Lecture on Maximum Likelihood Estimation

Lecturer: Your Name

Machine Learning Course for Research Students

May 21, 2025

Outline

Introduction

Motivation and Problem Statement

- Introduction to Maximum Likelihood Estimation (MLE)
- Motivation and Problem Statement
- Intuitive Approach
- Detailed Mathematical Derivations
- Examples and Applications
- Summary and Q&A

Introduction: What is Maximum Likelihood Estimation?

- MLE is a method for estimating the parameters of a statistical model.
- It selects the parameter values that maximize the likelihood of the observed data.
- Widely used in statistics and machine learning for fitting models.

Motivation for MLE

- In many real-world problems, we need to infer the underlying parameters that best explain the data.
- MLE provides a principled way to derive parameter estimates directly from the data.
- It is especially effective when the model is correctly specified.

The Problem: Parameter Estimation

Given a set of independent and identically distributed (i.i.d.) data points $\{x_1, x_2, \ldots, x_n\}$ drawn from a probability density function $f(x; \theta)$, our goal is to

- Find the parameter θ that maximizes the likelihood function:

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

- Alternatively, maximize the log-likelihood:

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

Intuitive Approach to MLE

- Interpret the likelihood as a measure of how probable the observed data is, given the parameters.
- The optimal parameter is the one under which the observed data is most 'expected'.
- Think of tuning the parameter until the model's prediction aligns with the real-world data.

Deriving the MLE: Detailed Math I

Consider a simple case where $x_i \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2. The likelihood function is:

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Taking the log-likelihood, we obtain:

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

Differentiate with respect to μ to find the maximum.

Deriving the MLE: Detailed Math II

Differentiate the log-likelihood with respect to μ:

$$\frac{d}{d\mu}\ell(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

Solving for μ yields:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

This is the well-known sample mean, the maximum likelihood estimator for the mean of a normal distribution.

MLE for Different Distributions

Beyond the normal distribution, MLE can be applied to various probability models.

- For a Bernoulli model: estimating the probability parameter p.
- For an Exponential distribution: determining the rate parameter λ.
- Other distributions including Poisson, Binomial, and Gamma.

These examples illustrate the versatility of the MLE method in statistical inference.

Case Study: Bernoulli Distribution

Consider $x_i \sim \text{Bernoulli}(p)$ with $x_i \in \{0, 1\}$. The likelihood function is given by:

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

Taking the log-likelihood, we have:

$$\ell(p) = \sum_{i=1}^{n} \Big[x_i \log p + (1 - x_i) \log(1-p) \Big].$$

Differentiating with respect to p and equating to zero gives the MLE:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Case Study: Exponential Distribution

Suppose $x_i \sim \text{Exponential}(\lambda)$ where $x_i \geq 0$. The likelihood function is:

$$L(\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right).$$

The corresponding log-likelihood is:

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

Differentiating and setting the derivative to zero, we obtain:

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \quad \Rightarrow \quad \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}.$$

Computational Considerations

- In many cases, solving for the MLE analytically is challenging.
- Numerical methods (e.g., Newton-Raphson, Expectation-Maximization) are often employed.
- Optimization algorithms help in navigating complex likelihood landscapes.
- Software tools and programming libraries play a crucial role in practical implementations.

Understanding these computational aspects is key for applying MLE to real-world problems.

MLE in Machine Learning

MLE forms the backbone for various machine learning techniques, including:

- Logistic Regression: Estimating the parameters in classification problems.
- Gaussian Mixture Models: Parameter estimation in clustering.
- Hidden Markov Models: Inferring the transition and emission probabilities.

These scenarios demonstrate the broad applicability of MLE in data-driven modeling.

Summary of MLE Concepts

- MLE provides a systematic approach for parameter estimation.
- It is applicable across a wide range of probability distributions.
- Both analytical and numerical techniques are valuable for solving MLE problems.
- Its principles underpin many modern machine learning algorithms.

This summary consolidates the key ideas discussed and prepares us for further exploration in subsequent slides.

Regularity Conditions for MLE

- For the MLE to have desirable properties, certain regularity conditions must be satisfied.
- Conditions include: **smoothness** of the likelihood function, **identifiability** of the model parameters, and the existence of required **moments**.
- These ensure that the log-likelihood function behaves well and that derivative-based methods yield valid estimators.

Properties of Maximum Likelihood Estimators

MLEs possess several important asymptotic properties:

1. **Consistency**: As the sample size increases, the MLE converges in probability to the true parameter value.
2. **Efficiency**: Under regularity conditions, the MLE achieves the minimum possible variance (asymptotically).
3. **Asymptotic Normality**: The distribution of the MLE (properly normalized) is approximately normal for large samples.

Fisher Information and the Cramér-Rao Lower Bound

- **Fisher Information** measures the amount of information that an observable variable carries about an unknown parameter.
- For a likelihood function $L(\theta)$, the Fisher information is given by: $I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right]$.
- **Cramér-Rao Bound**: Provides a lower bound on the variance of any unbiased estimator. That is, $\mathrm{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$.

