National Ph.D. Program in *Artificial Intelligence for Society*

Statistics for Machine Learning

Lesson 06 - Unbiased estimators. Efficiency and MSE. Maximum likelihood estimation.

Andrea Pugnana, Salvatore Ruggieri

Department of Computer Science
University of Pisa, Italy
andrea.pugnana@di.unipi.it salvatore.ruggieri@unipi.it

Statistical model for repeated measurement

- A dataset $x_1, \ldots, x_n$ consists of repeated measurements of a phenomenon we are interested in understanding
  - E.g., measurement of the speed of light
- We model a dataset as the realization of a random sample

Random sample

A *random sample* is a collection of i.i.d. random variables $X_1, \ldots, X_n \sim F(\alpha)$, where $F()$ is the distribution and α its parameter(s).

- Joint distribution $P(X_1, \ldots, X_n) = \prod_{i=1}^{m} \mathrm{P}(\mathrm{X}_i) \sim F^n(\alpha)$
- Challenging questions/inferences on a population given a sample:
  - How to determine $E[X]$, $Var(X)$, or other functions of X?
  - How to determine α, assuming to know the form of F?
  - How to determine both F and α?

An example

Table 17.1. Michelson data on the speed of light.

850	740	900	1070	930	850	950	980	980	880
1000	980	930	650	760	810	1000	1000	960	960
960	940	960	940	880	800	850	880	900	840
830	790	810	880	880	830	800	790	760	800
880	880	880	860	720	720	620	860	970	950
880	910	850	870	840	840	850	840	840	840
890	810	810	820	800	770	760	740	750	760
910	920	890	860	880	720	840	850	850	780
890	840	780	810	760	810	790	810	820	850
870	870	810	740	810	940	950	800	810	870

- What is an estimate of the true speed of light (estimand)?

$$x_1 = 850, \text{ or } min\ x_i, \text{ or } max\ x_i, \text{ or } \bar{x}_n = 852.4\ ?$$

An example

- Speed of light dataset as realization of

$$X_i = c + \epsilon_i$$

where ϵ_i is measurement error with $E[\epsilon_i] = 0$ and $Var(\epsilon_i) = \sigma^2$

- We are then interested in $E[X_i] = c$
- How to estimate it?
- Use some data. For X_1:

$$E[X_1] = c \qquad Var(X_1) = \sigma^2$$

- Use all data. For $\bar{X}_n = (X_1 + \ldots + X_n)/n$:

$$E[\bar{X}_n] = c \qquad Var(\bar{X}_n) = \frac{Var(X_1)}{n} = \frac{\sigma^2}{n}$$

Hence, for $n \to \infty$, $Var(\bar{X}_n) \to 0$

Estimate

Estimand and estimate

An *estimand* θ is an unknown parameter of a distribution $F()$.
An *estimate* t of θ is a value obtained as a function $h()$ over a dataset $x_1, \ldots, x_n$:

$$t = h(x_1, \ldots, x_n)$$

- $t = \bar{x}_n = 852.4$ is an estimate of the speed of light (estimand) $\quad t = x_1 = 850$ is another estimate
- Since $x_1, \ldots, x_n$ are modelled as realizations of $X_1, \ldots, X_n$, estimates are realizations of the corresponding sample statistics $h(X_1, \ldots, X_n)$

Statistics and estimator

A *statistics* is a function of $h(X_1, \ldots, X_n)$ of r.v.'s.
An *estimator* of a parameter θ is a statistics $T_n = h(X_1, \ldots, X_n)$ intended to provide information about θ.

