Parameter Estimation

Given samples $D = \{x^1, \ldots, X^N\}$, form of the distribution $p(x) = f(x; \mathbf{w})$ estimate values of the parameters w.

i.i.d assumption: each instance is independently and identically distributed.

Two methods:

1. Maximum likelihood estimation
2. Bayesian estimation

Aug 31	Sep 1	2	3	4	5
	0	1	1	0	0
Y	-100	-90	-75	-100	-98

$N = 5$

$= D$

$1(x) \sim \text{Bernoulli}(x; q)$

$p(\text{rain on Friday, 8th sep 2023}) = 0.5$

X

$Y =$ amount of rainfall in 1st 10 days of september.

$P(Y) \sim N(Y; \hat{\mu}; \sigma^2)$

$f(Y; w) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(Y-\mu)^2}{2\sigma^2}}$

$\hat{\mu}_{/D} = \frac{-\cancel{4}463}{5}$

$\hat{\sigma}^2_{/D} =$

$f(x; w) = w^x (1-w)^{1-x}$

w

q

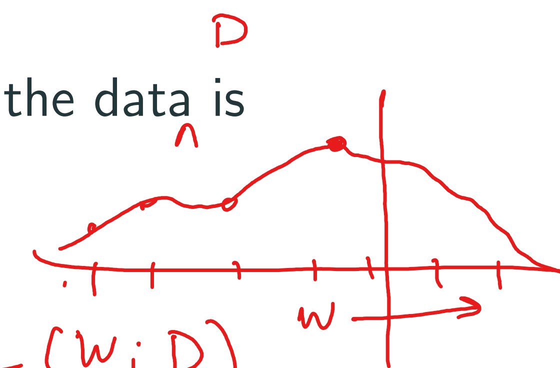
$= \begin{cases} w & \text{if } x = 1 \\ 1-w & \text{if } x = 0 \end{cases}$

$\hat{w} = 2/5$

Maximum Likelihood Estimation (MLE)

parameters

- Find the value **w** for which the probability (likelihood) of the data is maximized.



- Likelihood of data $L(w; D)$

$$L(w; D) = \prod_{i=1}^{N} f(x^i; w)$$

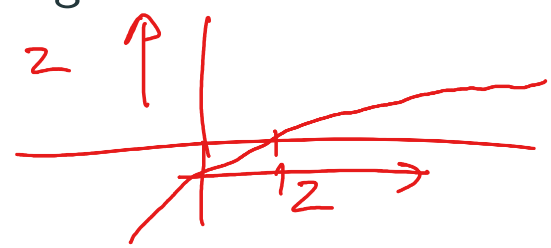
probability function given to us.

$$\hat{w} \equiv \underset{w}{\arg\max}\ L(w; D)$$

- Maximizing log-likelihood of data is equivalent to maximizing likelihood.

$$\boxed{\hat{w} \equiv \underset{w}{\arg\max}\ \log \prod_{i=1}^{N} f(x^i; w) = \sum_{i=1}^{N} \log f(x^i; w)}$$

log z



Solving the MLE objective

$$\max_{\mathbf{w}} LL(w; D)$$

Apply numerical optimization algorithms...e.g. stochastic gradient ascent

$w^0 =$ initial value of parameters.
for $t = 1$ to ....
$x^i \sim$ sample one example from D.
$w^{t+1} \leftarrow w^t + \eta \nabla_w \log f(x^i; w)$
η learning rate

For many simple distributions, the objective is concave in **w**. Maxima of **w** iff gradients w.r.t **w** is zero. Eg: $\log f(x^i; w)$ is concave for Bernoulli, Gaussian

If $\nabla_w LL(w^*; D) = 0$ then w^* is global maxima.

$$\iff \nabla_w \sum_{i=1}^{N} \log f(x^i; w) = 0$$

Parameter estimation for Bernoulli distribution

MLE for Bernoulli $\quad f(x;w) = w^{x}(1-w)^{1-x} \qquad x^{i} \in \{0, 1\}$

$$LL(w; D \equiv \{x^1, x^2, \dots x^N\}) = \sum_{i=1}^{N} \log w^{x^i}(1-w)^{1-x^i}$$

$$= \sum_{i=1}^{N} x^i \log w + (1-x^i)\log(1-w)$$

$$\nabla_w LL(w;D) = \frac{\partial}{\partial w}\sum_{i=1}^{N} x^i \log w + (1-x^i)\log(1-w)$$

$$= \sum_{i=1}^{N} x^i \frac{1}{w} + (1-x^i)\cdot\frac{1\,(-1)}{1-w} = \frac{\sum_{i=1}^{N} x^i}{w} - \frac{\sum_{i=1}^{N}(1-x^i)}{1-w} = 0$$

$$\Rightarrow (1-w)\,n_1(D) - w\,n_0(D) = 0$$

$$\Rightarrow n_1(D) = w(n_0(D) + n_1(D)) = wN \Rightarrow \hat{w} = \frac{n_1(D)}{N}$$

