CS5487 Problem Set 2
Parameter Estimation - Maximum Likelihood

Antoni Chan
Department of Computer Science
City University of Hong Kong

Maximum Likelihood Estimation

Problem 2.1 The Poisson distribution and flying bombs

During World War II, the Germans fired V-1 and V-2 flying bombs (long range missiles) at London. Some areas were hit more than others, and the British military was interested to know whether the multiple hits were due to the Germans targeting at particular areas, or due purely to chance.

To analyze this problem, the British statistician R.D. Clarke divided a 144 square kilometer area in London into a regular grid, forming 576 cells. Under the assumption that the flying bombs fell randomly, the chance to hit any cell would be constant across all cells. Hence, the hit counts of the cells are i.i.d samples from a common random variable x.

A natural distribution for modeling the number of events (bomb hits) occurring within a fixed time period is the Poisson distribution, given by

$$p(x = k|\lambda) = \frac{1}{k!} e^{-\lambda} \lambda^k. \tag{2.1}$$

where $k \in \{0, 1, 2, 3, \cdots\}$ is a counting number. The parameter λ is the average number of events, and the mean and variance are the same $\mathbb{E}[x] = \text{var}(x) = \lambda$.

(a) Derive the maximum-likelihood estimate of λ, given a set of i.i.d. samples $\{k_1, \cdots, k_N\}$.

(b) Show that the ML estimator is unbiased, and the estimator variance is $\frac{\lambda}{N}$.

The following table lists the number of cells that were observed to have k hits (this is Clarke's actual data!).

number of hits (k)	0	1	2	3	4	5 and over
number of cells with k hits	229	211	93	35	7	1

(c) Using the above data, calculate the ML estimate $\hat{\lambda}$ for the Poisson distribution.

(d) Use the estimate $\hat{\lambda}$ to predict the expected number of cells with k hits, for $k \in \{0, 1, 2, 3, 4, 5+\}$. Compare the expected counts with the observed data. What conclusions can you make?

. . . .

Problem 2.2 MLE for the exponential density

Let x have an exponential density,

$$p(x|\theta) = \begin{cases} \theta e^{-\theta x}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

(a) Plot $p(x|\theta)$ versus x for $\theta = 1$. Plot $p(x|\theta)$ versus θ for $x = 2$.

(b) For a set of samples $\{x_1, \cdots x_N\}$ drawn i.i.d from $p(x|\theta)$, show that the maximum-likelihood estimate for θ is

$$\hat{\theta} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N} x_i}. \tag{2.3}$$

(c) Now suppose that we reparameterize our exponential density using $\lambda = 1/\theta$,

$$p(x|\lambda) = \begin{cases} \frac{1}{\lambda} e^{-\frac{x}{\lambda}}, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.4}$$

Derive the maximum-likelihood estimate of λ. How is $\hat{\lambda}$ related to $\hat{\theta}$?

. . . .

Problem 2.3 Invariance property of MLE

Prove the *invariance property* of maximum-likelihood estimators – let θ be the parameter of a probability distribution $p(x|\theta)$, and $\lambda = g(\theta)$ be a reparameterization $p(x|\lambda)$, where $g(\theta)$ is a differentiable function. Show that if $\hat{\theta}$ is the ML estimate of θ, then $\hat{\lambda} = g(\hat{\theta})$ is the ML estimate of λ.

. . . .

Problem 2.4 MLE for a Laplace distribution

→ more peaky than gaussian

Let x have a Laplace distribution (also called a double exponential)

$$p(x|\mu,\lambda) = \frac{1}{2\lambda} e^{-\frac{|x-\mu|}{\lambda}} = \begin{cases} \frac{1}{2\lambda} e^{\frac{x-\mu}{\lambda}}, & x < \mu \\ \frac{1}{2\lambda} e^{\frac{\mu-x}{\lambda}}, & x \geq \mu \end{cases} \tag{2.5}$$

(a) Derive the maximum-likelihood estimates for μ and λ, given a dataset of iid samples $\{x_1, \cdots, x_n\}$.

Hint: When finding the estimate for μ, the following property might be useful: for any 2 numbers $a, b \in \mathbb{R}$ with $a \leq b$, then $|a - \mu| + |b - \mu| \geq b - a$, with equality when $a \leq \mu \leq b$. Furthermore, without loss of generality, we can assume that the samples are ordered such that $x_1 \leq x_2 \leq \cdots \leq x_n$.

. . . .