- An estimate $t = h(x_1, \ldots, x_n)$ is a realization of the estimator $T_n = h(X_1, \ldots, X_n)$
- $T_n = \bar{X}_n = (X_1 + \ldots, X_n)/n$ is an estimator of $\mu \quad T_n = X_1$ is another estimator

Unbiased estimator

- The probability distribution of an estimator T is called the *sampling distribution* of T

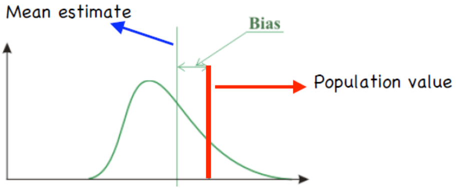
Unbiased estimator

An estimator $T_n = h(X_1, \ldots, X_n)$ of a parameter θ (estimand) is *unbiased* if:

$$E[T_n] = \theta$$

If the difference $E[T_n] - \theta$, called the *bias* of T_n, is non-zero, T_n is called a *biased* estimator.

- $E[T_n] > \theta$ is a positive bias, $E[T_n] < \theta$ is a negative bias
- **Asymptotically unbiased:** $\lim_{n\to\infty} E[T_n] = \theta$
- Sometimes, T_n written as $\hat{\theta}$, e.g., $\hat{\mu}$ estimator of μ



When is an estimator better than another one?

Efficiency of unbiased estimators

Let T_1 and T_2 be unbiased estimators of the same parameter θ. The estimator T_2 is *more efficient* than T_1 if:

$$Var(T_2) < Var(T_1)$$

- The *relative efficiency* of T_2 w.r.t. T_1 is ${}^{Var(T_1)}/_{Var(T_2)}$
- Speed of light example:
  - $E[X_1] = E[X_2] = \ldots = E[\bar{X}_n] = c$, i.e., all unbiased estimators

  The mean is more efficient than a single value

$$Var(\bar{X}_n) = \sigma^2/n < \sigma^2 = Var(X_1) \qquad \frac{Var(X_1)}{Var(\bar{X}_n)} = n$$

- The standard deviation of the sampling distribution is called the *standard error* (SE)
  - The SE of the mean estimator $\bar{X}_n$ is $\sigma/\sqrt{n}$

Unbiased estimators for expectation and variance

UNBIASED ESTIMATORS FOR EXPECTATION AND VARIANCE. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with finite expectation μ and finite variance σ^2. Then

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is an *unbiased estimator for* μ and

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

is an *unbiased estimator for* σ^2.

- Estimates: sample mean $\bar{x}_n$ and sample variance s_n^2
- $E[\bar{X}_n] = (E[X_1] + \ldots + E[X_n])/n = \mu$ and, by CLT, $Var(\bar{X}_n) \to 0$ for $n \to \infty$
- Why division by $n-1$ in S_n^2? *[Bessel's correction]*

Unbiasedness does not carry over (no functional invariance)

- $E[S_n^2] = \sigma^2$ implies $E[S_n] = \sigma$?
- Since $g(x) = x^2$ is convex, by Jensen's inequality:

$$\sigma^2 = E[S_n^2] = E[g(S_n)] > g(E[S_n]) = E[S_n]^2$$

  which implies $E[S_n] < \sigma$ *[Negative bias]*
- In general, if T unbiased for θ does not imply $g(T)$ unbiased for $g(\theta)$
  - But it holds for $g()$ linear transformation!

Estimators for the median and quantiles

- $T = Med(X_1, \ldots, X_n)$, for X_i with density function $f(x)$
- Let m be the true median, i.e., $F(m) = 0.5$: **[CLT for medians]**

$$\text{for } n \to \infty, T \sim \mathcal{N}(m, \frac{1}{4nf(m)^2})$$

and then for $n \to \infty$:

$$E[Med(X_1, \ldots, X_n)] = m$$

- $T = q_{X_1,\ldots,X_n}(p)$, for X_i with density function $f(x)$
- Let q_p be the true p-quantile, i.e., $F(q_p) = p$: **[CLT for quantiles]**

$$\text{for } n \to \infty, T \sim \mathcal{N}(q_p, \frac{p(1-p)}{nf(q_p)^2})$$

and then for $n \to \infty$:

$$E[q_{X_1,\ldots,X_n}(p)] = q_p$$

See R script

Example: estimating the probability of zero arrivals

- $X_1, \ldots, X_n$, for $n = 30$, observations:

