Estimation, MLE

Machine Learning Techniques

Karthik Thiagarajan

References and Credits

- The content presented in these slides is derived from professor Arun Rajkumar's lectures and slides in the MLT course. This is the "ground truth" for almost all the content in these slides. As a result, these slides should be viewed as a presentation of the same content in the professor's lectures using a different medium. At the same time, these slides are not meant to be a replacement for the lectures.
- These slides were prepared using the tool mathcha.io.

Story so far

Unsupervised Learning

- Representation learning
  - PCA
  - Kernel PCA
- Clustering
  - Lloyd's algorithm (K-means)
- Estimation

Story so far

Unsupervised Learning

- Representation learning
  - PCA
  - Kernel PCA
- Clustering
  - Lloyd's algorithm (K-means)
- Estimation

Story so far

Unsupervised Learning

- Representation learning
  - PCA
  - Kernel PCA
- Clustering
  - Lloyd's algorithm (K-means)
- Estimation

Comprehension via Compression

- PCA
  -

Story so far

Unsupervised Learning

- Representation learning
  - PCA
  - Kernel PCA
- Clustering
  - Lloyd's algorithm (K-means)
- Estimation

Comprehension via Compression

- PCA
  - k principal components
- Clustering
  -

Story so far

Unsupervised Learning

- Representation learning
  - PCA
  - Kernel PCA
- Clustering
  - Lloyd's algorithm (K-means)
- Estimation

Comprehension via Compression

- PCA
  - k principal components
- Clustering
  - k means
- Estimation
  -

Story so far

Unsupervised Learning

- Representation learning
  - PCA
  - Kernel PCA
- Clustering
  - Lloyd's algorithm (K-means)
- Estimation

Comprehension via Compression

- PCA
  - k principal components
- Clustering
  - k means
- Estimation
  - k parameters

Estimation

Estimation

- Model the data-generation process
  - Where does the data come from?
  - What is the process that generates the data?
  - What are the parameters that govern this process?

Estimation

- Model the data-generation process
  - Where does the data come from?
  - What is the process that generates the data?
  - What are the parameters that govern this process?

- How is it different from clustering and representation learning?
  -

Estimation

- Model the data-generation process
  - Where does the data come from?
  - What is the process that generates the data?
  - What are the parameters that govern this process?

- How is it different from clustering and representation learning?
  - Present $\rightarrow$ Future
    - Start with what is given and try to explain patterns in it

Estimation

- Model the data-generation process
  - Where does the data come from?
  - What is the process that generates the data?
  - What are the parameters that govern this process?

- How is it different from clustering and representation learning?
  - Present $\rightarrow$ Future
    - Start with what is given and try to explain patterns in it
  - Present $\rightarrow$ Past
    - Explore the process that could have given rise to the data

Generative Process

$D = \{x_1, \cdots, x_n\}$

Bag of marbles $\longrightarrow$ Pick with replacement $\longrightarrow$ Marbles

Generative Process

$D = \{x_1, \cdots, x_n\}$

Bag of marbles → Pick with replacement → Marbles

Distribution → Sampling → Dataset

Generative Process

$$D = \{x_1, \cdots, x_n\}$$

$$
\begin{tikzcd}
\text{Bag of marbles} \arrow[r] \arrow[d] & \text{Pick with replacement} \arrow[r] \arrow[d] & \text{Marbles} \\
\text{Identitcally distributed} & \text{Independence} & \\
\text{Distribution} \arrow[r] & \text{Sampling} \arrow[r] & \text{Dataset}
\end{tikzcd}
$$

Maximum Likelihood

$$\begin{tikzcd}
\text{Distribution} \arrow[r] & \text{Sampling} \arrow[r] & \text{Dataset} \arrow[ll, bend left]
\end{tikzcd}$$

Maximum Likelihood

$$\begin{tikzcd}
\text{Distribution} \arrow[r] & \text{Sampling} \arrow[r] & \text{Dataset} \arrow[ll, bend left]
\end{tikzcd}$$

(1) Choose a distribution

(2) Estimate the parameters

Maximum Likelihood

$$\begin{tikzcd}
\text{Distribution} \arrow[r] & \text{Sampling} \arrow[r] & \text{Dataset} \arrow[ll, bend left]
\end{tikzcd}$$