Problem 2.5 MLE for a univariate Gaussian

Derive the ML estimate for a univariate Gaussian for samples $\{x_1, \cdots, x_N\}$. In particular,

(a) Show that the ML estimate of the mean is $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

(b) Show that the ML estimate of the variance is $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$.

....

Problem 2.6 MLE for a multivariate Gaussian

In this problem you will derive the ML estimate for a multivariate Gaussian. Given samples $\{x_1, \cdots, x_N\}$,

(a) Derive the ML estimate of the mean μ of a multivariate Gaussian.

(b) Derive the ML estimate of the covariance Σ.

$f(X) = tr(AX^{-1})$

$df = tr(dAX^{-1})$

$= tr(A \cdot dX^{-1})$

$= tr(-AX^{-1}dXX^{-1})$

$= -tr(X^{-1}AX^{-1} \cdot dX)$

$= -(X^{-T}A^TX^{-T}) \cdot dX$

$d\,tr(X^{-1})$

$= tr\lim_{t\to 0} \frac{(X+tM)^{-1} - X^{-1}}{t}$

$= tr(\lim (X+tM)^{-1} \frac{X - (X+tM)}{t} \cdot X^{-1})$

$= tr(-X^{-1} \cdot dX \cdot X^{-1})$

You may find the following vector and matrix derivatives helpful:

$a^Tb = tr(a^Tb)$

$= tr(ba^T)$

- $\frac{\partial}{\partial x} a^T x = a$, for vectors $x, a \in \mathbb{R}^d$.
- $\frac{\partial}{\partial x} x^T A x = Ax + A^T x$, for vector $x \in \mathbb{R}^d$ and matrix $A \in \mathbb{R}^{d\times d}$.
- $\frac{\partial}{\partial X} \log|X| = X^{-T}$, for a square matrix X.
- $\frac{\partial}{\partial X} \mathrm{tr}(AX^{-1}) = \frac{\partial}{\partial X}\mathrm{tr}(X^{-1}A) = -(X^{-T}A^TX^{-T})$, for matrices A, X.

$\frac{\partial}{\partial X_{jk}} = \sum_i A_{ij} \cdot B_{ki}$

$= [BA]_{kj}$

$\Rightarrow \frac{\partial}{\partial X}(tr(AXB)) = B^TA^T$

analogy: $\frac{\partial}{\partial x}\frac{1}{x} = \frac{-1}{x^2}$

Hint: remember Σ is symmetric!

$ABx = \lambda x$

$BABx = \lambda Bx$ ⇒ same eigenvalue

$BAy = \lambda y$ ⇒ hence same trace.

....

$tr(AXB) = \sum_i [AXB]_{ii}$

$= \sum_i\sum_j A_{ij}[XB]_{ji}$

$= \sum_i\sum_j A_{ij}(\sum_k X_{jk} \cdot B_{ki})$

$= \sum_i\sum_j\sum_k A_{ij}X_{jk}B_{ki}$

? Problem 2.7 MLE for a multinomial distribution

In this problem we will consider the ML estimate of the parameters of a multinomial distribution. Consider a discrete random variable x such that $p(x = j) = \pi_j, j \in \{1, \ldots, K\}$. Suppose we draw N independent observations from x and form a random vector $c = [c_1, \ldots, c_K]^T$ where c_j is the number of times that the observed value is j (i.e. c is the histogram of the sample of observations). Then, c has a multinomial distribution

$$p(c_1, \ldots, c_K) = \frac{N!}{\prod_{i=1}^{K} c_i!}\prod_{j=1}^{K}\pi_j^{c_j}. \tag{2.6}$$

(a) Derive the ML estimator for the parameters $\pi_j, j = 1, \ldots, K$. (Hint: notice that these parameters are probabilities, which makes this an optimization problem with a constraint. If you know about Lagrange multipliers feel free to use them. Otherwise, note that minimizing a function $f(a, b)$ under the constraint $a + b = 1$ is the same as minimizing the function $f(a, 1 - a)$).

(b) Is the estimator derived in (a) unbiased? What is its variance? Is this a good estimator? Why?

....