$$X_i = \text{ number of arrivals (of a packet, of a call, etc.) in a minute}$$

- $X_i \sim Pois(\mu)$, where $p(k) = P(X = k) = \frac{\mu^k}{k!} e^{-\mu}$ $[E[X] = \mu]$
- We want to estimate $p_0 = p(0)$, probability of zero arrivals
- Frequentist-based estimator S:

$$S = \frac{|\{i \mid X_i = 0\}|}{n}$$

  - Takes values $0/30, 1/30, \ldots, 30/30$ ... may not exactly be p_0
  - $S = Y/n$ where $Y = \mathbb{1}_{X_1=0} + \ldots + \mathbb{1}_{X_n=0} \sim Bin(n, p_0)$
  - Hence, $E[S] = \frac{1}{n} E[Y] = \frac{n}{n} p_0 = p_0$ *[S is unbiased]*

Example: estimating the probability of zero arrivals

- Since $p_0 = p(0) = e^{-\mu}$, we devise a mean-based estimator T:

$$T = e^{-\bar{X}_n}$$

  - By Jensen's inequality:

$$E[T] = E[e^{-\bar{X}_n}] > e^{-E[\bar{X}_n]} = e^{-\mu} = p_0$$

    Hence T is biased!
  - However, T is asymptotically unbiased!
- Let's look at the variances

See R script

MSE: Mean Squared Error of an estimator

- What if one estimator is unbiased and the other is biased but with a smaller variance?

MSE

The Mean Squared Error of an estimator T for a parameter θ is defined as:

$$MSE(T) = E[(T - \theta)^2]$$

- An estimator T_1 performs better than T_2 if $MSE(T_1) < MSE(T_2)$
- Note that:

$$MSE(T) = E[(T - E[T] + E[T] - \theta)^2] =$$
$$= E[(T - E[T])^2] + (E[T] - \theta)^2 + 2E[T - E[T]](E[T] - \theta) = Var(T) + (E[T] - \theta)^2$$

- $E[T] - \theta$ is called the *bias* of the estimator
- Hence, $MSE = Var + Bias^2$
- A biased estimator with a small variance may be better than an unbiased one with a large variance!

See R script

Best estimators

Consistent estimator

An estimator T_n is a squared error consistent estimator if:

$$\lim_{n \to \infty} MSE(T_n) = 0$$

- Hence, for $n \to \infty$, both *Bias* and *Var* converge to 0
- $\bar{X}_n$ is a squared error consistent estimator of μ
- What if there is no consistent estimator or if there are more than once?

MVUE

An unbiased estimator T_n is a Minimum Variance Unbiased Estimators (MVUE) if:

$$Var(T_n) \leq Var(S_n)$$

for all unbiased estimators S_n.

- **Corollary.** $MSE(T_n) \leq MSE(S_n)$
- $\bar{X}_n$ is a MVUE of μ if $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

Example: number of German tanks

- Tanks' ID drawn at random without replacement from $1, \ldots, N$. Objective: estimate N.

Example: number of German tanks

- Let $x_1, \ldots, x_n$ be the observed ID's
- E.g., $61, 19, 56, 24, 16$ with $n = 5$
- They are realizations of $X_1, \ldots, X_n$ draws without replacement from $1, \ldots, N$
  - $X_1, \ldots, X_n$ is **not a random sample**, as they are not independent!
  - The marginal distribution is $X_i \sim U(1, N)$ [**prove it**, *or see Sect. 9.3 of [T]*]
- **Estimator based on the mean**
  - Since:

    $$E[\bar{X}_n] = E[X_i] = \frac{N+1}{2}$$

    we can define an estimator:

    $$T_1 = 2\bar{X}_n - 1$$

  - T_1 is unbiased:

    $$E[T_1] = 2E[\bar{X}_n] - 1 = N$$

  - E.g., $t_1 = 2(61 + 19 + 56 + 24 + 16)/5 - 1 = 69.4$

Estimators

- So far, estimators were derived from parameter definition through the plug-in method
- A general principle to derive estimators will be shown today
- Example

Table 21.1. Observed numbers of cycles up to pregnancy.