(1) Choose a distribution

(2) Estimate the parameters

$$L(\theta; \{x_1, \cdots, x_n\}) = P(\{x_1, \cdots, x_n\}; \theta)$$

Maximum Likelihood

$$\begin{tikzcd}
\text{Distribution} \arrow[r] & \text{Sampling} \arrow[r] & \text{Dataset} \arrow[ll, bend left]
\end{tikzcd}$$

(1) Choose a distribution

(2) Estimate the parameters

$$L(\theta; \{x_1, \cdots, x_n\}) = P(\{x_1, \cdots, x_n\}; \theta)$$

Likelihood is treated as a function of θ

Maximum Likelihood

$$\begin{tikzcd} \text{Distribution} \arrow[r] & \text{Sampling} \arrow[r] & \text{Dataset} \arrow[ll, bend left] \end{tikzcd}$$

(1) Choose a distribution

(2) Estimate the parameters

$$\max_{\theta} \quad L(\theta; \{x_1, \cdots, x_n\})$$

$$L(\theta; \{x_1, \cdots, x_n\}) = P(\{x_1, \cdots, x_n\}; \theta)$$

$$\widehat{\theta} = \arg\max_{\theta} \quad L(\theta; \{x_1, \cdots, x_n\})$$

Likelihood is treated as a function of θ

Likelihood and Log-Likelihood

$D = \{x_1, \cdots, x_n\}$

$$L(\theta; \{x_1, \cdots, x_n\}) = P(\{x_1, \cdots, x_n\}; \theta)$$

Likelihood and Log-Likelihood

$$D = \{x_1, \cdots, x_n\}$$

$$\begin{aligned} L(\theta; \{x_1, \cdots, x_n\}) &= P(\{x_1, \cdots, x_n\}; \theta) \\ &= P(x_1; \theta) \cdots P(x_n; \theta) \end{aligned}$$

Likelihood and Log-Likelihood

$D = \{x_1, \cdots, x_n\}$

$$
\begin{aligned}
L(\theta; \{x_1, \cdots, x_n\}) &= P(\{x_1, \cdots, x_n\}; \theta) \\
&= P(x_1; \theta) \cdots P(x_n; \theta) \\
&= \prod_{i=1}^{n} P(x_i; \theta)
\end{aligned}
$$

Likelihood and Log-Likelihood

$D = \{x_1, \cdots, x_n\}$

$$\begin{aligned} L(\theta; \{x_1, \cdots, x_n\}) &= P(\{x_1, \cdots, x_n\}; \theta) \\ &= P(x_1; \theta) \cdots P(x_n; \theta) \\ &= \prod_{i=1}^{n} P(x_i; \theta) \end{aligned}$$

$$\begin{aligned} l(\theta;\ \{x_1, \cdots, x_n\}) &= \log L(\theta; \{x_1, \cdots, x_n\}) \\ &= \sum_{i=1}^{n} \log P(x_i; \theta) \end{aligned}$$

Likelihood and Log-Likelihood

$D = \{x_1, \cdots, x_n\}$

$$\begin{aligned} L(\theta; \{x_1, \cdots, x_n\}) &= P(\{x_1, \cdots, x_n\}; \theta) \\ &= P(x_1; \theta) \cdots P(x_n; \theta) \\ &= \prod_{i=1}^{n} P(x_i; \theta) \end{aligned}$$

$$\begin{aligned} l(\theta;\ \{x_1, \cdots, x_n\}) &= \log L(\theta; \{x_1, \cdots, x_n\}) \\ &= \sum_{i=1}^{n} \log P(x_i; \theta) \end{aligned}$$

- **Problem:** Likelihood involves product of quantities less than 1

Likelihood and Log-Likelihood

$D = \{x_1, \cdots, x_n\}$

$$\begin{aligned} L(\theta; \{x_1, \cdots, x_n\}) &= P(\{x_1, \cdots, x_n\}; \theta) \\ &= P(x_1; \theta) \cdots P(x_n; \theta) \\ &= \prod_{i=1}^{n} P(x_i; \theta) \end{aligned}$$

$$\begin{aligned} l(\theta;\ \{x_1, \cdots, x_n\}) &= \log L(\theta; \{x_1, \cdots, x_n\}) \\ &= \sum_{i=1}^{n} \log P(x_i; \theta) \end{aligned}$$

- **Problem:** Likelihood involves product of quantities less than 1
- **Solution**: Use $\log$ as it converts products to sums