Problem 2.8 Least-squares regression and MLE

In this problem we will consider the issue of linear regression and the connections between maximum likelihood and least squares solutions. Consider the polynomial function of $x \in \mathbb{R}$,

$$f(x,\theta) = \sum_{k=0}^{K} x^k \theta_k = \underset{\underset{\text{features}}{\downarrow}}{\phi(x)}^T \theta, \rightarrow \text{parameters} \tag{2.7}$$

where we define the feature transformation $\phi(x)$ and the parameter vector θ (both of dimension $D = K + 1$) as

$$\phi(x) = \left[1, x, x^2, \cdots, x^K\right]^T \in \mathbb{R}^D, \qquad \theta = \left[\theta_0, \cdots, \theta_K\right]^T \in \mathbb{R}^D. \tag{2.8}$$

Given an input x, instead of observing the actual function value $f(x,\theta)$, we observe a noisy version y,

$$y = f(x,\theta) + \epsilon, \tag{2.9}$$

where ϵ is an Gaussian random variable of zero mean and variance σ^2. Our goal is to obtain the best estimate of the function given iid samples $\mathcal{D} = \{(x_1,y_1),\ldots,(x_n,y_n)\}$.

(a) Formulate the problem as one of least squares, i.e define

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \Phi = \left[\phi(x_1), \cdots, \phi(x_n)\right] = \begin{bmatrix} 1 & \ldots & 1 \\ x_1^1 & & x_n^1 \\ \vdots & & \vdots \\ x_1^K & \ldots & x_n^K \end{bmatrix} \tag{2.10}$$

and find the value of θ that minimizes the sum-squared-error,

transpose of each other

$$\sum_{i=1}^{n} (y_i - \phi(x_i)^T\theta)^2 = \left\|y - \Phi^T\theta\right\|^2 = (y-\Phi^T\theta)^T(y-\Phi\theta) \tag{2.11}$$

$$= y^Ty \underbrace{-(\Phi^T\theta)^Ty - y^T\Phi\theta}_{\text{scalar}} + \theta^T\Phi\Phi^T\theta$$

(b) Formulate the problem as one of ML estimation, i.e. write down the likelihood function $p(y|x,\theta)$, and compute the ML estimate, i.e. the value of θ that maximizes $p(y_1,\cdots,y_n|x_1,\cdots,x_n,\theta)$. Show that this is equivalent to (a).

$$= \underbrace{y^Ty}_{x} - \underbrace{2y^T\Phi^T\theta}_{a^T\theta} + \underbrace{\theta^T\Phi\Phi^T\theta}_{\theta^TA\theta}$$

Hint: the vector derivatives listed in Problem 2.6 might helpful.

$$\frac{\partial}{\partial\theta}(\,) = -2\Phi y + 2\Phi\Phi^T\theta = 0$$

$$\hat{\theta} = (\Phi\Phi^T)^{-1}\Phi y$$

why don't expand inverse?

Φ $(k+1)\times n$

not necessary square

$$\sum \log p(y_i|x_i,\theta) \quad \ldots\ldots\ldots$$

$$= \sum \log N(y_i|f(x_i), a^2) \Rightarrow \text{least square formulation}$$

Problem 2.9 Weighted least-squares and MLE

The advantage of the statistical formulation of least-squares regression is that it makes the assumptions explicit. We will now challenge some of these assumptions.

Assume that instead of a fixed variance σ^2 we now have a variance that depends on the sample point, i.e.

$$y_i = f(x_i,\theta) + \epsilon_i, \tag{2.12}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. This means that our sample is independent but no longer identically distributed. It also means that we have different degrees of confidence in the different measurements (y_i, x_i).

Extra information: eg. low variance / high variance region.

(a) Formulate the problem as one of ML estimation, and compute the ML estimate, i.e., the value of θ that maximizes $p(y_1, \cdots, y_n | x_1, \cdots, x_n)$.

different weight when data point is biased ☆!

(b) Consider the weighted least squares problem where the goal is to minimize the weighted sum-squared error

$$\sum_{i=1}^{n} w_i (y_i - \phi(x_i)^T \theta)^2 = (y - \Phi^T \theta)^T W (y - \Phi^T \theta) \tag{2.13}$$

where W is a diagonal matrix with diagonal entries w_i. Compute the optimal θ for this situation. What is the equivalent maximum likelihood problem? Rewrite the model (2.9), making explicit all the assumptions that lead to the new problem. What is the statistical interpretation of W and w_i? What is an alternative interpretation of the weight w_i (e.g., if w_i is restricted to be a positive integer)?

① $\frac{1}{\sigma^2}$

② eg. $w_i = 3 \Rightarrow$ treat it as multiple samples.

. . . .

