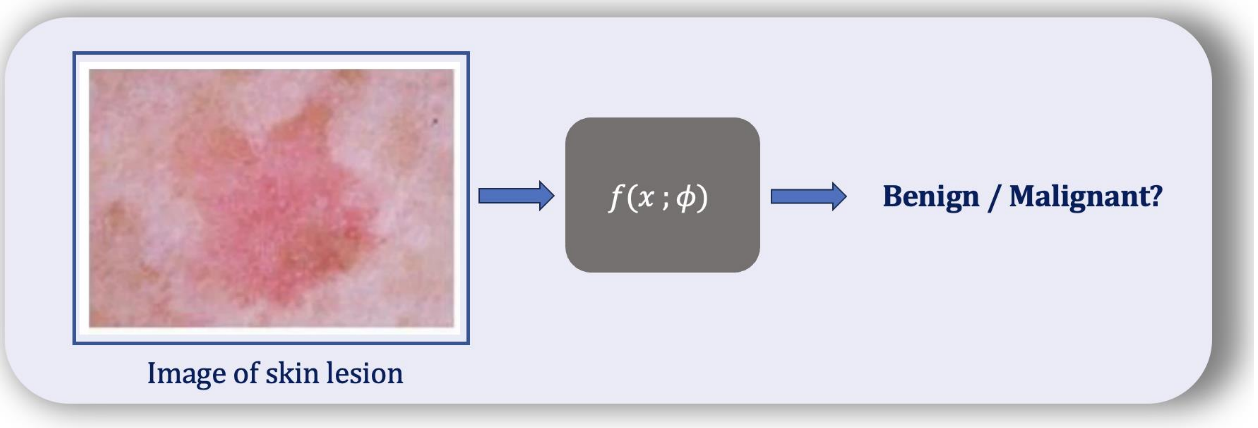
$$L[\phi] = \sum_{i=1}^{I} \left(\mathrm{f}[x_i, \phi] - y_i\right)^2$$