Number of cycles	1	2	3	4	5	6	7	8	9	10	11	12	>12
Smokers	29	16	17	4	3	9	4	5	1	1	1	3	7
Nonsmokers	198	107	55	38	18	22	7	9	5	3	6	6	12

- Assume that the data is generated from geometric distributions:

$$P(X_i = k) = (1-p)^{k-1} p$$

  where p is distinct for smokers and non smokers.
- What is an estimator for p? *[parametric inference]*
  - E.g., since $p = P(X_i = 1)$, we could use $S = \frac{|\{i \mid X_i=1\}|}{n}$, and show $E[S] = p$
  - $p = 29/100$ for smokers, and $p = 198/486 = 0.41$ for non-smokers
  - But we did not use all of the available data!

The maximum likelihood principle

The maximum likelihood principle

Given a dataset, choose the parameter(s) of interest in such a way that the data are most likely.

Table 21.1. Observed numbers of cycles up to pregnancy.

Number of cycles	1	2	3	4	5	6	7	8	9	10	11	12	>12
Smokers	29	16	17	4	3	9	4	5	1	1	1	3	7
Nonsmokers	198	107	55	38	18	22	7	9	5	3	6	6	12

- For $k = 1, \ldots, 12$, $P(X_i = k) = (1-p)^{k-1}p$. Moreover, $P(X_i > 12) = (1-p)^{12}$
- Since the X_i's are independent, we can write the probability of observing the smokers as:

$$L(p) = C \cdot P(X_i = 1)^{29} \cdot P(X_i = 2)^{16} \cdot \ldots \cdot P(X_i = 12)^{3} \cdot P(X_i > 12)^{7} = Cp^{93}(1-p)^{322}$$

  - C is the number of ways we can assign 29 ones, 16 twos, ..., 3 twelves, and 7 numbers larger than 12 to 100 smokers
- ML principle: choose $\hat{p} = arg\max_p L(p)$

Example

- ML principle: choose $\hat{p} = arg\max_p L(p) = arg\max_p Cp^{93}(1-p)^{322}$
- $L'(p) = C(93p^{92}(1-p)^{322} - 322p^{93}(1-p)^{321}) = Cp^{92}(1-p)^{321}(93-415p)$
- $L'(p) = 0$ for $p = 0$ or $p = 1$ or $p = 93/415 = 0.224$
- ML estimate is $arg\max_p L(p) = 0.224 < 0.41$ (estimate using S)
- Equivalent formulation for maximization:

$$arg\max_p L(p) = arg\max_p \log L(p)$$

- $\log L(p) = \log C + 93\log p + 322\log(1-p)$
- $\log' L(p) = \frac{93}{p} - \frac{322}{1-p}$
- $\log' L(p) = 0$ for $322p = 93(1-p)$, i.e., $p = 93/(322+93) = 0.224$

Likelihood and log-likelihood

Likelihood, log-likelihood, and MLE

Let $x_1, \ldots, x_n$ be a dataset, i.e., realizations of a random sample $X_1, \ldots, X_n$ where the density/p.m.f of X_i's is $f_\theta()$, parametric on θ. The likelihood function is:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

and the log-likelihood function is:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i)$$

Maximum likelihood estimates

The *maximum likelihood estimates* of θ is the value $t = \arg\max_\theta L(\theta) = \arg\max_\theta \ell(\theta)$. The statistics over the random sample:

$$\hat{\theta}_{ML} = \arg\max_\theta L(\theta) = \arg\max_\theta \ell(\theta)$$

is called the *maximum likelihood estimator* for θ.