Case Study: Poisson Distribution

Consider data $x_i \sim \text{Poisson}(\lambda)$ with $x_i \in \{0, 1, 2, ...\}$. The likelihood function is:

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} \exp(-\lambda)}{x_i!}$$

Taking the log-likelihood, we get:

$$\ell(\lambda) = \sum_{i=1}^{n} \left[x_i \log \lambda - \lambda - \log(x_i!) \right]$$

Differentiating with respect to λ and setting it to zero yields:

$$\frac{d}{d\lambda}\ell(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0 \quad \Rightarrow \quad \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Case Study: Gamma Distribution

Consider observations x_i drawn from a Gamma distribution with shape parameter k and scale parameter θ, so that

$$f(x; k, \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} \exp\left(-\frac{x}{\theta}\right), \quad x > 0.$$

The log-likelihood for a given sample $\{x_1, \ldots, x_n\}$ becomes

$$\ell(k, \theta) = -n \log \Gamma(k) - nk \log \theta + (k-1) \sum_{i=1}^{n} \log x_i - \frac{1}{\theta} \sum_{i=1}^{n} x_i.$$

Estimating k and θ typically requires numerical optimization techniques.

Numerical Optimization: Newton-Raphson Method

When closed-form solutions are intractable, iterative methods can be applied. **Newton-Raphson Update:**

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\ell'(\theta^{(t)})}{\ell''(\theta^{(t)})},$$

where $\ell'(\theta)$ and $\ell''(\theta)$ are the first and second c

This method is often used to compute MLE where derivatives of the log-likelihood can be analytically derived.

EM Algorithm in MLE

For models with latent variables or incomplete data, the Expectation-Maximization (EM) algorithm offers a powerful approach.

- **E-step**: Estimate the expected value of the log-likelihood with respect to the latent variables.
- **M-step**: Maximize this expectation to update parameter estimates.

This iterative process guarantees a non-decreasing likelihood and is central in mixture models and hidden Markov models.

Robustness and Limitations of MLE

- While MLE is asymptotically efficient, its performance can degrade in small samples or under model misspecification.
- Sensitivity to outliers is a common issue, motivating the use of robust statistics as alternatives.
- Regularization methods may be incorporated to stabilize estimates in complex models.

Recap: Theoretical Properties of MLE

1. **Consistency**: Estimates converge to the true parameter as $n \to \infty$.
2. **Efficiency**: Achieves the lowest possible variance among unbiased estimators (asymptotically).
3. **Asymptotic Normality**: For large samples, the estimator is normally distributed around the true parameter.

These properties provide the statistical foundation for the reliability of MLE.

Application: Logistic Regression

In logistic regression, the probability of a binary outcome $y_i \in \{0, 1\}$ given a vector of predictors $\mathbf{x}_i$ is modeled as

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x}_i)}.$$

The log-likelihood is written as

$$\ell(\mathbf{w}) = \sum_{i=1}^{n} \Big[y_i \log P(y_i = 1|\mathbf{x}_i) + (1 - y_i) \log(1 - P(y_i = 1|\mathbf{x}_i)) \Big].$$

Maximizing this log-likelihood with respect to $\mathbf{w}$ gives the MLE for the model parameters.

MLE vs. Bayesian Estimation

- **MLE:** Provides point estimates by maximizing the likelihood of observed data.
- **Bayesian:** Incorporates prior beliefs through a prior distribution and obtains a posterior distribution.

The key difference lies in the treatment of uncertainty and the incorporation of prior knowledge in Bayesian methods.

Practical Implementation of MLE

- Programming libraries such as R, Python (with SciPy, Statsmodels), and MATLAB provide built-in functions for MLE.
- Convergence criteria and initialization can significantly affect numerical optimization outcomes.
- It is important to conduct diagnostic checks to verify the validity of the estimated parameters.

Further Studies and Open Questions

- How do MLE estimators behave under severe model misspecification?
- What are the best practices for integrating regularization with MLE in high-dimensional settings?
- Recommended readings:
  - "The Elements of Statistical Learning" by Hastie, Tibshirani, and Friedman.
  - "Pattern Recognition and Machine Learning" by Christopher Bishop.
  - Relevant journal articles on robust and regularized MLE.

Lecture Summary

- Reviewed the fundamental concept and intuition behind MLE.
- Derived MLE for various distributions including Normal, Bernoulli, Exponential, Poisson, and Gamma.
- Discussed numerical optimization methods and their role in complex models.
- Compared MLE with alternative approaches such as Bayesian estimation.
- Highlighted practical aspects of implementing MLE in real-world scenarios.

QA Session

Any Questions?

Feel free to ask for clarifications, additional examples, or any other points of discussion regarding MLE.

Closing Remarks and References

- Thank you for your attention.
- For further queries, contact: `your.email@institution.edu`
- References:

  - Casella, G., Berger, R. L. (2002). Statistical Inference.
  - Wasserman, L. (2004). All of Statistics: A Concise Course in Statistical Inference.

Goodbye!