$$= \sum_{i=1}^{I} \left(\phi_0 + \phi_1 x_i - y_i\right)^2$$

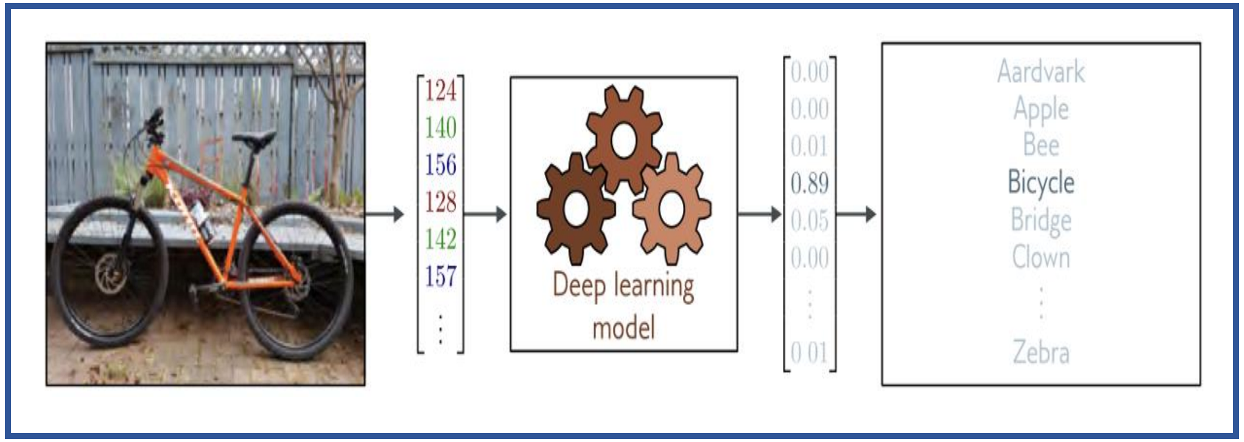
$$y = f(x;\phi)$$

$$Pr(y|x)$$

Conditional Probability Model



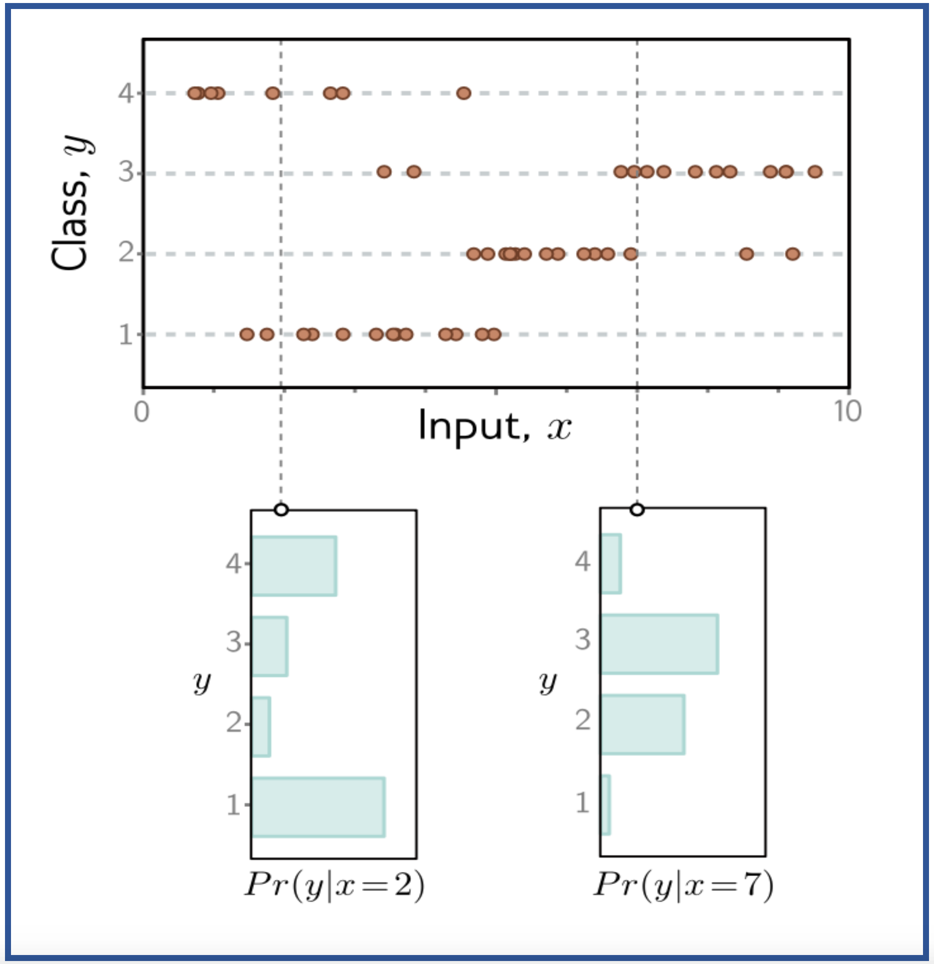
Binary Classification?

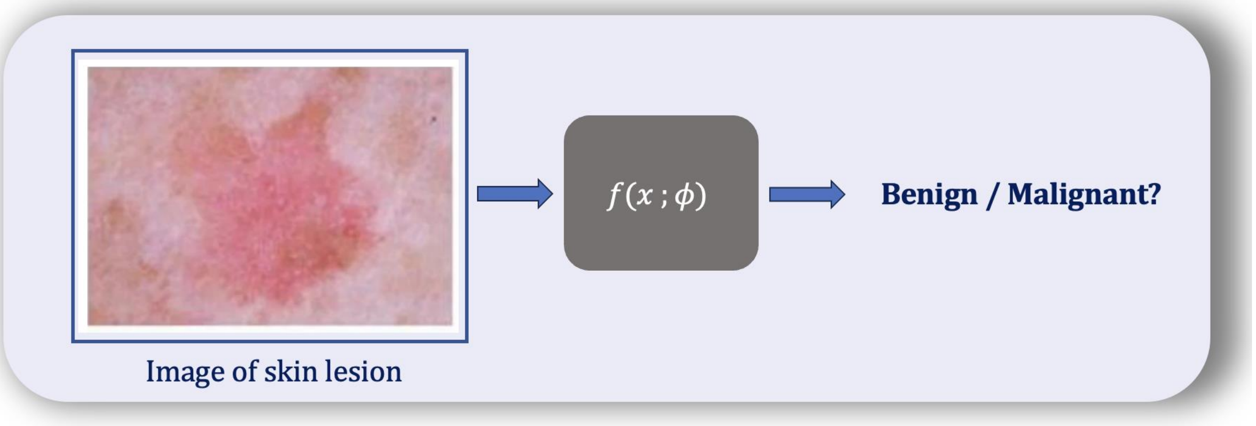


Multiclass Classification?