MLE for Gaussian distribution

$$f(x^i; \mathbf{w} \equiv (\mu, \sigma^2)) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad D \equiv \{x^1, x^2 \dots x^N\} \quad x^i \in R$$

$$LL(w; D) = \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x^i-\mu)^2}{2\sigma^2}}$$

$$= \sum_{i=1}^{N} -\frac{(x^i-\mu)^2}{2\sigma^2} - \log\sigma - \log\sqrt{2\pi}$$

← constant
– ignore.

$$\nabla_w LL(w; D) = 0 = \begin{bmatrix} \frac{\partial}{\partial\mu} LL(w;D) \\ \frac{\partial}{\partial\sigma} LL(w;D) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\frac{\partial}{\partial\mu}\left[-\sum_{i=1}^{N} \frac{(x^i-\mu)^2}{2\sigma^2} - \underbrace{\log\sigma}_{=0}\right] \Rightarrow -\sum_{i=1}^{N} \frac{\cancel{2}(x^i-\mu)}{\cancel{2}\sigma^2}(-1) = 0 \Rightarrow \sum_{i=1}^{N} x^i = \sum_{i=1}^{N} \mu = N\mu$$

$$\Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{N} x^i}{N}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} (x^i - \hat{\mu})^2}{N}$$

← MLE estimate.
was derived using $\hat{\mu}$ estimated from same data

As a result $\hat{\sigma}$ is known to have some "bias". statistical correction for the bias is obtained using

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{N} (x^i - \hat{\mu})^2}{N-1}.$$

MLE for Multinomial distribution

$x \in \{1, 2, \dots K\}$ $\quad K \geq 2$ $\quad$ x is also denoted as a "one-hot" vector

$x^i = \begin{bmatrix} x^i_1 \\ x^i_2 \\ \vdots \\ x^i_k \end{bmatrix}^K$ $\quad$ K $\quad$ $X = 3$ $\quad \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

$f(x \in \mathbb{R}^K; W = \{p_1, p_2 \dots p_K\}) = \prod_{j=1}^{K} p_j^{x_j}$

$\sum_{j=1}^{K} p_j = 1 \quad p_j \geq 0$

$D = \{x^1, x^2, \dots x^N\} \quad x^i \in \{0, 1\}^K$

$\max_{W = \{p_1 \dots p_K\}} \sum_{i=1}^{N} \log \prod_{j=1}^{K} p_j^{x^i_j}$ $\quad$ s.t $\underline{p_1 + p_2 + \dots p_K = 1}$

using Lagrangian multiplier:

$\underbrace{\max_{p_1, p_2 \dots p_K} \sum_{i=1}^{N} \sum_{j=1}^{K} x^i_j \log p_j + \lambda \left(\sum_{j=1}^{K} p_j - 1\right)}_{F(W; \lambda)}$

Lagrangian multiplier

$\nabla_{p_j} F(W; \lambda) = \sum_{i=1}^{N} x^i_j / p_j + \lambda \cdot 1 = 0 \Rightarrow \underline{p_j} = \dfrac{\boxed{\sum_{i=1}^{N} x^i_j}}{(-\lambda)} = \dfrac{n_j(D)}{(-\lambda)}$

$\because \sum_j p_j = 1$ we get $\sum_{j=1}^{K} \dfrac{n_j(D)}{(-\lambda)} = 1 \Rightarrow (-\lambda) = \left(\sum_{j=1}^{K} n_j(D)\right) = N$

$$\hat{p}_j = \frac{n_j(D)}{N}$$

Example: states from which students in a class came

$K = 30$; $\{1 = \text{Gujarat},\ 2 = \text{TN};\ 3 = \text{Mah},\ 4 = \text{Odisha}, \ldots\}$

We took three samples.

$$x^1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}; \quad x^2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}; \quad x^3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

MLE estimates

$$\hat{p}_1 = \frac{1}{3};\ \hat{p}_2 = \frac{1}{3};\ \hat{p}_3 = 0;\ \hat{p}_4 = \frac{1}{3},\ \hat{p}_5 = \hat{p}_6 = \cdots = \hat{p}_{30} = 0$$