Refresh:

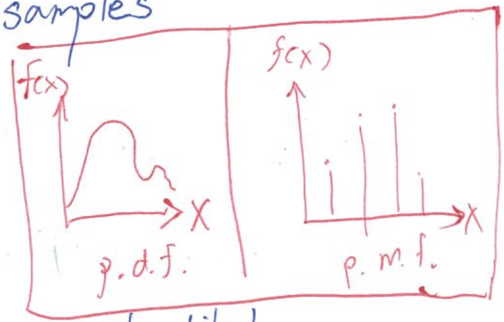
- In generative learning, the goal is to estimate

$$P(X, Y) = P(X|y)\,P(y).$$

- After that, we can use Bayes classifier that return $\arg\max_y P(y|x)$

Estimating Probability from Data:

- Let θ be a parameter of the modelling distribution. $\underbrace{f(x|\theta)}_{\text{p.m.f / p.d.f.}}$
- The data set D contains n random samples from $f(x|\theta)$.

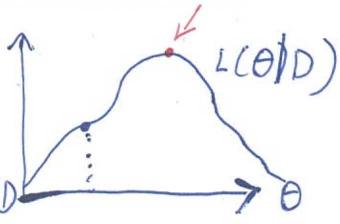


- Recall that the n samples are i.i.d.
- Likelihood estimation: Try to guess what is the most likely value that theta could be given the data set D we have observed.

Likelihood Function:

$L(\theta|D)$ = joint pmf of $x_1, \dots, x_n$ in D

Plug in the value — constant



The most likely value of θ that maximizes the likelihood function

$$L(\theta | D = \{(x_1,y_1), \dots, (x_n,y_n)\}) = f((x_1,y_1), \dots, (x_n,y_n) | \theta)$$

$$= f((x_1,y_1)|\theta) \cdots f((x_n,y_n)|\theta) \quad \text{(Each sample is i.i.d.)}$$

$$= \prod_{i}^{n} f((x_i,y_i)|\theta)$$

Maximum Likelihood Estimation (MLE):

- Principle: find $\theta_{max} = \arg\max_{\theta} (L(\theta | D))$
- The p.m.f. $f(D|\theta)$ ~~is~~ represents the modelling distribution assumption

---

Simple senario: Freethrows:

- Let's say we attempt $n=6$ freethrows and $n_H = 5$ make it in.

$$D = \{ \underbrace{1, 0, 1, 1, 1, 1}_{\text{i.i.d.}} \}$$

x_1 x_n

Intuity, we should guess

$$P(X=1) = \frac{5}{6}$$

- Each $x_i \sim P(X)$ ← To estimate

X: r.v. denoting the ~~~~ success ~~~~ at freethrows

- To estimate $P(X)$, we assume $x_i \sim f(D|\theta) = Bin(n, \theta)$

$$\Rightarrow f(D|\theta) = \binom{n}{n_H} \theta^{n_H} (1-\theta)^{n-n_H}$$

$$\Rightarrow L(\theta | D) = \binom{n}{n_H} \theta^{n_H} (1-\theta)^{n-n_H}$$

- In MLE, we will try to find θ_{max}

$$\theta_{max} = \arg\max_{\theta} \left(\binom{n}{n_H} \theta^{n_H} (1-\theta)^{n-n_H} \right)$$

$$= \arg\max_{\theta} \left(\log_e \left(\binom{n}{n_H} \theta^{n_H} (1-\theta)^{n-n_H} \right) \right)$$

$$= \arg\max_{\theta} \left(\underbrace{n_H \cdot \log\theta + (n - n_H) \log(1-\theta)}_{\log L(\theta|D)} \right)$$

- To find θ_{max}, we ~~...~~ consider the first derivative

$$\frac{dL}{d\theta} \overset{set}{=} 0$$

$$\Rightarrow \frac{dL}{d\theta} = \frac{d(n_H \cdot \log\theta + (n - n_H)\cdot \log(1-\theta))}{d\theta}$$

$$= \frac{n_H}{\theta} + \frac{n - n_H(-1)}{1-\theta} \overset{set}{=} 0$$

$$\Rightarrow \frac{n_H}{\theta} = \frac{n - n_H}{1-\theta}$$

$$\Rightarrow n_H(1-\theta) = (n - n_H)(\theta)$$

$$\Rightarrow n_H - n_H\theta = (n - n_H)(\theta)$$

$$\Rightarrow n\theta = n_H \Rightarrow \theta = \frac{n_H}{n} = \frac{5}{6}$$

Recap of MLE:

- Step 1: Make an explicit modeling assumption about what type of distribution the data was sampled from.
- Step 2: Set the parameter of the distribution so that the data observed is as likely as possible.

