Machine Learning

Maximum Likelihood Estimation

Indian Institute of Information Technology
Sri City, Chittoor

Today's Agenda

- Introduction to Maximum Likelihood Estimation
- Maximum likelihood Estimation (MLE) method
- General Principle of MLE
- General Principle of MLE: Example
- MLE - The Gaussian Case: Unknown μ
- MLE - The Gaussian Case: Unknown μ and Σ
- Additional Resources

Introduction to Maximum Likelihood Estimation

- We can design an optimal classifier if we knew the prior probabilities $P(\omega_i)$ and the class-conditional densities $p(\mathbf{x}/\omega_i)$.

- Unfortunately, in ML applications we rarely if ever have this kind of complete knowledge about the probabilistic structure of the problem.

- In a typical case we merely have some vague, general knowledge about the situation, together with a number of *design samples* or *training data.*

- If we know the number of parameters in advance and our general knowledge about the problem permits us to parameterize the conditional densities, then the severity of these problems can be reduced significantly.

Introduction to Maximum Likelihood Estimation

- Suppose, for example, that we can reasonably assume that $p(x|\omega_i)$ is a normal density with mean μ_i and covariance matrix Σ_i, although we do not know the exact values of these quantities.

- This knowledge simplifies the problem from one of estimating an unknown function $p(x|\omega_i)$ to one of estimating the parameters μ_i and Σ_i.

- We shall consider two common and reasonable procedures:
  - *Maximum likelihood* estimation (MLE)
  - *Bayesian* estimation.

Introduction to Maximum Likelihood Estimation

- Maximum likelihood method view the parameters as quantities whose values are fixed but unknown.
  - The best estimate of their value is defined to be the one that maximizes the probability of obtaining the samples actually observed.

- In contrast, Bayesian methods view the parameters as random variables having some known a priori distribution.
  - Observation of the samples converts this to a posterior density, thereby revising our opinion about the true values of the parameters

Maximum likelihood Estimation (MLE) method

- Maximum likelihood estimation is a method that determines values for the parameters of a model.

- Maximum Likelihood Estimation involves treating the problem as an optimization or search problem, where we seek a set of parameters that results in the best fit for the joint probability of the data sample (X).

- In Maximum Likelihood Estimation, we wish to maximize the probability of observing the data from the joint probability distribution given a specific probability distribution and its parameters

Maximum likelihood Estimation (MLE) method

- Maximum likelihood estimation is a method that determines values for the parameters of a model.
- Maximum Likelihood Estimation involves treating the problem as an optimization or search problem, where we seek a set of parameters that results in the best fit for the joint probability of the data sample (X).
- In Maximum Likelihood Estimation, we wish to maximize the probability of observing the data from the joint probability distribution given a specific probability distribution and its parameters.

- Why MLE Methods?
  - Always have good convergence properties as the number of training samples increases.
  - Simpler than alternate methods, such as Bayesian techniques

General Principle of MLE

- Suppose that we separate a collection of samples according to class, so that we have c sets, D_1, ...,D_c, with the samples in D_j having been drawn independently according to the probability law $p(x|\omega_j)$.

- We say such samples are i.i.d. — independent identically distributed random variables.

- We assume that $p(x|\omega_j)$ has a known parametric form, and is therefore determined uniquely by the value of a parameter vector θ_j.

- Our problem is to use the information provided by the training samples to obtain good estimates for the unknown parameter vectors θ_1, ..., θ_c associated with each category.

General Principle of MLE: example

- Assume that we might have $p(x|\omega_j) \sim N(\mu_j, \Sigma_j)$, where θ_j consists of the components of μ_j and Σ_j .

- Our problem is to use the information provided by the training samples (x) to obtain good estimates for the unknown parameter vector θ_j, that is, estimating μ_j and Σ_j for each θj.

- To show the dependence of $p(x|\omega_j)$ on θ_j explicitly, we write $p(x|\omega_j)$ as $p(x|\omega_j, \theta_j)$.