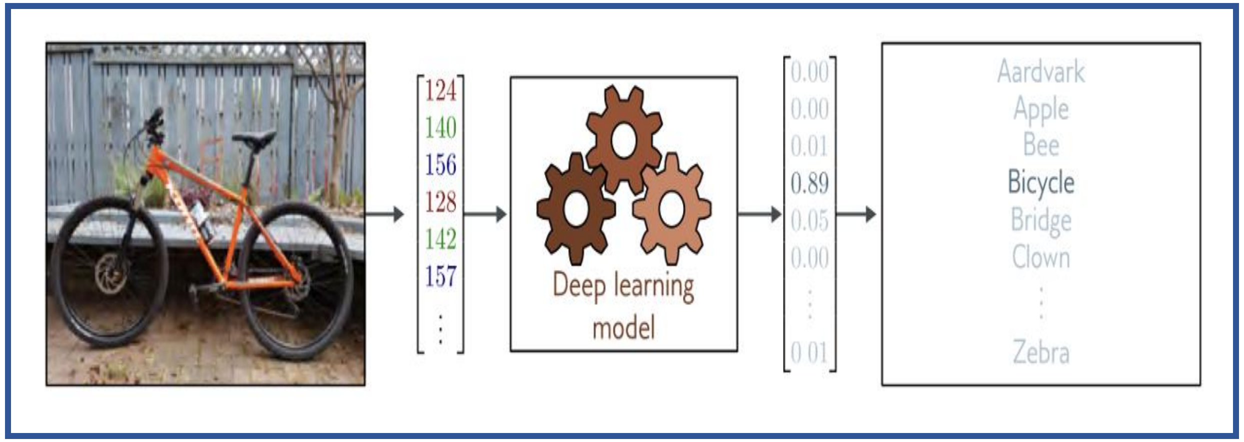
[Image from Prince 2023]

Loss Function





Binary Classification?

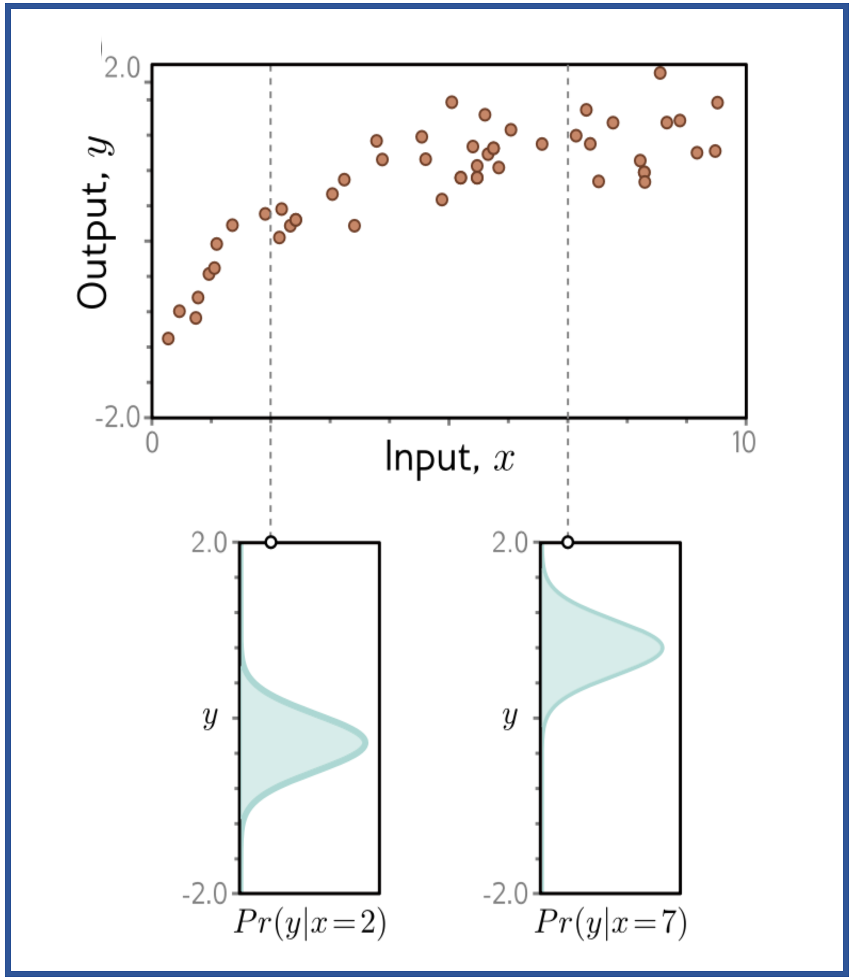


Multiclass Classification?

