1 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a statistical model given observations, by finding the parameters that maximize the likelihood of the observations. Concretely, given observations $y_1, y_2, \ldots, y_n$ distributed according to $p_\theta(y_1, y_2, \ldots, y_n)$ (here p_θ can be a probability mass function for discrete observations or a density for continuous observations), the likelihood function is defined as $L(\theta) = p_\theta(y_1, y_2, \ldots, y_n)$ and the MLE is

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta L(\theta).$$

We often make the assumption that the observations are *independent and identically distributed* or iid, in which case $p_\theta(y_1, y_2, \ldots, y_n) = p_\theta(y_1) \cdot p_\theta(y_2) \cdot \cdots \cdot p_\theta(y_n)$.

(a) Your friendly TA recommends maximizing the log-likelihood $\ell(\theta) = \log L(\theta)$ instead of $L(\theta)$. **Why does this yield the same solution $\hat{\theta}_{\text{MLE}}$? Why is it easier to solve the optimization problem for $\ell(\theta)$ in the iid case? Write down both $L(\theta)$ and $\ell(\theta)$ for the Gaussian $f_\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}$ with $\theta = (\mu, \sigma)$.**

(b) **What is $\int p_\theta(y_1, y_2, \dots, y_n)\, dy_1 \cdots dy_n$? Can we say anything about $\int p_\theta(y_1, y_2, \dots, y_n) d\theta$?**

(c) Let's practice performing MLE with a Poisson distribution, with a PMF given as: $f_\lambda(y) = \frac{\lambda^y e^{-\lambda}}{y!}$. Let $Y_1, Y_2, \dots, Y_n$ be a set of independent and identically distributed random variables with Poisson distribution with parameter λ.

Find the joint distribution of $Y_1, Y_2, \dots, Y_n$. Find the maximum likelihood estimate of λ as a function of observations $y_1, y_2, \dots, y_n$.

2 Linear Regression from MLE

In this problem, we will use maximum likelihood to motivate optimizing certain types of loss functions when performing linear regression.

To review, the goal of linear regression is to learn the parameters of a model $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ so that we can predict a label y that corresponds to input features $\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$. We are given a dataset of n input feature vectors and output labels $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$. (Note: we will ignore the bias term in this problem for simplicity.)

When performing linear regression on a dataset such as this one, we assume that the labels y_i are noisy: $y_i = \mathbf{w}^{*\top}\mathbf{x}_i + \varepsilon_i$, where $\mathbf{w}^*$ is the true linear model parameter we are trying to estimate.

There are many possible assumptions we could make about the noise ε_i, and in this problem, we will explore the implications of assuming noise sampled from certain probability distributions.

(a) **Assuming Gaussian noise variables $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, write the likelihood function $\mathcal{L}(\mathbf{w})$ of the dataset given a set of parameters $\mathbf{w}$.**

$$\mathcal{L}(\mathbf{w}) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n, y_1, \ldots, y_n; \mathbf{w}) \tag{1}$$

The PDF of the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ is given as:

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(z-\mu)^2}{\sigma^2}\right\} \tag{2}$$

Hint: What is the probability distribution that gives $p(y_i|\mathbf{x}_i; \mathbf{w})$?

Hint: Remember that we can exclude all terms in our likelihood that don't depend on the parameter $\mathbf{w}$.

(b) **Compute the log likelihood $\ell(\mathbf{w}) = \log \mathcal{L}(\mathbf{w})$ and simplify to matrix notation to show that the maximum likelihood objective is equivalent to:**

$$\hat{\mathbf{w}}_{MLE} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 \tag{3}$$

$\mathbf{X}$ is an $n \times d$ matrix where each row represents a data point, and $\mathbf{y}$ is a n-dimensional vector of labels for all the data points.

(c) Now, we will assume a different noise distribution (still zero-mean and independently sampled). **Given noise variables sampled from a Laplace distribution $\varepsilon_i \overset{i.i.d.}{\sim} \text{Laplace}(0, b)$, what is the likelihood function $\mathcal{L}(\mathbf{w})$?**

The PDF of the Laplace distribution Laplace(μ, b) is given as:

$$f(z) = \frac{1}{2b} \exp\left\{-\frac{|z - \mu|}{b}\right\} \tag{4}$$

(d) **Compute the log likelihood $\ell(\mathbf{w})$ and simplify to matrix notation to show that the maximum likelihood objective is equivalent to:**

$$\hat{\mathbf{w}}_{MLE} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_1 \tag{5}$$

(e) The ℓ-1 (least absolute deviations) objective is known to be more robust in the presence of outlier labels that fall far from the true distribution. **Why, from the viewpoint of maximum likelihood and probability densities, does it make sense that the ℓ-1 objective is less sensitive to outliers than the ℓ-2 (least squares) objective?**

3 Maximum Likelihood Estimation for Reliability Testing

Suppose we are reliability testing n units taken randomly from a population of identical appliances. We want to estimate the mean failure time of the population. We assume the failure times come from an exponential distribution with parameter $\lambda > 0$, whose probability density function is $f(t) = \lambda e^{-\lambda t}$ on the domain $t \geq 0$.

(a) In an ideal (but impractical) scenario, we run the units until they all fail.

The failure times $T_1, T_2, \ldots, T_n$ for units $1, 2, \ldots, n$ are observed to be $t_1, t_2, \ldots, t_n$.

Formulate the likelihood function $\mathcal{L}(\lambda; t_1, \ldots, t_n)$ for our data. Then, find the maximum likelihood estimate $\hat{\lambda}$ for the distribution's parameter.

(Remember that it's equivalent, and usually easier, to optimize the log-likelihood.)

(b) In a more realistic scenario, we run the units for a fixed time h. The failure time for $T_1, T_2, \ldots, T_r$ are observed to be $t_1, t_2, \ldots, t_r$, where $0 \leq r \leq n$. The remaining $n - r$ units survive the entire time h without failing. Let's find the maximum likelihood estimate $\hat{\lambda}$ for our model distribution parameters!

(a) **What is the probability that a unit will not fail during time h?**

(b) **Write the new likelihood function $\mathcal{L}(\lambda; h, n, r, t_1, \ldots, t_r)$ and optimize to find the MLE estimate $\hat{\lambda}$.**

(c) **Compare the two MLE estimates, and explain the difference with a physical interpretation.**