Problem 2.10 Robust regression and MLE

The L_2 norm is known to be prone to large estimation error if there are *outliers* in the training sample. These are training examples (y_i, x_i) for which, due to measurement errors or other extraneous causes, the residual error $|y_i - \phi(x_i)^T \theta|$ is much larger than for the remaining examples (the *inliers*). In fact, it is known that a single outlier can completely derail the least squares solution, a highly undesirable behavior. It is also well known that other norms lead to much more robust estimators. One of such distance metrics is the L_1-norm

$$L_1 = \sum_{i=1}^{n} |y_i - \phi(x_i)^T \theta| = \left\| y - \Phi^T \theta \right\|_1 . \tag{2.14}$$

laplacian distribution

(a) In the maximum likelihood framework, what is the statistical assumption that leads to the L_1 norm? Once again, rewrite the model (2.9), making explicit all the assumptions that lead to the new problem. Can you justify why this alternative formulation is more robust? In particular, provide a justification for **i)** why the L_1 norm is more robust to outliers, and **ii)** why the associated statistical model copes better with them.

Unlike the previous least-squares formulations, the minimization of (2.14) does not have a closed-form solution. Instead, we must solve it numerically when we are given the data. To facilitate this, we need to cast our L1 minimization problem into the standard form of a *linear program* (an optimization problem with a linear objective function and linear inequality constraints). This will allow us to use a standard optimization toolbox to perform the minimization numerically (e.g., `linprog` in Matlab). Our original optimization problem is

$$\min_{\theta} \sum_{i=1}^{n} |y_i - \phi(x_i)^T \theta|. \tag{2.15}$$

(b) Introduce an auxiliary variable $t \in \mathbb{R}^n$. Verify that the optimization problem in (2.15) is equivalent to

$$\begin{aligned} &\min_{\theta,t} \sum_{i=1}^{n} t_i \\ &\text{s.t. } |y_i - \phi(x_i)^T\theta| \le t_i \end{aligned} \tag{2.16}$$

if $t_i > |y_i - \phi(x_i)^T\theta|$ we can always decrease t_i $\downarrow$ so that.

What is the role of each auxiliary variable t_i? use t_i to represent $|y_i - \phi(x_i)^T\theta|$

(c) Define $\mathbf{x} = \begin{bmatrix}\theta \\ t\end{bmatrix} \in \mathbb{R}^{D+n}$. Show that (2.16) can be turned into a standard *linear program*,

$$\begin{aligned} &\min_{\mathbf{x}} \mathbf{f}^T\mathbf{x} \\ &\text{s.t. } \mathbf{A}\mathbf{x} \le \mathbf{b} \end{aligned} \tag{2.17}$$

$t_i \geq y_i - \phi(x_i)^T\theta$

$t_i \geq -(y_i - \phi(x_i)^T\theta)$

where

$$\mathbf{f} = \begin{bmatrix}0_D \\ 1_n\end{bmatrix} \in \mathbb{R}^{D+n}, \qquad \mathbf{A} = \begin{bmatrix}-\Phi^T & -I_n \\ \Phi^T & -I_n\end{bmatrix} \in \mathbb{R}^{2n\times(D+n)}, \qquad \mathbf{b} = \begin{bmatrix}-y \\ y\end{bmatrix} \in \mathbb{R}^{2n}, \tag{2.18}$$

and 0_n is the vector of n zeros, 1_n is the vector of n ones, and I_n is the $n \times n$ identity matrix. We now have a form of our original L1 optimization problem that can be solved numerically using `linprog` in the MATLAB optimization toolbox. Note: we could probably get faster performance if we develop a custom solver for our original problem in (2.15).

. . . .

Estimator Bias and Variance

Problem 2.11 Simple estimator

Let $\mathcal{D} = \{x_1, \cdots, x_N\}$ be a set of i.i.d samples from a Gaussian r.v. x. Suppose we define an estimator of the mean as the first point in the set,

$$\hat{\mu} = x_1. \tag{2.19}$$

(a) Show that the estimator $\hat{\mu}$ is unbiased.

(b) Derive the variance of the estimator.

(c) Why is this an undesirable estimator?

. . . .

Problem 2.12 Bias of the ML estimator for variance of a Gaussian

The ML estimator of the variance of a Gaussian is

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2, \tag{2.20}$$

where $\hat{\mu}$ is the ML estimator of the mean.