Simple Scenario II: Salmon or Mackerel

- For each type of fishes, we assume the Guassion distribution

$$\Rightarrow x_i \sim P(X)$$

r.v. denotes the fish length

Our intuition is we compute the mean and s.d.
$\Rightarrow$ We have the guassion distribution

- To estimate $P(X)$, we assume $x_i \sim \underbrace{f(x|\theta)}_{\text{p.d.f.}}$,
where $\theta = \{\mu, \sigma\}$, and

$$f(x|\theta) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- We define the likelihood function

$$L(\theta|D) = f(D|\theta)$$
$$= f(x_1|\theta)\cdot f(x_2|\theta)\cdots f(x_n|\theta)$$
$$= \prod_{i=1}^{n} f(x_i|\theta)$$

- Now, let's solve for θ_{max} that maximizes $L(\theta|D)$

$$\theta_{max} = \underset{\theta}{\arg\max}\left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right)$$

$$= \underset{\theta}{\arg\max}\left(\underbrace{\sum_{i=1}^{n} \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right)}\right)$$

$$\cancel{= \sum_{i=1}^{n} \left(-\frac{1}{2\sigma}(x_i-\mu)^2\right) \cdot \log\left(\frac{1}{\sigma\sqrt{2\pi}} e\right)}$$

$$\cancel{= \sum_{i=1}^{n} -\frac{1}{2\sigma}(x_i-\mu)^2 = \left[-\frac{1}{2\sigma}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{n}{}\right.}$$

$$\log\left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right) = \sum_{i=1}^{n} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right)$$

$$= \sum_{i=1}^{n} \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \log\left(e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right)$$

$$= \sum_{i=1}^{n} \overset{0}{\cancel{\log 1}} - \log(\sigma\sqrt{2\pi}) + \log\left(e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right)$$

$$= \sum_{i=1}^{n} -\log(\sigma) - \log(\sqrt{2\pi}) + \left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)$$

$$= -n\log(\sigma) - n\log(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2$$

$$l(\theta) = -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2$$

$$\frac{\partial l}{\partial \sigma}$$

$$\frac{\partial \log(L)}{\partial \mu} = \frac{\partial}{\partial \mu}\left(\cancel{\frac{-x_i^2}{2\sigma^2} + \frac{2\mu x_i}{2\sigma^2} + \frac{-\mu^2}{2\sigma^2}}\right)\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

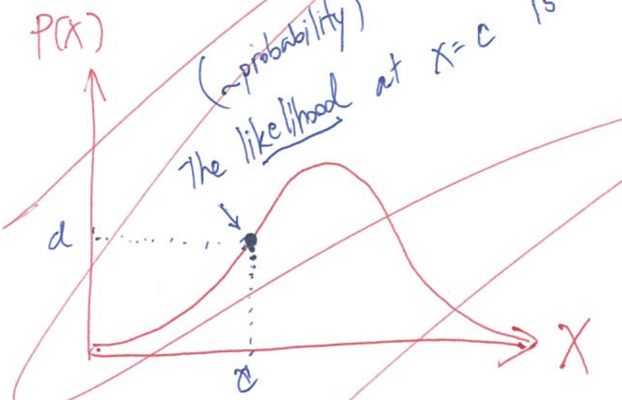
$$= \frac{\partial}{\partial \mu}\left(\frac{-1}{2\sigma^2} \cdot \sum_{i=1}^{n}\left(x_i^2 - 2\mu x_i + \mu^2\right)\right)$$

$$= \frac{-1}{2\sigma^2} \cdot \sum_{i=1}^{n} -2x_i + 2\mu = \frac{-1}{\cancel{2}\sigma^2}(-\cancel{2})\sum_{i=1}^{n}(x_i - \mu)$$

$$= \frac{1}{\sigma^2}\sum_{i}^{n}(x_i - \mu)$$

$$\frac{1}{\sigma^2}\sum_{i}^{n}(x_i-\mu) \overset{set}{=} 0 \longrightarrow \mu_{max} = \frac{1}{n}\sum_{i}^{n} x_i$$

p. d. f. $P(X)$



$$-\frac{1}{2}\cdot\log\left(\frac{e}{\sigma\sqrt{2\pi}}\right)\cdot\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

$$\sum_{i=1}^{n}-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\log\left(\frac{e}{\sigma\sqrt{2\pi}}\right)$$

$$= -\frac{1}{2}\cdot\left[1+\log(\sigma)+\log(2\pi)\right]\cdot\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

$$= -\frac{1}{2}+\frac{\log\sigma}{2}-\frac{\log(2\pi)}{2}.$$

$$= \frac{-1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\mu)^2\left(\log^{1} e-\log\sigma\sqrt{2\pi}\right)\right]$$

$$= \left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)-\cancel{\frac{\log\sigma}{2\sigma}}$$

$$\frac{\partial \log L}{\partial \sigma} = \frac{\partial}{\partial \sigma}\left(-n\log(\sigma) - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right)$$

Fact:
$\frac{\partial}{\partial x}\ln(x) = \frac{1}{x}$

$$= \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 \overset{set}{=} 0$$

$$\Rightarrow \quad \frac{n}{\sigma} = \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\Rightarrow \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

$$\sigma_{max} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2} \quad !!!$$

Note:
- $\mu_{max}, \sigma_{max} \Rightarrow \theta_{max}$ and they are what our intuition is about!!!

Summary of MLE:

- MLE gives the explanation of the data we observed.
- If n is large and your choise of distribution is correct, then MLE finds the "true" parameters
- MLE can overfit the data if n is small, it works well when n is large
- If you don't have the correct model, the MLE can be terribly wrong