Likelihood and Log-Likelihood

$D = \{x_1, \cdots, x_n\}$

$$\begin{aligned} L(\theta; \{x_1, \cdots, x_n\}) &= P(\{x_1, \cdots, x_n\}; \theta) \\ &= P(x_1; \theta) \cdots P(x_n; \theta) \\ &= \prod_{i=1}^{n} P(x_i; \theta) \end{aligned}$$

$$\begin{aligned} l(\theta;\ \{x_1, \cdots, x_n\}) &= \log L(\theta; \{x_1, \cdots, x_n\}) \\ &= \sum_{i=1}^{n} \log P(x_i; \theta) \end{aligned}$$

- **Problem:** Likelihood involves product of quantities less than 1
- **Solution**: Use $\log$ as it converts products to sums
- **Allowed**: $\log$ is strictly increasing

Likelihood and Log-Likelihood

$D = \{x_1, \cdots, x_n\}$

$$\begin{aligned} L(\theta; \{x_1, \cdots, x_n\}) &= P(\{x_1, \cdots, x_n\}; \theta) \\ &= P(x_1; \theta) \cdots P(x_n; \theta) \\ &= \prod_{i=1}^{n} P(x_i; \theta) \end{aligned}$$

$$\begin{aligned} l(\theta;\ \{x_1, \cdots, x_n\}) &= \log L(\theta; \{x_1, \cdots, x_n\}) \\ &= \sum_{i=1}^{n} \log P(x_i; \theta) \end{aligned}$$

$$x^* \text{ maximizes } f \iff x^* \text{ maximizes } \log f$$

- **Problem:** Likelihood involves product of quantities less than 1
- **Solution**: Use $\log$ as it converts products to sums
- **Allowed**: $\log$ is strictly increasing

Likelihood and Log-Likelihood

$$D = \{x_1, \cdots, x_n\}$$

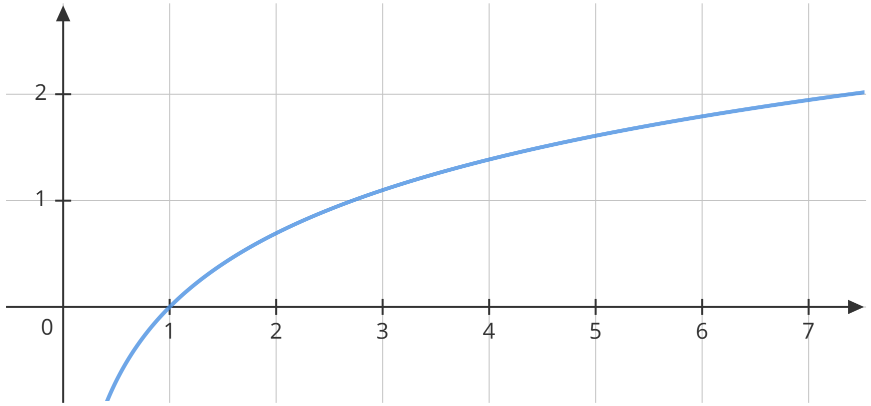
$$\begin{aligned} L(\theta; \{x_1, \cdots, x_n\}) &= P(\{x_1, \cdots, x_n\}; \theta) \\ &= P(x_1; \theta) \cdots P(x_n; \theta) \\ &= \prod_{i=1}^{n} P(x_i; \theta) \end{aligned}$$

$$\begin{aligned} l(\theta;\ \{x_1, \cdots, x_n\}) &= \log L(\theta; \{x_1, \cdots, x_n\}) \\ &= \sum_{i=1}^{n} \log P(x_i; \theta) \end{aligned}$$

$$x^* \text{ maximizes } f \iff x^* \text{ maximizes } \log f$$

$$f(x) \leqslant f\big(x^*\big) \iff \log(f(x)) \leqslant \log\big(f\big(x^*\big)\big)$$

- **Problem:** Likelihood involves product of quantities less than 1
- **Solution**: Use $\log$ as it converts products to sums
- **Allowed**: $\log$ is strictly increasing