(a) Suppose that the true distribution has mean μ and variance σ^2, show that

$$\mathbb{E}_{x_1,\cdots,x_N}[\hat{\sigma}^2] = \left(1 - \frac{1}{N}\right)\sigma^2, \tag{2.21}$$

and hence $\hat{\sigma}^2$ is a biased estimator.

(b) Using (a), propose an unbiased estimator of the variance.

....

Problem 2.13 Bias-Variance Tradeoff

In general, there is always a tradeoff between bias and variance. We can reduce the bias only by increasing the variance, and vice versa. In this problem, we will consider an example of this.

Suppose we want a faster decay of the variance of the mean estimator for a Gaussian. We could scale the original mean estimator to form a new estimator $\tilde{\mu}$,

$$\tilde{\mu} = \frac{\alpha}{N}\sum_{i=1}^{n} X_i. \tag{2.22}$$

(a) Show that the mean value of the estimator is $\mathbb{E}[\tilde{\mu}] = \alpha\mu$, and hence the bias and variance are

$$\text{Bias}(\tilde{\mu}) = (1-\alpha)\mu, \qquad \text{Var}(\tilde{\mu}) = \frac{\alpha^2}{N}\sigma^2. \tag{2.23}$$

Hence, selecting $\alpha < 1$ will decrease the variance but also causes $\tilde{\mu}$ to be biased!

....

Problem 2.14 Weak law of large numbers

The weak law of large numbers states that the sample mean of a large number of i.i.d. random variables is very close to the true mean, with high probability. Let's prove it!

Let $\{x_1, \cdots, x_n\}$ be a set of n i.i.d. random variables with mean μ_x and variance σ_x^2. Define the sample mean as,

$$M_n = \frac{1}{n}(x_1 + \cdots + x_n) \tag{2.24}$$

(a) Show that the mean and variance of the sample mean are

$$\mathbb{E}[M_n] = \mu_x, \qquad \text{var}(M_n) = \frac{\sigma_x^2}{n}. \tag{2.25}$$

(b) The *Chebyshev inequality* states that for a random variable x with mean μ and variance σ^2,

$$P(|x-\mu| \geq c) \leq \frac{\sigma^2}{c^2}, \qquad \text{for all } c > 0. \tag{2.26}$$

In other words, if a r.v. has small variance, then the probability it will take a value far from the mean is small.

Use the Chebyshev inequality and your result above to show that, for every $\epsilon > 0$,

$$P(|M_n - \mu_x| \geq \epsilon) \to 0, \qquad \text{as } n \to \infty. \tag{2.27}$$

In other words, there is high probability that M_n will fall within the interval $[\mu_x - \epsilon, \mu_x + \epsilon]$, for large n. That is, the bulk of the distribution of M_n is concentrated around μ_x, and converges to μ_x as $n \to \infty$.

....

Proof of the Chebyshev inequality (continuous case):

Given: X a real continuous random variables with $E(X) = \mu$, $V(X) = \sigma^2$, real number $\epsilon > 0$.
To show: $P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$.

Then

$$\begin{aligned}
\sigma^2 &= V(X) \\
&= \int_{-\infty}^{\infty} (t-\mu)^2 f_X(t)\,dt \\
&\geq \int_{-\infty}^{\mu-\epsilon} (t-\mu)^2 f_X(t)\,dt + \int_{\mu+\epsilon}^{\infty} (t-\mu)^2 f_X(t)\,dt,
\end{aligned}$$

where the last line is by restricting the region over which we integrate a positive function. Then this is

$$\geq \int_{-\infty}^{\mu-\epsilon} \epsilon^2 f_X(t)\,dt + \int_{\mu+\epsilon}^{\infty} \epsilon^2 f_X(t)\,dt,$$

since $t \leq \mu - \epsilon \implies \epsilon \leq |t - \mu| \implies \epsilon^2 \leq (t-\mu)^2$. But we rearrange and use the definition of the density function to get

$$\begin{aligned}
&= \epsilon^2 \left(\int_{-\infty}^{\mu-\epsilon} f_X(t)\,dt + \int_{\mu+\epsilon}^{\infty} f_X(t)\,dt \right) \\
&= \epsilon^2 P(X \leq \mu - \epsilon \text{ or } X \geq \mu + \epsilon) \\
&= \epsilon^2 P(|X - \mu| \geq \epsilon).
\end{aligned}$$

Thus,

$$\sigma^2 \geq \epsilon^2 P(|X - \mu| \geq \epsilon),$$

and dividing through by ϵ^2 gives the desired. □