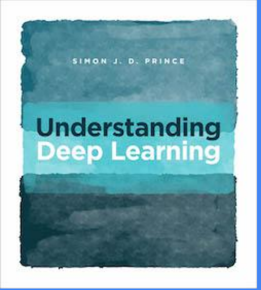
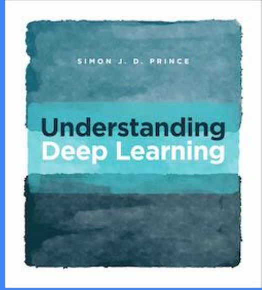
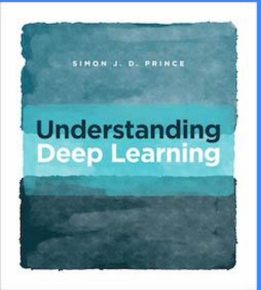
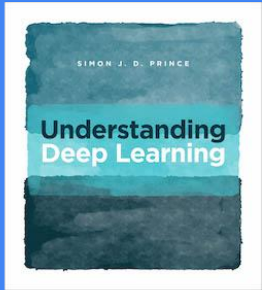




Loss functions

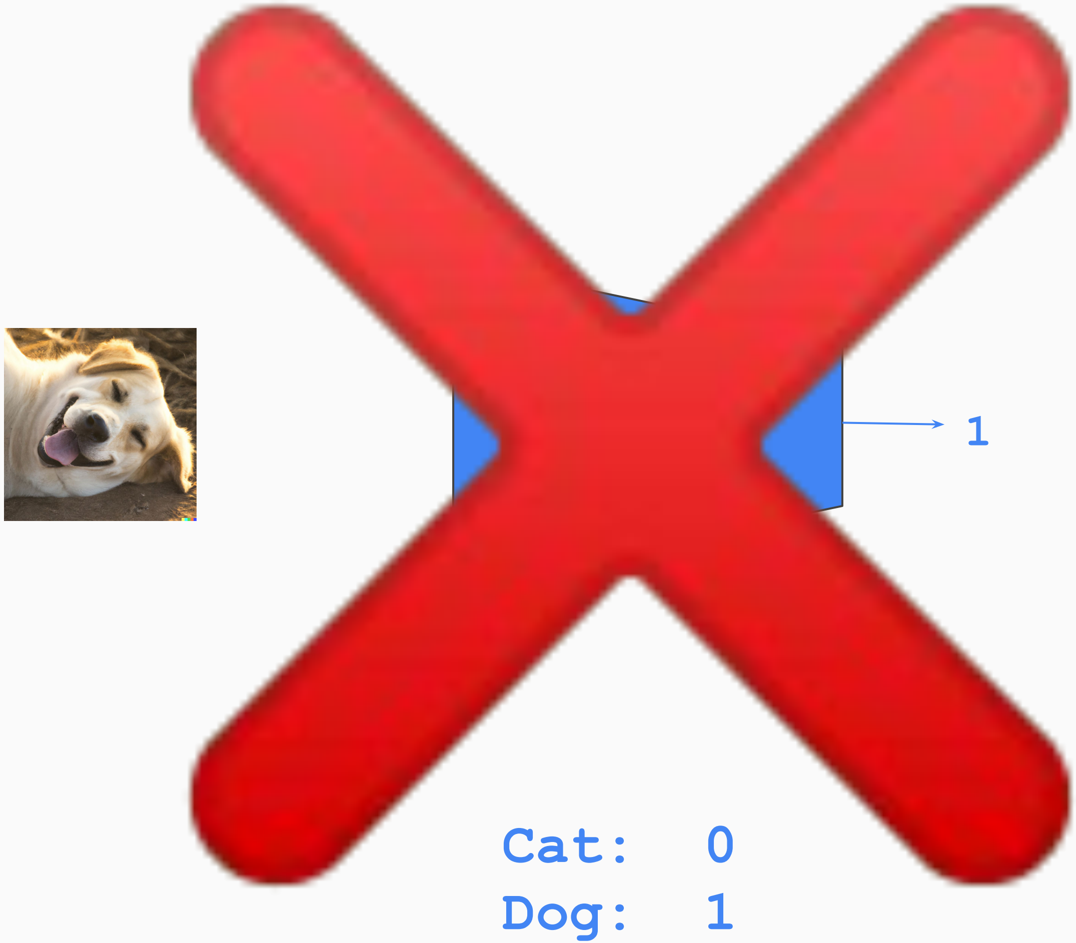




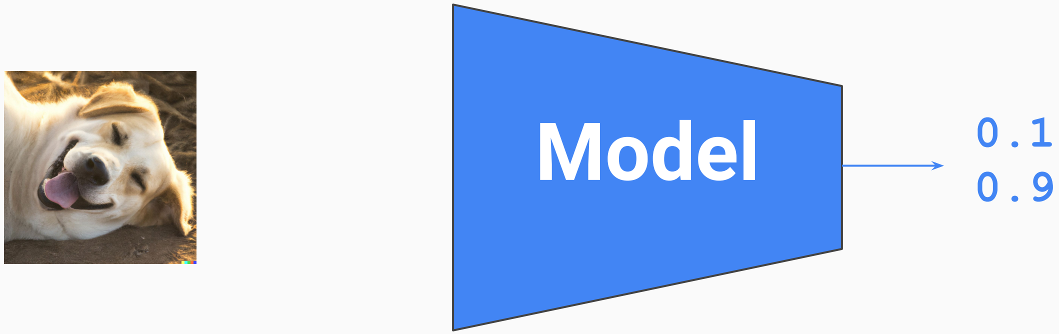
Loss or Cost function

Purpose:

To measure the difference between the predicted values and the actual values (or labels). The larger the difference, the larger the loss.


1
Cat: 0
Dog: 1

Conditional Probability Distribution



P(0 | ) = 0.1

Cat: 0
Dog: 1

P(1 | ) = 0.9

Terminology

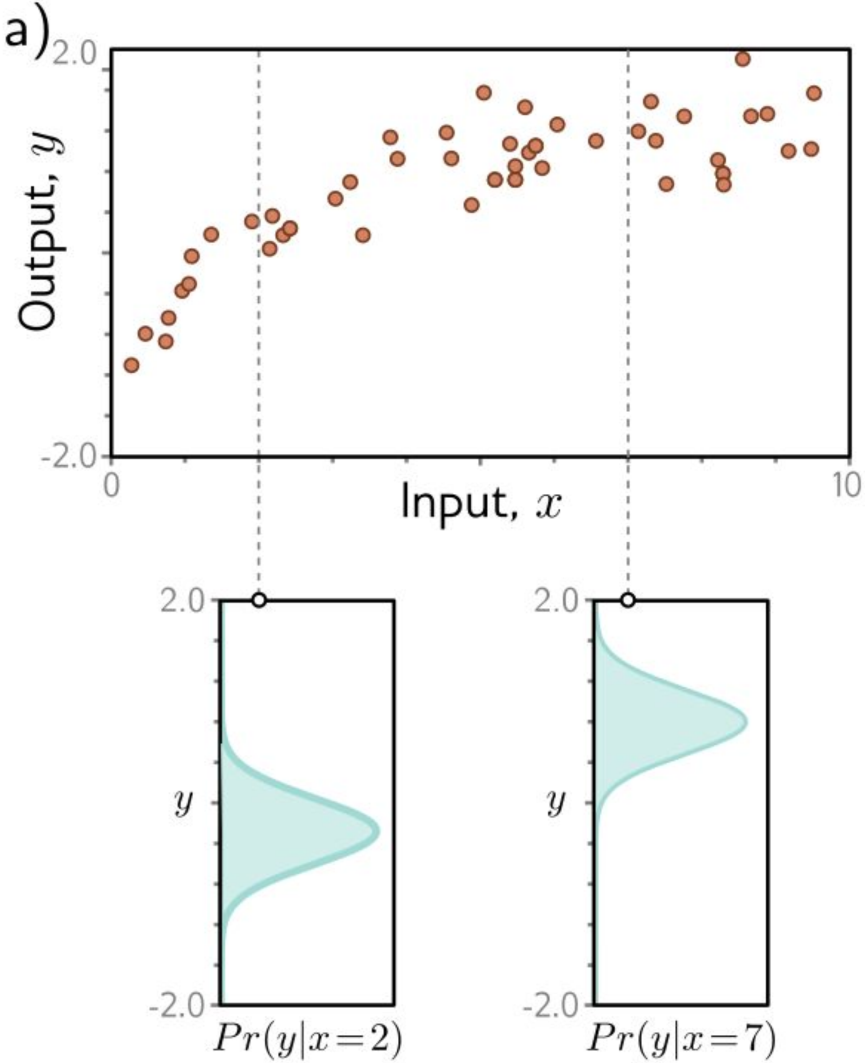
- Binary Classification $y \in \{0, 1\}$
- multiclass classification $y \in \{1, 2, \ldots, K\}$