Bernoulli

$X_i \sim Br(p)$

Bernoulli

$$P(X_i = 1) = p$$
$$P(X_i = 0) = 1 - p$$

$X_i \sim Br(p)$

Bernoulli

$$P(X_i = 1) = p$$
$$P(X_i = 0) = 1 - p$$

$$X_i \sim Br(p)$$

$$P(X_i = x_i) = p^{x_i} \cdot (1 - p)^{1 - x_i}$$

Bernoulli

$$L(p; D) = \prod_{i=1}^{n} P(X_i = x_i)$$

$$P(X_i = 1) = p$$
$$P(X_i = 0) = 1 - p$$

$$X_i \sim Br(p)$$

$$P(X_i = x_i) = p^{x_i} \cdot (1 - p)^{1 - x_i}$$

Bernoulli

$$L(p; D) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

$$P(X_i = 1) = p$$
$$P(X_i = 0) = 1 - p$$

$$X_i \sim Br(p)$$

$$P(X_i = x_i) = p^{x_i} \cdot (1-p)^{1-x_i}$$

Bernoulli

$$L(p; D) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$
$$= p^{n_h}(1-p)^{n-n_h}$$

$$P(X_i = 1) = p$$
$$P(X_i = 0) = 1 - p$$

$$X_i \sim Br(p)$$

$$P(X_i = x_i) = p^{x_i} \cdot (1-p)^{1-x_i}$$

Bernoulli

$$L(p; D) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$
$$= p^{n_h}(1-p)^{n-n_h}$$

$$P(X_i = 1) = p$$
$$P(X_i = 0) = 1 - p$$

$$l(p; D)$$

$$X_i \sim Br(p)$$

$$P(X_i = x_i) = p^{x_i} \cdot (1-p)^{1-x_i}$$

Bernoulli

$$L(p; D) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$
$$= p^{n_h}(1-p)^{n-n_h}$$

$$P(X_i = 1) = p$$
$$P(X_i = 0) = 1 - p$$

$$l(p; D) = n_h \log p + (n - n_h)\log(1-p)$$

$$X_i \sim Br(p)$$

$$P(X_i = x_i) = p^{x_i} \cdot (1-p)^{1-x_i}$$

Bernoulli

$$L(p; D) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$
$$= p^{n_h}(1-p)^{n-n_h}$$

$$P(X_i = 1) = p$$
$$P(X_i = 0) = 1 - p$$

$$l(p; D) = n_h \log p + (n - n_h)\log(1-p)$$
$$\frac{dl(p; D)}{dp} =$$

$$X_i \sim Br(p)$$

$$P(X_i = x_i) = p^{x_i} \cdot (1-p)^{1-x_i}$$

Bernoulli

$$L(p;D) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$
$$= p^{n_h}(1-p)^{n-n_h}$$

$$P(X_i = 1) = p$$
$$P(X_i = 0) = 1 - p$$

$$l(p;D) = n_h \log p + (n - n_h)\log(1-p)$$
$$\frac{dl(p;D)}{dp} = \frac{n_h}{p} - \frac{n-n_h}{1-p}$$

$$X_i \sim Br(p)$$

$$P(X_i = x_i) = p^{x_i} \cdot (1-p)^{1-x_i}$$

Bernoulli

$$L(p; D) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$
$$= p^{n_h}(1-p)^{n-n_h}$$

$$P(X_i = 1) = p$$
$$P(X_i = 0) = 1 - p$$

$$l(p; D) = n_h \log p + (n - n_h)\log(1-p)$$
$$\frac{dl(p; D)}{dp} = \frac{n_h}{p} - \frac{n - n_h}{1 - p}$$

$$X_i \sim Br(p)$$

$$\frac{dl(p; D)}{dp} = 0$$

$$P(X_i = x_i) = p^{x_i} \cdot (1-p)^{1-x_i}$$

Bernoulli

$$L(p; D) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$
$$= p^{n_h} (1-p)^{n-n_h}$$

$$P(X_i = 1) = p$$
$$P(X_i = 0) = 1 - p$$

$$l(p; D) = n_h \log p + (n - n_h) \log(1-p)$$
$$\frac{dl(p; D)}{dp} = \frac{n_h}{p} - \frac{n - n_h}{1 - p}$$

$$X_i \sim Br(p)$$

$$\frac{dl(p; D)}{dp} = 0$$
$$(1-p)n_h - (n - n_h)p = 0$$
$$p = \frac{n_h}{n}$$

$$P(X_i = x_i) = p^{x_i} \cdot (1-p)^{1-x_i}$$