Could we still do regression with this model?

Loss Function

❑ Regression with probabilistic model



[Image from Prince 2023]

Loss Function

❑ Maximum Likelihood

$$\boxed{Pr(y|x)}$$

Maximize the probability of $Pr(y_i\,|x_i)$ from training data $(y_i\,, x_i)$.

[Image from Prince 2023]

Loss Function

- **Maximum Likelihood**

$$Pr(y|\theta)$$

$$\theta = f(x; \phi)$$

Parameterized probability distribution

Training data: $(x_1, y_1), (x_2, y_2), \ldots, (x_I, y_I)$

$$Pr(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_I | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_I)$$

$$\begin{aligned}
\hat{\boldsymbol{\phi}} &= \underset{\boldsymbol{\phi}}{\operatorname{argmax}} \left[\prod_{i=1}^{I} Pr(\mathbf{y}_i | \mathbf{x}_i) \right] \\
&= \underset{\boldsymbol{\phi}}{\operatorname{argmax}} \left[\prod_{i=1}^{I} Pr(\mathbf{y}_i | \boldsymbol{\theta}_i) \right] \\
&= \underset{\boldsymbol{\phi}}{\operatorname{argmax}} \left[\prod_{i=1}^{I} Pr(\mathbf{y}_i | \mathbf{f}[\mathbf{x}_i, \boldsymbol{\phi}]) \right]
\end{aligned}$$

[Image from Prince 2023]

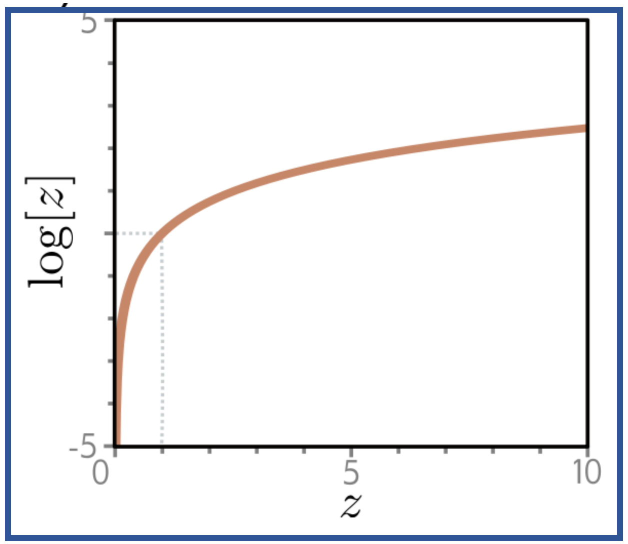
Loss Function

❑ Maximum Likelihood

$$
\begin{aligned}
\hat{\boldsymbol{\phi}} &= \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\left[\prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{f}[\mathbf{x}_i, \boldsymbol{\phi}])\right] \\
&= \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\left[\log\left[\prod_{i=1}^{I} Pr(\mathbf{y}_i|\mathbf{f}[\mathbf{x}_i, \boldsymbol{\phi}])\right]\right] \\
&= \underset{\boldsymbol{\phi}}{\operatorname{argmax}}\left[\sum_{i=1}^{I} \log\left[Pr(\mathbf{y}_i|\mathbf{f}[\mathbf{x}_i, \boldsymbol{\phi}])\right]\right]
\end{aligned}
$$

$$
\begin{aligned}
\hat{\boldsymbol{\phi}} &= \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[-\sum_{i=1}^{I} \log\left[Pr(\mathbf{y}_i|\mathbf{f}[\mathbf{x}_i, \boldsymbol{\phi}])\right]\right] \\
&= \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[L[\boldsymbol{\phi}]\right]
\end{aligned}
$$

