Maximum Likelihood Estimation

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

References:
Sample Statistics and Population Parameters

• A Schematic Depiction
Introduction

• Statistic
  – Any value (or function) that is calculated from a given sample
  – Statistical inference: make a decision using the information provided by a sample (or a set of examples/instances)

• Parametric methods
  – Assume that examples are drawn from some distribution that obeys a known model \( p(x) \)
  – Advantage: the model is well defined up to a small number of parameters
    • E.g., mean and variance are sufficient statistics for the Gaussian distribution
  – Model parameters are typically estimated by either maximum likelihood estimation or Bayesian (MAP) estimation
Maximum Likelihood Estimation (MLE) (1/2)

- Assume the instances $x = \{x^1, x^2, \ldots, x^t, \ldots, x^N\}$ are independent and identically distributed (iid), and drawn from some known probability distribution $X$
  
  - $X^t \sim p(x^t | \theta)$
  - $\theta$: model parameters (assumed to be fixed but unknown here)

- MLE attempts to find $\theta$ that make $x$ the most likely to be drawn
  
  - Namely, maximize the likelihood of the instances

\[
l(\theta | x) = p(x | \theta) = p(x^1, \cdots, x^N | \theta) = \prod_{t=1}^{N} p(x^t | \theta)
\]
MLE (2/2)

• Because logarithm will not change the value of \( \theta \) when it take its maximum \( \text{(monotonically increasing/decreasing)} \)
  – Finding \( \theta \) that maximizes the likelihood of the instances is equivalent to finding \( \theta \) that maximizes the log likelihood of the samples

\[
L(\theta | x) = \log l(\theta | x) = \sum_{t=1}^{N} \log p(x^t | \theta)
\]

  – As we shall see, logarithmic operation can further simplify the computation when estimating the parameters of those distributions that have exponents

\[ a \geq b \implies \log a \geq \log b \]
MLE: Bernoulli Distribution (1/3)

- Bernoulli Distribution
  - A random variable $X$ takes either the value $x=1$ (with probability $r$) or the value $x=0$ (with probability $1-r$).
  - Can be thought of as $X$ is generated from two distinct states.
  - The associated probability distribution:
    \[ P(x) = r^x(1-r)^{1-x}, \quad x \in \{0,1\} \]

- The log likelihood for a set of iid instances $\mathbf{x}$ drawn from Bernoulli distribution:
  \[ L(r | \mathbf{X}) = \log \prod_{t=1}^{N} r^{x^t}(1 - r)^{(1-x^t)} \]
  \[ \theta = \sum_{t=1}^{N} \log \left[ r^{x^t}(1 - r)^{(1-x^t)} \right] \]
  \[ = \left( \sum_{t=1}^{N} x^t \right) \log r + \left( N - \sum_{t=1}^{N} x^t \right) \log (1 - r) \]
MLE: Bernoulli Distribution (2/3)

- **MLE of the distribution parameter** $r$
  
  $$\hat{r} = \frac{\sum_{t=1}^{N} x^t}{N}$$

  - The estimate for $r$ is the ratio of the number of occurrences of the event ($x^t = 1$) to the number of experiments

- **The expected value for $X$**
  
  $$E[X] = \sum_{x \in \{0,1\}} x \cdot P(x) = 0 \cdot (1 - r) + 1 \cdot r = r$$

- **The variance value for $X$**
  
  $$\text{var}(X) = E[X^2] - (E[X])^2 = r - r^2 = r(1 - r)$$
MLE: Bernoulli Distribution (3/3)

- Appendix A

\[
\left( \sum_{t=1}^{N} x^t \right) \log r + \left( N - \sum_{t=1}^{N} x^t \right) \log \left( 1 - r \right)
\]

\[
\frac{dL(r|X)}{dr} = \frac{\partial}{\partial r} \int \left( \sum_{t=1}^{N} x^t \right) \log r + \left( N - \sum_{t=1}^{N} x^t \right) \log \left( 1 - r \right) dr = 0
\]

\[
\Rightarrow \frac{\left( \sum_{t=1}^{N} x^t \right)}{r} - \frac{\left( N - \sum_{t=1}^{N} x^t \right)}{1 - r} = 0
\]

\[
\Rightarrow \hat{r} = \frac{\sum_{t=1}^{N} x^t}{N}
\]

The maximum likelihood estimate of the mean is the sample average.
MLE: Multinomial Distribution (1/4)

- Multinomial Distribution
  - A generalization of Bernoulli distribution
  - The value of a random variable $X$ can be one of $K$ mutually exclusive and exhaustive states $x \in \{s_1, s_2, \ldots, s_K\}$ with probabilities $r_1, r_2, \ldots, r_K$, respectively
  - The associated probability distribution
    $$p(x) = \prod_{i=1}^{K} r_{i}^{s_i}, \quad \sum_{i=1}^{K} r_i = 1$$
    $$s_i = \begin{cases} 
    1 & \text{if } X \text{ choose state } s_i \\
    0 & \text{otherwise} 
    \end{cases}$$