General Principle of MLE: example

- Suppose that D contains n samples, x_1, ..., x_n. Then, since the samples were drawn independently, we have

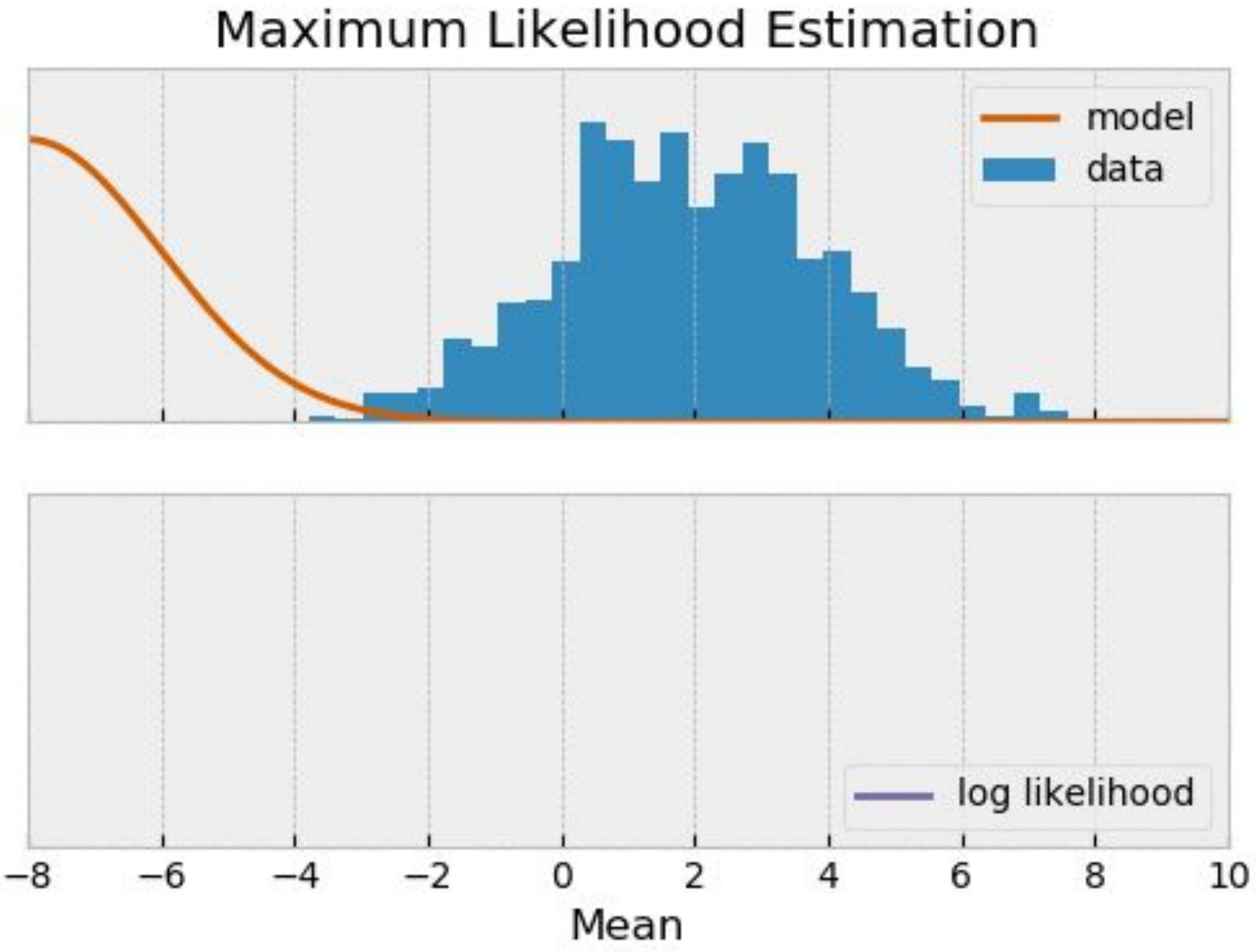
$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}). \qquad \text{Eq. 1}$$

i.e, p(D|θ) is called the likelihood of θ with respect to the set of samples.