Negative Log Likelihood Loss

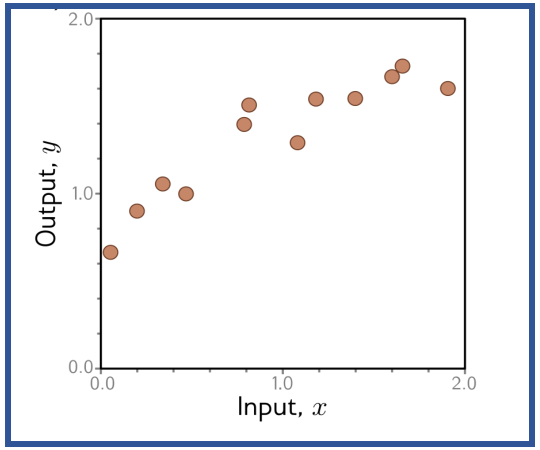


[Image from Prince 2023]

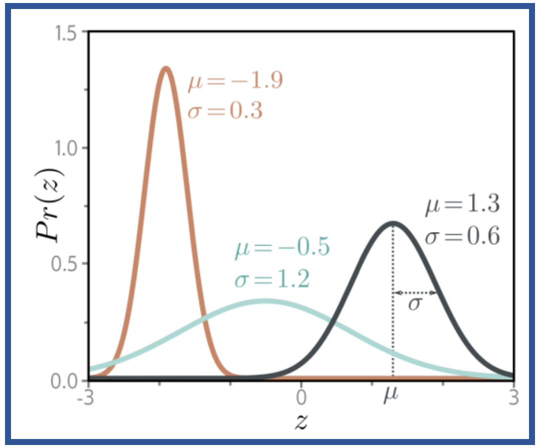
Loss Function

❑ Univariate Regression with Probabilistic Model

$$Pr(y|f(x;\phi))$$?



$$Pr(y|\mathrm{f}[\mathbf{x},\boldsymbol{\phi}],\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y-\mathrm{f}[\mathbf{x},\boldsymbol{\phi}])^2}{2\sigma^2}\right]$$



$$\begin{aligned}L[\boldsymbol{\phi}] &= -\sum_{i=1}^{I}\log\left[Pr(y_i|\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}],\sigma^2)\right]\\ &= -\sum_{i=1}^{I}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i-\mathrm{f}[\mathbf{x}_i,\boldsymbol{\phi}])^2}{2\sigma^2}\right]\right]\end{aligned}$$

$$Pr(y|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

Normal Distribution

[Image from Prince 2023]

Loss Function

❑ Univariate Regression with Probabilistic Model

$$
\begin{aligned}
\hat{\phi} &= \underset{\phi}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i-\mathrm{f}[\mathbf{x}_i,\phi])^2}{2\sigma^2}\right]\right]\right]\\
&= \underset{\phi}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}\left(\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]-\frac{(y_i-\mathrm{f}[\mathbf{x}_i,\phi])^2}{2\sigma^2}\right)\right]\\
&= \underset{\phi}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}-\frac{(y_i-\mathrm{f}[\mathbf{x}_i,\phi])^2}{2\sigma^2}\right]\\
&= \underset{\phi}{\operatorname{argmin}}\left[\sum_{i=1}^{I}(y_i-\mathrm{f}[\mathbf{x}_i,\phi])^2\right],
\end{aligned}
$$

Least Square Error Loss

$$
\hat{\phi},\hat{\sigma}^2 = \underset{\phi,\sigma^2}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i-\mathrm{f}[\mathbf{x}_i,\phi])^2}{2\sigma^2}\right]\right]\right]
$$

[Image from Prince 2023]

Loss Function

❑ Univariate Regression with Probabilistic Model

$$
\begin{aligned}
\hat{\phi} &= \underset{\phi}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i-\mathrm{f}[\mathbf{x}_i,\phi])^2}{2\sigma^2}\right]\right]\right] \\
&= \underset{\phi}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}\left(\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]-\frac{(y_i-\mathrm{f}[\mathbf{x}_i,\phi])^2}{2\sigma^2}\right)\right] \\
&= \underset{\phi}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}-\frac{(y_i-\mathrm{f}[\mathbf{x}_i,\phi])^2}{2\sigma^2}\right] \\
&= \underset{\phi}{\operatorname{argmin}}\left[\sum_{i=1}^{I}(y_i-\mathrm{f}[\mathbf{x}_i,\phi])^2\right],
\end{aligned}
$$