- The log likelihood for a set of iid instances $x$ drawn from a multinomial distribution $X$
  $$L(r|x) = \log \prod_{t=1}^{N} \prod_{i=1}^{K} r_{i}^{s_{i}} \quad x = \{x^1, x^2, \ldots, x^t, \ldots, x^N\}$$
MLE: Multinomial Distribution (2/4)

• MLE of the distribution parameter $r_i$

$$\hat{r}_i = \frac{\sum_{t=1}^{N} s_i^t}{N}$$

- The estimate for $r_i$ is the ratio of the number of experiments with outcome of state $i$ ($s_i^t = 1$) to the number of experiments
MLE: Multinomial Distribution (3/4)

- Appendix B

\[ L(r|x) = \log \prod_{i=1}^{N} \prod_{t=1}^{K} r_{si} = \sum_{i=1}^{N} \sum_{t=1}^{K} \log r_{si} , \text{ with constraint } \sum_{i=1}^{K} r_{i} = 1 \]

\[ \frac{\partial \bar{L}(r|x)}{\partial r_{i}} = \frac{\partial}{\partial r_{i}} \left[ \sum_{i=1}^{N} \sum_{t=1}^{K} s_{i} \cdot \log r_{i} + \lambda \left( \sum_{i=1}^{K} r_{i} - 1 \right) \right] = 0 \]

Lagrange Multiplier

\[ \Rightarrow \sum_{i=1}^{N} s_{i} \cdot \frac{1}{r_{i}} + \lambda = 0 \]

\[ \Rightarrow r_{i} = -\frac{1}{\lambda} \sum_{i=1}^{N} s_{i} \]

\[ \Rightarrow \sum_{i=1}^{K} r_{i} = 1 = -\frac{1}{\lambda} \sum_{i=1}^{N} \left( \sum_{t=1}^{K} s_{t} \right) \]

\[ \Rightarrow \lambda = -N \]

\[ \Rightarrow \hat{r}_{i} = \frac{1}{N} \sum_{t=1}^{N} s_{i} \]


Berlin Chen 11
MLE: Multinomial Distribution (4/4)

P(B)=3/10
P(W)=4/10
P(R)=3/10
MLE: Gaussian Distribution (1/3)

• Also called Normal Distribution
  – Characterized with mean \( \mu \) and variance \( \sigma^2 \)

\[
p(x) = \frac{1}{\sqrt{2\pi \sigma}} \exp\left[ -\frac{(x - \mu)^2}{2\sigma^2} \right], \quad -\infty < x < \infty
\]

  – Recall that mean and variance are sufficient statistics for Gaussian

• The log likelihood for a set of \( iid \) instances drawn from Gaussian distribution \( X \)

\[
L(\mu, \sigma \mid x) = \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi \sigma}} e^{-\frac{\left( x^t - \mu \right)^2}{2\sigma^2}} \quad x = \{x_1, x_2, \ldots, x^t, \ldots, x^N\}
\]

\[
= -\frac{N}{2} \log (2\pi) - N \log \sigma - \frac{\sum_{t=1}^{N} \left( x^t - \mu \right)^2}{2\sigma^2}
\]
MLE: Gaussian Distribution (2/3)

- MLE of the distribution parameters $\mu$ and $\sigma^2$

$$m = \hat{\mu} = \frac{\sum_{t=1}^{N} x^t}{N}$$  
sample average

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{t=1}^{N} (x^t - m)^2}{N}$$  
sample variance

- Remind that $\mu$ and $\sigma^2$ are still fixed but unknown
MLE: Gaussian Distribution (3/3)

- Appendix C

\[
L(\mu, \sigma | x) = -\frac{N}{2} \log (2\pi) - \frac{N}{2} \log \sigma^2 - \frac{\sum_{t=1}^{N} (x^t - \mu)^2}{2\sigma^2}
\]

\[
\frac{\partial L(\mu, \sigma | x)}{\partial \mu} = 0 \Rightarrow \frac{1}{\sigma^2} \sum_{t=1}^{N} (x^t - \mu) = 0 \Rightarrow \hat{\mu} = \frac{\sum_{t=1}^{N} x^t}{N}
\]

\[
\frac{\partial L(\mu, \sigma | x)}{\partial \sigma^2} = 0 \Rightarrow -N + \frac{1}{\sigma^2} \sum_{t=1}^{N} (x^t - \mu)^2 = 0 \Rightarrow \hat{\sigma}^2 = \frac{\sum_{t=1}^{N} (x^t - \hat{\mu})^2}{N}
\]
Evaluating an Estimator: Bias and Variance (1/6)