Maximum Likelihood

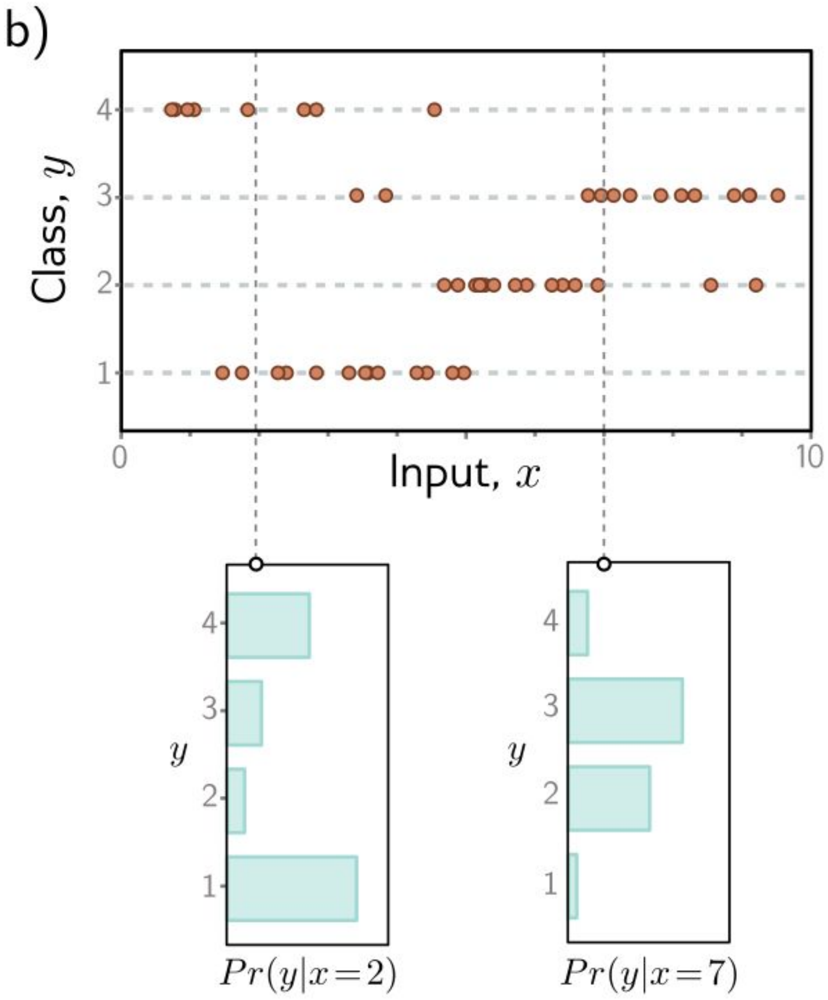
Conditional probability distribution P r(y|x)