Example: MLE of exponential distribution

- Random sample of $Exp(\lambda)$ $E[X] = {}^1/_\lambda$
- Since $f_\lambda(x) = \lambda e^{-\lambda x}$ for $x \geq 0$:

$$\ell(\lambda) = \sum_{i=1}^{n} (\log \lambda - \lambda x_i) = n \log \lambda - \lambda(x_1 + \ldots + x_n) = n(\log \lambda - \lambda \bar{x}_n)$$

- $\ell'(\lambda) = 0$ iff $n({}^1/_\lambda - \bar{x}_n) = 0$ iff $\lambda = {}^1/_{\bar{x}_n}$
- $\hat{\lambda}_{ML} = {}^1/_{\bar{X}_n}$ is the MLE of λ for a $Exp(\lambda)$-distributed random sample
- It is biased!: $E[\hat{\lambda}_{ML}] \geq 1/E[\bar{X}_n] = \lambda$ *[Jensen's inequality]*
- **Exercise at home**
  - show that $\bar{X}_n$ is an unbiased MLE of θ for a $Exp(1/\theta)$-distributed random sample

Example: MLE of normal distribution

- Random sample of $\mathcal{N}(\mu, \sigma^2)$
- MLE of $\theta = (\mu, \sigma^2)$ where $f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ *[we work on σ^2, not on σ]*

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

- Partial derivatives:

$$\frac{d}{d\mu} \ell(\mu, \sigma) = \frac{n}{\sigma^2} (\bar{x}_n - \mu) \qquad \frac{d}{d\sigma^2} \ell(\mu, \sigma) = \frac{1}{2\sigma^2} \left(\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - n \right)$$

- Partial derivatives at 0 for $\mu = \bar{x}_n$ and $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$
- MLE estimators $\hat{\mu}_{ML} = \bar{X}_n$ (*unbiased*) and $\hat{\sigma}^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$ *[biased]*

Loss functions (to be minimized)

- Negative log-likelihood (nLL)

$$nLL(\theta) = -\ell(\theta)$$

- How to compare estimators that use different numbers of parameters?
  - T_1 assuming a $Ber(p)$ vs T_2 assuming $Bin(n, p)$
  - Neural network with 10 nodes vs with 100 nodes
- Akaike information criterion (AIC), balances model fit against model simplicity

$$AIC(\theta) = 2|\theta| - 2\ell(\theta)$$

- Bayesian information criterion (BIC), stronger balances over model simplicity

$$BIC(\theta) = |\theta| \log n - 2\ell(\theta)$$

Cross entropy and nLL

- X, Y discrete random variables with p.m.f. p_X and p_Y:
- Cross entropy of X w.r.t. Y: $H(X;Y) = E_X[-\log p(Y)]$ *[see Lesson 4]*

$$H(X;Y) = -\sum_i p_X(a_i) \log p_Y(a_i)$$

- $H(X;Y)$ is the "information" or "uncertainty" or "loss" when using Y to encode X
- Negative log-likelihood:

$$nLL(\theta) = -\sum_{i=1}^{n} \log f_\theta(x_i) = H(X,Y)$$

  where $X \sim F_n$ (empirical distribution) and $Y \sim F_\theta$
- Minimizing nLL is equivalent to minimizing cross-entropy (or KL-divergence) between the empirical and the theoretical distributions!

Properties of MLE estimators

- MLE estimators can be biased, but under mild assumptions, they are asyntotically unbiased! *[Asyntotic unbiasedness]*

$$\lim_{n \to \infty} E[\hat{\theta}_{ML}] = \theta$$

- If $\hat{\theta}_{ML}$ is the MLE estimator of θ and $g()$ is an invertible function, then $g(\hat{\theta}_{ML})$ is the MLE estimator of $g(\theta)$ *[Invariance principle]*
  - E.g., MLE of σ for normal data is $\hat{\sigma}_{ML} = \sqrt{\hat{\sigma}^2_{ML}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2}$
  - but, $E[\hat{\theta}_{ML}] = \theta$ does **NOT** necessarily imply $E[g(\hat{\theta}_{ML})] = g(\theta)$
- Under mild assumptions, MLE estimators have asymptotically the smallest variance among unbiased estimators *[Asymptotic minimum variance]*