• The mean square error of the estimator \( d \) can be further decomposed into two parts respectively composed of bias and variance:

\[
gr(d, \theta) = E[(d - \theta)^2] \\
= E[(d - E[d] + E[d] - \theta)^2] \\
= E[(d - E[d])^2 + (E[d] - \theta)^2 + 2(d - E[d])(E[d] - \theta)] \\
= E[(d - E[d])^2] + E[(E[d] - \theta)^2] + 2E[(d - E[d])(E[d] - \theta)] \\
= E[(d - E[d])^2] + (E[d] - \theta)^2 + 2E[(d - E[d])(E[d] - \theta)] \\
= E[(d - E[d])^2] + (E[d] - \theta)^2 \\
\]

Constant terms cancel out, leaving:

\[
gr(d, \theta) = \text{variance} + \text{bias}^2
\]
Evaluating an Estimator: Bias and Variance (2/6)

Figure 4.1: $\theta$ is the parameter to be estimated. $d_i$ are several estimates (denoted by ‘\times’) over different samples. Bias is the difference between the expected value of $d$ and $\theta$. Variance is how much $d_i$ are scattered around the expected value. We would like both to be small.
Example 1: sample average and sample variance

- Assume samples $\{x^1, x^2, \ldots, x^t, \ldots, x^N\}$ are independent and identically distributed (iid), and drawn from some known probability distribution $X$ with mean $\mu$ and variance $\sigma^2$.

Mean: $\mu = E[X] = \sum_x x \cdot p(x)$

Variance: $\sigma^2 = E[(X - \mu)^2] = E[X^2] - (E[X])^2$

Sample average (mean) for the observed samples: $m = \frac{1}{N} \sum_{t=1}^{N} x^t$

Sample variance for the observed samples: $s^2 = \frac{1}{N} \sum_{t=1}^{N} (x^t - m)^2$

or $s^2 = \frac{1}{N-1} \sum_{t=1}^{N} (x^t - m)^2$?
Evaluating an Estimator: Bias and Variance (4/6)

• Example 1 (count.)
  – Sample average \( m \) is an unbiased estimator of the mean \( \mu \)
    \[
    E[m] = E\left[\frac{1}{N} \sum_{t=1}^{N} X^t\right] = \frac{1}{N} \sum_{t=1}^{N} E[X] = \frac{N \cdot \mu}{N} = \mu
    \]
    \[
    \therefore E[m] - \mu = 0
    \]
  
  • \( m \) is also a consistent estimator: \( \text{Var}(m) \to 0 \) as \( N \to \infty \)
    \[
    \text{Var}(m) = \text{Var}\left(\frac{1}{N} \sum_{t=1}^{N} X^t\right) = \frac{1}{N^2} \sum_{t=1}^{N} \text{Var}(X) = \frac{N \cdot \sigma^2}{N^2} = \sigma^2 \frac{N}{N} \xrightarrow{N=\infty} 0
    \]

\[
\text{Var}(aX + b) = a^2 \cdot \text{Var}(X)
\]
\[
\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)
\]
Evaluating an Estimator: Bias and Variance (5/6)

• Example 1 (count.)
  – Sample variance $s^2$ is an asymptotically unbiased estimator of the variance $\sigma^2$

$$E[s^2] = E \left[ \frac{1}{N} \sum_{t=1}^{N} (X^t - m)^2 \right]$$

$$= E \left[ \frac{1}{N} \sum_{t=1}^{N} (X - m)^2 \right] \quad (X^t's \ are \ i.i.d.)$$

$$= E \left[ \frac{1}{N} \sum_{t=1}^{N} (X^2 - 2X \cdot m + m^2) \right]$$

$$= E \left[ \frac{N \cdot X^2 - 2N \cdot m^2 + Nm^2}{N} \right]$$

$$= E \left[ \frac{N \cdot X^2 - N \cdot m^2}{N} \right] = \frac{N \cdot E[X^2] - N \cdot E[m^2]}{N}$$

$s^2 = \frac{1}{N} \sum_{t=1}^{N} (x^t - m)^2$

$\sum_{t=1}^{N} x^t = N \cdot m$
Evaluating an Estimator: Bias and Variance (6/6)

- Example 1 (count.)
  - Sample variance $s^2$ is an asymptotically unbiased estimator of the variance $\sigma^2$

\[
E \left[ S^2 \right] = \frac{N \cdot E \left[ X^2 \right] - N \cdot E \left[ m^2 \right]}{N}
\]

\[
= \frac{N \left( \sigma^2 + \mu^2 \right) - N \left( \frac{\sigma^2}{N} + \mu^2 \right)}{N}
\]

\[
= \frac{(N - 1) \sigma^2}{N} \quad \text{as } N = \infty \rightarrow \sigma^2
\]

The size of the observed sample set
Bias and Variance: Example 2

![Diagram showing different samples for an unknown population: \( X \rightarrow (x, y) \), \( y = F(x) \)]

\[
y' = F(x) + \varepsilon
\]

error of measurement
Simple is Elegant?