Speech Recognition

Assignment 3

Maximum Likelihood Estimation

Name: 赵卓冰
Student Number: 2252750

Maximum Likelihood Estimate

For a multivariate Gaussian model, the pdf is formulated as:

$$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

where μ is the mean vector (parameter to estimate)

$\Sigma \in R^{d \times d}$ is the covariance matrix

$x \in R^d$ is a sample vector

Goal: estimate μ using MLE given a dataset $X = \{x_1, x_2, \cdots x_n\}$

Step 1: Write the Likelihood Function

Likelihood function for n independent samples is the product of all pdf:

$$L(\mu, \Sigma; X) = \prod_{i=1}^{n} p(x_i \mid \mu, \Sigma)$$

Step 2: Log

Usually, we take the log-likelihood function to turn the product to a sum

$$\ell(\mu, \Sigma; X) = \log L(\mu, \Sigma; X) = \sum_{i=1}^{n} \log p(x_i \mid \mu, \Sigma)$$

$$= \sum_{i=1}^{n} \log \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$$= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x-\mu)^T \Sigma^{-1} (x-\mu)$$

Step 3: Focus on log

To estimate μ, we suppose Σ is fixed and focus only on μ. The relevant part is

$$\ell(\mu; X) = -\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T \Sigma^{-1} (x_i-\mu)$$

Step 4: Expand the Quadratic Term

Expanding:

$$(x_i-\mu)^T \Sigma^{-1} (x_i-\mu) = x_i^T \Sigma^{-1} x_i - 2x_i^T \Sigma^{-1}\mu + \mu^T \Sigma^{-1}\mu$$

Substitute this, we get

$$\ell(\mu; X) = -\frac{1}{2}\sum_{i=1}^{n}\left(x_i^T \Sigma^{-1} x_i - 2x_i^T \Sigma^{-1}\mu + \mu^T \Sigma^{-1}\mu\right)$$

Simplify:

$$l(\mu; x) = -\frac{1}{2}\sum_{i=1}^{n} x_i^T \Sigma^{-1} x_i + \sum_{i=1}^{n} x_i^T \Sigma^{-1} \mu - \frac{n}{2} \mu^T \Sigma^{-1} \mu$$

Step 5: Take the Gradient with respect to μ.

$$x_i^T \Sigma^{-1} = (\Sigma^{-T} x_i)^T$$

$$\frac{\partial l}{\partial \mu} = -\frac{1}{2} \times 0 + \sum_{i=1}^{n} \underbrace{\frac{\partial (x_i^T \Sigma^{-1} \mu)}{\partial \mu}}_{a_1} - \frac{n}{2} \underbrace{\frac{\partial (\mu^T \Sigma^{-1} \mu)}{\partial \mu}}_{a_2}$$

Theorem 1: if $f(x) = A^T x$, then $\frac{\partial f}{\partial x} = A$

Theorem 2: if $f(x) = x^T A^T x$, then $\frac{\partial f}{\partial x} = Ax + A^T x$

when $A = A^T$, then $\frac{\partial f}{\partial x} = 2Ax$

We know Σ^{-1} is a symmetric matrix i.e. $\Sigma^{-1} = (\Sigma^{-1})^T$

According to Theorem 1, $a_1 = (x_i^T \Sigma^{-1})^T = (\Sigma^{-1})^T x_i = \Sigma^{-1} x_i$

According to Theorem 2, $a_2 = n \Sigma^{-1} \mu$

Set the gradient to zero for maximization:

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} \Sigma^{-1} x_i - n \Sigma^{-1} \mu$$

Step 6: Solve for μ

Factor out Σ^{-1}:

$$\Sigma^{-1}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0$$

Since Σ^{-1} is non-singular, we can simplify:

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

Solve for μ:

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

Final Result:

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