$y \in \{0, 1\}$

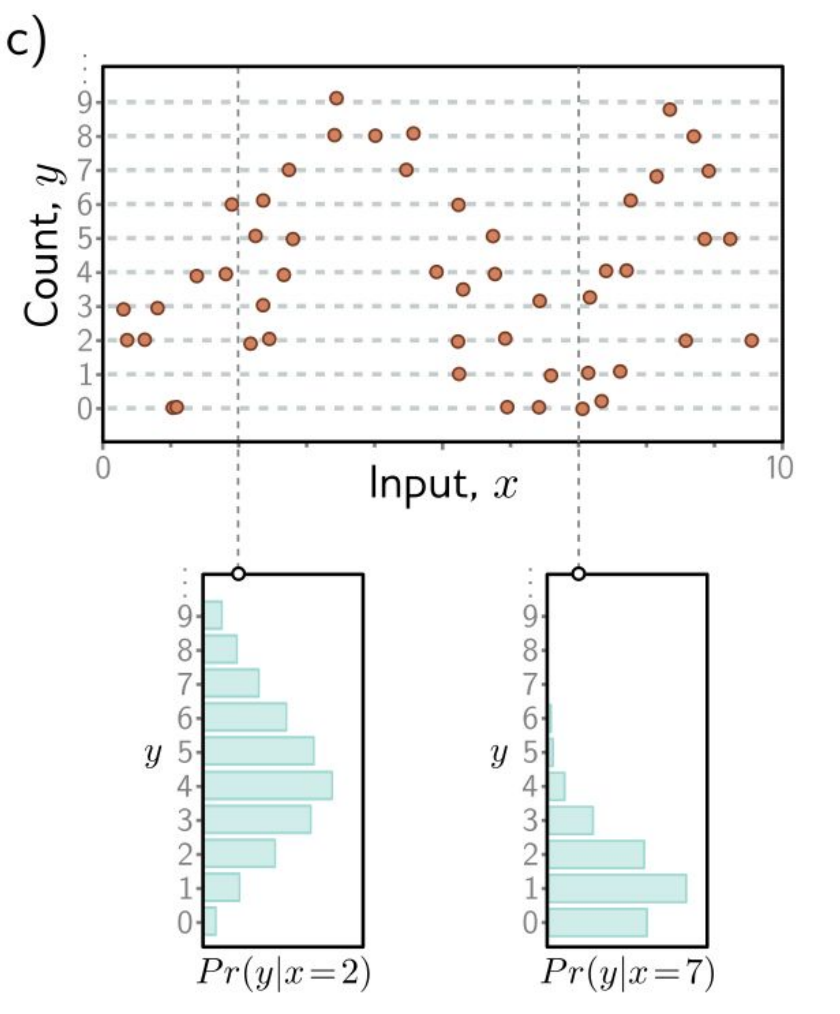
P(y | )

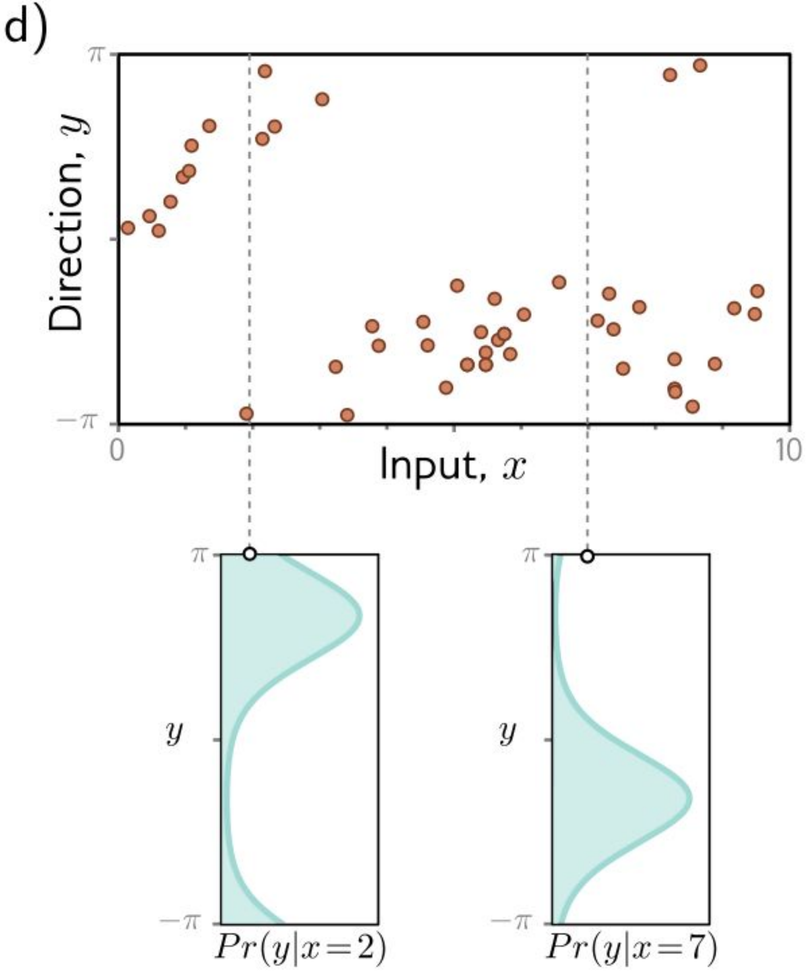


a)
2.0
-2.0
Output, y
0
10
Input, x
2.0
y
-2.0
$Pr(y|x=2)$
2.0
y
-2.0
$Pr(y|x=7)$