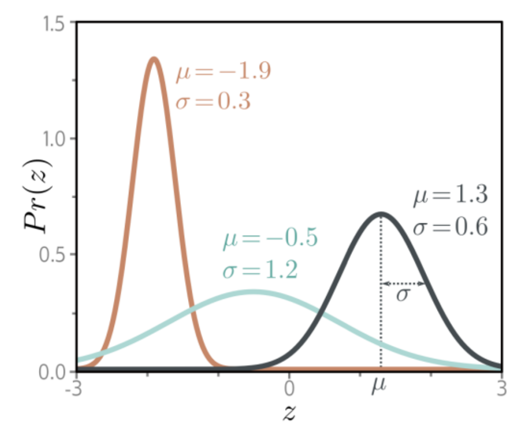
Least Square Error Loss

$$
\hat{\phi},\hat{\sigma}^2 = \underset{\phi,\sigma^2}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i-\mathrm{f}[\mathbf{x}_i,\phi])^2}{2\sigma^2}\right]\right]\right]
$$

❑ Inference

$$
\hat{y} = \underset{y}{\operatorname{argmax}}\ P(\cdot|f(x_i,\phi),\sigma)
$$

$$
\hat{y} = f(x_i,\phi)
$$



[Image from Prince 2023]

Loss Function

❑ Binary Classification

$$Pr(y|f(x;\phi))$$?

$$\lambda = f(x;\phi)$$?

$$\lambda = \text{sig}[f(x;\phi)]$$

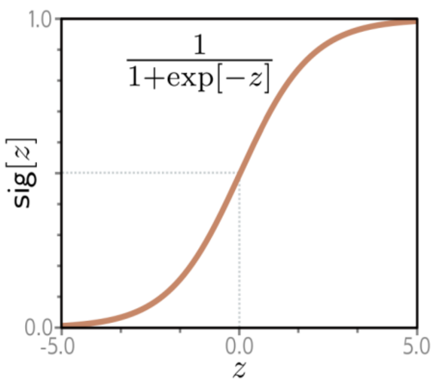
- **Likelihood Function**

$$Pr(y|\mathbf{x}) = (1 - \text{sig}[\text{f}[\mathbf{x}, \phi]])^{1-y} \cdot \text{sig}[\text{f}[\mathbf{x}, \phi]]^y.$$

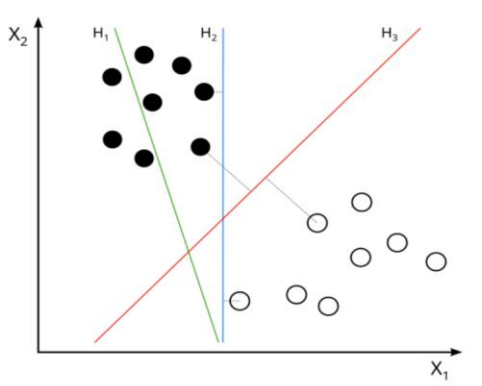
- **Negative Likelihood Loss**

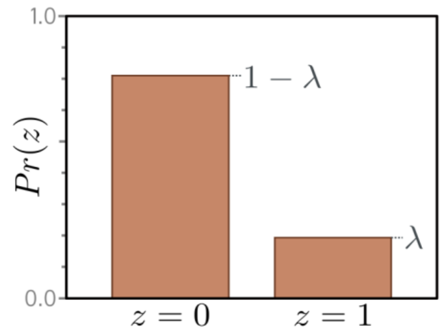
$$L[\phi] = \sum_{i=1}^{I} -(1-y_i)\log\Big[1 - \text{sig}[\text{f}[\mathbf{x}_i, \phi]]\Big] - y_i \log\Big[\text{sig}[\text{f}[\mathbf{x}_i, \phi]]\Big]$$