- The maximum likelihood estimate of θ is, by definition, the value $\hat{\theta}$ that maximizes p(D|θ).

- Intuitively, this estimate corresponds to the value of θ that in some sense best agrees with the actually observed training

MLE: example

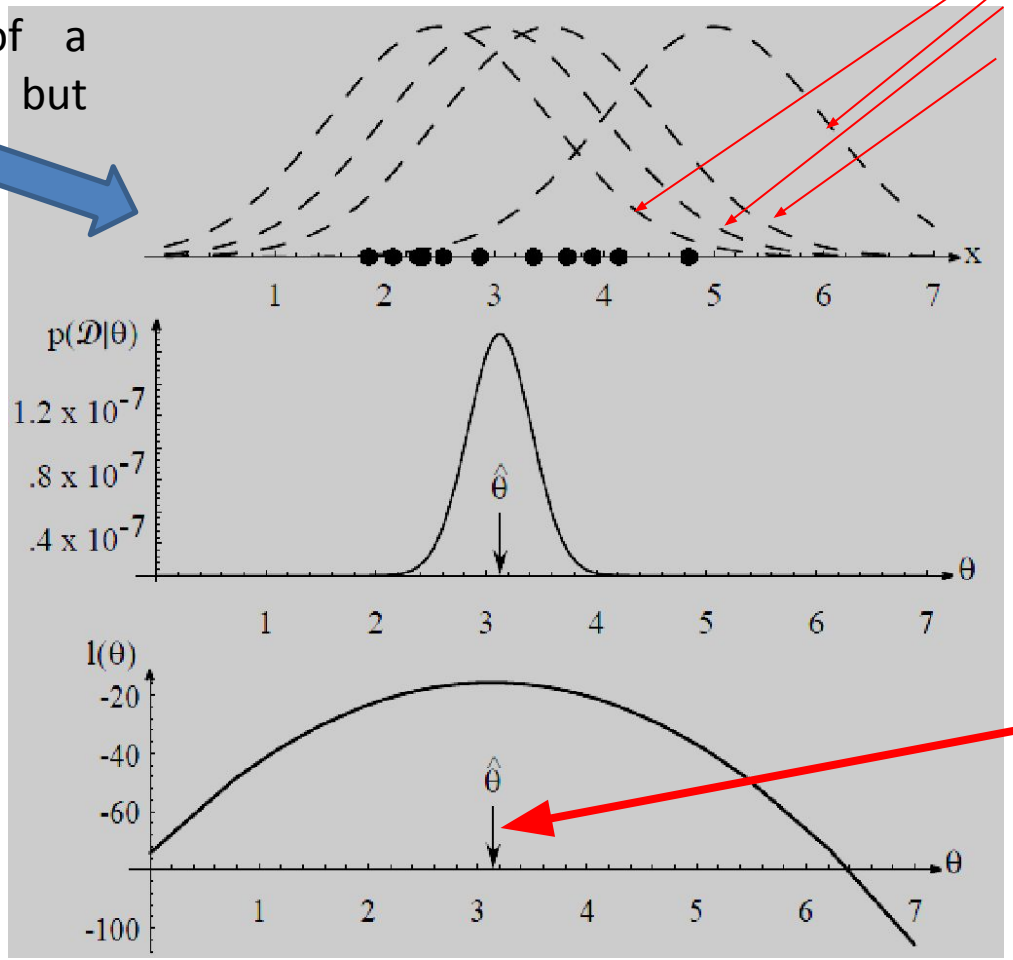


Credits: William Fleshman
https://towardsdatascience.com/maximum-likelihood-estimation-984af2dcfcac

MLE: Example

Several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean

Four infinite no of sour distribution in trainset



likelihood $p(D|\theta)$ as a function of the mean

The value that maximizes the likelihood.
It also Maximizes the log-likelihood $l(\theta)$.