b)
Class, y
4
3
2
1
0
10
Input, x
y
4
3
2
1
$Pr(y|x=2)$
y
4
3
2
1
$Pr(y|x=7)$


c)
Count, y
Input, x
0
10
0
1
2
3
4
5
6
7
8
9
y
$Pr(y|x=2)$
$Pr(y|x=7)$



d)
Direction, y
Input, x
π
$-\pi$
0
10
y
$Pr(y|x=2)$
$Pr(y|x=7)$

Assumption

- We assume that the data are identically distributed
- We assume that the conditional distributions Pr(yi | x i) of the output given the input are independent

$$Pr(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_I | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_I) = \prod_{i=1}^{I} Pr(\mathbf{y}_i | \mathbf{x}_i)$$

Independent vs Dependent Variable

Loosely speaking:

An independent variable is a variable that is manipulated or categorized to observe its effect on another variable, typically called the dependent variable. In an experiment or study, the **independent variable is adjusted** or separated into categories **to examine its influence on the dependent variable**. By doing this, researchers can test hypotheses or theories and draw conclusions about the relationship between the two variables.

Independent vs Dependent Variables

Mathematically speaking:

An independent variable is a variable that represents a quantity that is being manipulated in an equation, function, or model. The independent variable is often considered as the "input" of a function. When the independent variable is altered, the function's output (the dependent variable) may change as a result.

Example:

$$y = 5x^3 - 2x^2 + 1$$

Independent vs Dependent **Random** Variables

Statistically speaking:

> For random variables X and Y, statistical independence can be formally defined in terms of their joint probability distribution. Two random variables X and Y are independent if and only if their joint probability distribution can be expressed as the product of their individual probability distributions

$$P(X_1 = a, X_2 = b) = P(X_1 = a) \times P(X_2 = b)$$
$$P(X_1 = a | X_2 = b) = P(X_1 = a)$$

Assumption

- We assume that the data are identically distributed
- We assume that the conditional distributions Pr(yi | x i) of the output given the input are independent

$$Pr(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_I | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_I) = \prod_{i=1}^{I} Pr(\mathbf{y}_i | \mathbf{x}_i)$$

Maximum likelihood criterion

$$\hat{\phi} = \underset{\phi}{\operatorname{argmax}} \left[\prod_{i=1}^{I} Pr(\mathbf{y}_i | \mathbf{x}_i) \right]$$

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

A few points

- If $f(x^*)$ is the maximum value of $f(x)$, then $log(f(x^*))$ will be the maximum value of $log(f(x))$, and it will occur at the same point $x = x^*$

- $log\ (a \times b) = log(a) + log\ (b)$

- $\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \ldots x_n$

Minimizing negative log-likelihood

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

Note the presence of the - sign and the transition from a maximization to a minimization function

Inference

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\left[Pr(\mathbf{y}|\mathbf{f}[\mathbf{x}, \hat{\phi}])\right]$$

Cross-entropy loss

Entropy

Entropy measures the **amount of uncertainty or randomness** in a random variable. The entropy H(X) of a discrete random variable X with probability mass function p(x) is given by:

$$H(X) = -\Sigma_x p(x) \log_2 p(x)$$

Cross-entropy

Given a **true distribution** ***p*** and an **estimated distribution** ***q***, the cross-entropy between them is defined as:

$$H(p, q) = \Sigma_i p(i) log(q_i)$$

What does it mean

Cross-entropy

If the model's ***predictions q*** are perfect, then ***q=p*** and the cross-entropy is simply the entropy of the true distribution ***p***

If the model's predictions deviate from the truth, the cross-entropy will be greater than the entropy of ***p***, with the difference indicating the extent of that deviation.

KL divergence

$$D_{\mathrm{KL}}(P \parallel Q) = -\sum_{x \in \mathcal{X}} P(x) \log\left(\frac{Q(x)}{P(x)}\right)$$

$$\begin{aligned} D_{\mathrm{KL}}(P \parallel Q) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)} - \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &= \mathrm{H}(P, Q) - \mathrm{H}(P) \end{aligned}$$