Sigmoid Function



$$\text{sig}[z] = \frac{1}{1+\exp[-z]}$$





$$Pr(y|\lambda) = \begin{cases} 1-\lambda & y=0 \\ \lambda & y=1 \end{cases}$$

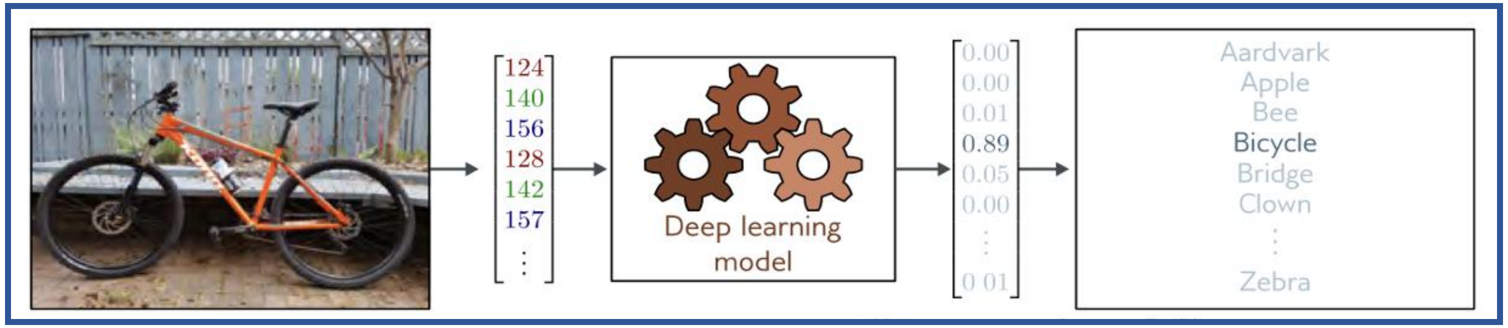
$$Pr(y|\lambda) = (1-\lambda)^{1-y} \cdot \lambda^y$$

Bernoulli's Distribution

[Image from Prince 2023]

Loss Function

❑ Multiclass Classification



Homework Exercise

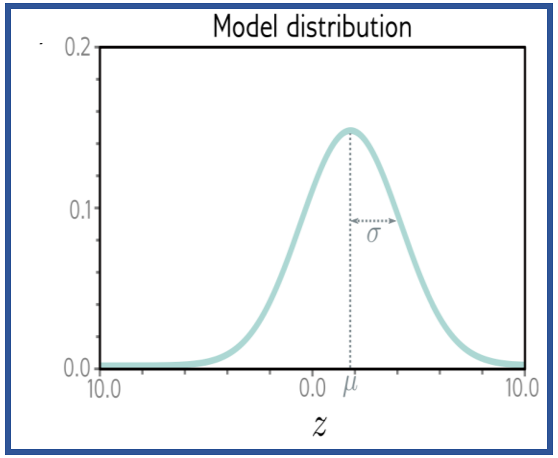
[Image from Prince 2023]

Loss Function

$q(y)$ ← **Training data:** $(x_1, y_1), (x_2, y_2), \ldots, (x_I, y_I)$

Empirical Data Distribution

$p(y;\theta)$



Where is the true distribution?

Loss Function

$q(y) \leftarrow$ **Training data:** $(x_1, y_1), (x_2, y_2), \ldots, (x_I, y_I)$