General Principle of MLE: example

- We can then write our solution formally as the argument θ that maximizes the loglikelihood
  - Since the logarithm is monotonically increasing, the $\hat{\theta}$ that maximizes the log-likelihood also maximizes the likelihood.
- If the number of parameters to be set is *p*, then we let θ denote the *p*-component vector $\theta = (\theta_1, ..., \theta_p)^t$, and ∇_θ be the gradient operator,

$$\nabla_\theta \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}.$$

- We define *l*(θ) as the *log-likelihood* function,

$$l(\theta) \equiv \ln p(\mathcal{D}|\theta).$$

General Principle of MLE: example

- We can then write our solution formally as the argument θ that maximizes the loglikelihood, i.e.,

$$\hat{\theta} = \arg\max_{\theta} l(\theta)$$

- Thus we have from **Eq. 1** $p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}).$

$$l(\boldsymbol{\theta}) = \sum_{k=1} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta}) \tag{5}$$

and

$$\nabla_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta}). \tag{6}$$

Thus, set of necessary conditions for the maximum likelihood estimate for θ can be obtained from the set of *p* equations $\nabla_{\boldsymbol{\theta}} l = 0.$

MLE - The Gaussian Case: Unknown μ

- To see how maximum likelihood methods results apply to a specific case, suppose that the samples are drawn from a multivariate normal population with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. *(For simplicity, consider first the case where only the mean is unknown.)*
- Under this condition, we consider a sample point $\mathbf{x}_k$ and find,

$$\ln p(\mathbf{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}) \quad (8)$$

and

$$\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k|\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}). \quad (9)$$

MLE - The Gaussian Case: Unknown μ

- Identifying θ with μ, we see from Eq. 9 that the maximum likelihood estimate for μ must satisfy

$$\sum_{k=1}^{n} \Sigma^{-1}(\mathbf{x}_k - \hat{\mu}) = 0$$

that is, each of the *d* components of $\hat{\boldsymbol{\mu}}_n$ must vanish. Multiplying by **Σ** and rearranging, we obtain

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k.$$

- It says that the maximum likelihood estimate for the unknown population mean is just the arithmetic average of the training samples — the *sample mean*, sometimes written $\hat{\boldsymbol{\mu}}_n$ to clarify its dependence on the number of samples.

MLE - The Gaussian Case: Unknown μ and Σ

- In the more general (and more typical) multivariate normal case, neither the mean μ nor the covariance matrix Σ is known.
  - these unknown parameters constitute the components of the parameter vector θ.

- Consider first the univariate case with $\theta_1 = \mu$ and $\theta_2 = \sigma_2$. Here the log-likelihood of a single point is

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

- and its derivative is,

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix}.$$

- Consider applying $\nabla_{\boldsymbol{\theta}} l = 0$. to the full log-likelihood.

MLE - The Gaussian Case: Unknown μ and Σ

- That leads to,

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0$$

And

$$-\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0,$$

- where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the maximum likelihood estimates for θ_1 and θ_2 , respectively. By substituting $\hat{\boldsymbol{\mu}}_n = \hat{\theta}_1$, $\hat{\sigma}_n = \hat{\theta}_2$ and doing a little rearranging we obtain the following maximum likelihood estimates for μ and σ^2:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2.$$

MLE - The Gaussian Case: Unknown μ and Σ

- While the analysis of the multivariate case is basically very similar: maximum likelihood estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t.$$

- Thus, once again we find that the maximum likelihood estimate for the mean vector is the sample mean.
- The maximum likelihood estimate for the covariance matrix is the

Additional Resources:

- https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1
- https://machinelearningmastery.com/what-is-maximum-likelihood-estimation-in-machine-learning/
- Deep Learning, Ian Goodfellow, 2016