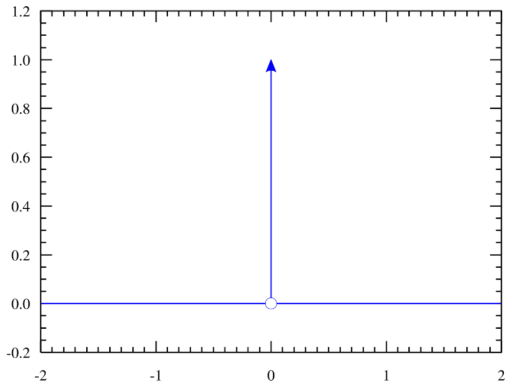
Empirical Data Distribution

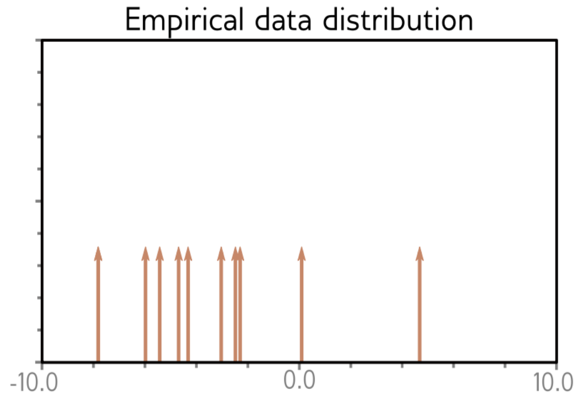
- **Dirac Delta Function**

$$\delta(x) = \begin{cases} 0, & x \neq 0 \\ \infty, & x = 0 \end{cases}$$

such that

$$\int_{-\infty}^{\infty} \delta(x)dx = 1.$$

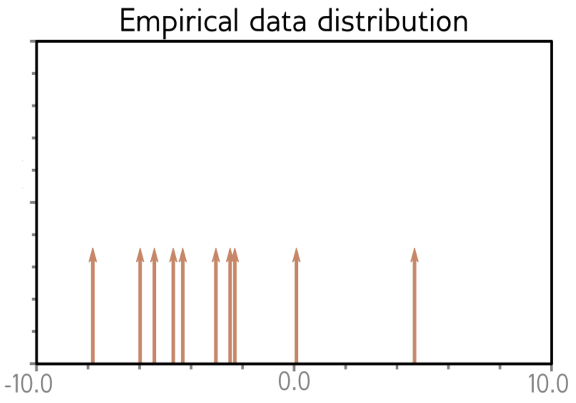




$$q(y) = \frac{1}{I}\sum_{i=1}^{I} \delta[y - y_i]$$

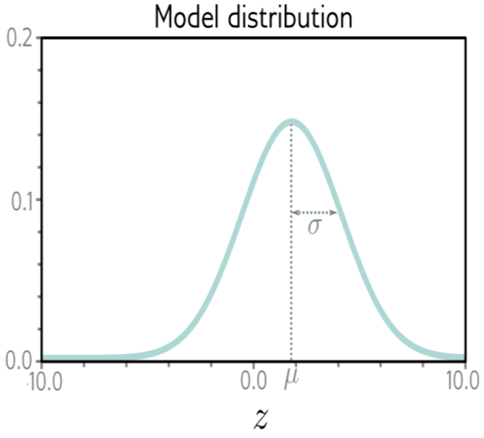
Loss Function

Empirical Data Distribution



$$q(y) = \frac{1}{I}\sum_{i=1}^{I}\delta[y - y_i]$$

Next Step?



$p(y;\theta)$

Minimize the distance between $p(y;\theta)$ and $q(y)$

$$D_{KL}\left[q||p\right] = \int_{-\infty}^{\infty} q(z)\log\left[q(z)\right]dz - \int_{-\infty}^{\infty} q(z)\log\left[p(z)\right]dz.$$

Loss Function

$$D_{KL}[q||p] = \int_{-\infty}^{\infty} q(z)\log\big[q(z)\big]dz - \int_{-\infty}^{\infty} q(z)\log\big[p(z)\big]dz.$$

$$q(y) = \frac{1}{I}\sum_{i=1}^{I}\delta[y - y_i]$$

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\left[\int_{-\infty}^{\infty} q(y)\log\big[q(y)\big]dy - \int_{-\infty}^{\infty} q(y)\log\big[Pr(y|\boldsymbol{\theta})\big]dy\right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\left[-\int_{-\infty}^{\infty} q(y)\log\big[Pr(y|\boldsymbol{\theta})\big]dy\right] \\
&= \underset{\theta}{\operatorname{argmin}}\left[-\int_{-\infty}^{\infty}\left(\frac{1}{I}\sum_{i=1}^{I}\delta[y-y_i]\right)\log\big[Pr(y|\boldsymbol{\theta})\big]dy\right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\left[-\frac{1}{I}\sum_{i=1}^{I}\log\big[Pr(y_i|\boldsymbol{\theta})\big]\right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}\log\big[Pr(y_i|\boldsymbol{\theta})\big]\right]
\end{aligned}$$

Cross-Entropy

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[-\sum_{i=1}^{I}\log\big[Pr(y_i|\mathbf{f}[\mathbf{x}_i,\boldsymbol{\phi}])\big]\right]$$

?

Negative Log Likelihood Loss

$\Updownarrow$

Cross Entropy Loss

Loss Function

Negative Likelihood Loss ⟷ Maximizing the likelihood of the observed data.

Cross Entropy Loss ⟷ Minimizing the distance between model distribution and empirical distribution.

- **Negative Likelihood Loss**

$$L[\phi] = \sum_{i=1}^{I} -(1-y_i)\log\Big[1-\text{sig}[\text{f}[\mathbf{x}_i,\phi]]\Big] - y_i \log\Big[\text{sig}[\text{f}[\mathbf{x}_i,\phi]]\Big]$$

Binary Cross-Entropy Loss