HW 4

Karl Hiner

November 30, 2023

1 Maximum Likelihood

Suppose we have m i.i.d. samples from the following probability distribution. This problem asks you to build a log-likelihood function, and find the maximum likelihood estimator of the parameter(s)

1.1 Multinomial distribution

The probability density function of Multinomial distribution is given by

$$f(x_1, x_2, \dots, x_k; n, \theta_1, \theta_2, \dots, \theta_k) = \frac{n!}{x_1!x_2!\dots x_k!} \prod_{j=1}^{k} \theta_j^{x_j},$$

where $\sum_{j=1}^k \theta_j = 1$ and $\sum_{j=1}^k x_j = n$. What is the maximum likelihood estimator of θ_j, $j = 1, \dots, k$?

Answer: Let $\boldsymbol{x}^i = \{x_1^i, x_2^i, \dots, x_k^i\}$ be the i-th sample, and let $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_k\}$ be the parameter vector. Given m i.i.d. samples $\{\boldsymbol{x}^i\}_{i=1}^m$, the log-likelihood function is

$$\begin{aligned}
\ell(\boldsymbol{\theta}) &= \log\left(\prod_{i=1}^{m} f(\boldsymbol{x}^i; n, \boldsymbol{\theta})\right) \\
&= \sum_{i=1}^{m} \log\left(\frac{n!}{x_1^i!x_2^i!\dots x_k^i!} \prod_{j=1}^{k} \theta_j^{x_j^i}\right) \\
&= \sum_{i=1}^{m} \left(\log(n!) - \sum_{j=1}^{k} \log(x_j^i!) + \sum_{j=1}^{k} x_j^i \log(\theta_j)\right).
\end{aligned}$$

To find the MLE, we need to maximize $\ell(\boldsymbol{\theta})$ subject to $\sum_{j=1}^k \theta_j = 1$. Let λ be the Lagrange multiplier for this constraint. The Lagrangian is given by

$$\mathcal{L}(\boldsymbol{\theta}, \lambda) = \ell(\boldsymbol{\theta}) - \lambda\left(\sum_{j=1}^{k} \theta_j - 1\right).$$

Taking the partial derivatives and setting them to zero gives

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta_j} &= \sum_{i=1}^{m} \frac{x_j^i}{\theta_j} - \lambda = 0, \quad j = 1, \dots, k \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= \sum_{j=1}^{k} \theta_j - 1 = 0.
\end{aligned}$$

Solving these equations for θ_j, $j = 1, \dots, k$:

$$\begin{aligned}
\sum_{i=1}^{m} \frac{x_j^i}{\theta_j} - \lambda = 0 & \\
\hat{\theta}_j &= \frac{\sum_{i=1}^{m} x_j^i}{\lambda} \\
&= \frac{\sum_{i=1}^{m} x_j^i}{\sum_{i=1}^{m} \sum_{j=1}^{k} x_j^i} \qquad \left(\text{solve for } \lambda \text{ using } \frac{\partial \mathcal{L}}{\partial \lambda}\right) \\
&= \frac{\sum_{i=1}^{m} x_j^i}{mn} \qquad \left(\text{since } \sum_{j=1}^{k} x_j^i = n, \forall i\right).
\end{aligned}$$

Thus, the MLE of θ_j is

$$\hat{\theta}_j = \frac{\sum_{i=1}^{m} x_j^i}{mn}, \quad j = 1, \dots, k.$$

1.2 Gaussian normal distribution

Suppose we have m i.i.d. samples from a multivariate Gaussian normal distribution on $\mathbb{R}^d$, $\mathcal{N}(\mu, \Sigma)$, which is given by

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{|\Sigma|^{1/2} (2\pi)^{d/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right).$$

What is the maximum likelihood estimator of μ and Σ? Is the ML estimator of Σ biased?

Answer: We first find the log-likelihood for the m samples:

$$\begin{aligned}
\ell(\mu, \Sigma) &= \log\left(\prod_{i=1}^{m} \mathcal{N}(x^i; \mu, \Sigma)\right) \\
&= \sum_{i=1}^{m} \log\left(\frac{1}{|\Sigma|^{1/2} (2\pi)^{d/2}} \exp\left(-\frac{1}{2}(x^i - \mu)^T \Sigma^{-1} (x^i - \mu)\right)\right) \\
&= -\frac{md}{2} \log(2\pi) - \frac{m}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{m} (x^i - \mu)^T \Sigma^{-1} (x^i - \mu)
\end{aligned}$$

To maximize $\ell(\mu, \Sigma)$ w.r.t. μ, we take the derivative and set it to zero:

$$\begin{aligned}
\frac{\partial \ell}{\partial \mu} &= -\frac{1}{2}\sum_{i=1}^{m} \frac{\partial}{\partial \mu}\big((x^i-\mu)^T\Sigma^{-1}(x^i-\mu)\big) \\
0 &= \Sigma^{-1}\sum_{i=1}^{m}(x^i-\mu) && \text{(set to zero and simplify)} \\
\hat{\mu} &= \frac{1}{m}\sum_{i=1}^{m} x^i && \text{(solve for } \mu\text{)}
\end{aligned}$$

To find the MLE of Σ, we differentiate the log-likelihood $\ell(\mu, \Sigma)$ w.r.t. Σ and set it to zero:

$$\begin{aligned}
\frac{\partial \ell}{\partial \Sigma} &= \frac{\partial}{\partial \Sigma}\left(-\frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(x^i-\mu)^T\Sigma^{-1}(x^i-\mu)\right) \\
&= -\frac{m}{2}\frac{\partial}{\partial \Sigma}\log|\Sigma| - \frac{1}{2}\frac{\partial}{\partial \Sigma}\sum_{i=1}^{m}(x^i-\mu)^T\Sigma^{-1}(x^i-\mu) && \text{(linearity of derivative)} \\
&= -\frac{m}{2}\Sigma^{-1} - \frac{1}{2}\frac{\partial}{\partial \Sigma}\sum_{i=1}^{m}(x^i-\mu)^T\Sigma^{-1}(x^i-\mu) && \left(\frac{\partial \log|\Sigma|}{\partial \Sigma} = \left(\Sigma^{-1}\right)^T, \Sigma = \Sigma^T\right) \\
&= -\frac{m}{2}\Sigma^{-1} + \frac{1}{2}\sum_{i=1}^{m}\Sigma^{-1}(x^i-\mu)(x^i-\mu)^T\Sigma^{-1} && \left(\tfrac{\partial}{\partial X}A^T\Sigma^{-1}A = -\Sigma^{-1}AA^T\Sigma^{-1}\right) \\
m\Sigma^{-1} &= \Sigma^{-1}\left(\sum_{i=1}^{m}(x^i-\mu)(x^i-\mu)^T\right)\Sigma^{-1} && \text{(set to zero and rearrange)} \\
m\Sigma &= \sum_{i=1}^{m}(x^i-\mu)(x^i-\mu)^T && \text{(left \& right multiply by } \Sigma\text{)} \\
\hat{\Sigma} &= \frac{1}{m}\sum_{i=1}^{m}(x^i-\hat{\mu})(x^i-\hat{\mu})^T && \text{(solve for } \Sigma \text{ and use } \mu = \hat{\mu}\text{)}
\end{aligned}$$

Note that the matrix calculus identity used in the fourth step requires that Σ is invertible. Since the covariance matrix Σ is symmetric positive definite, it is invertible.

Thus, the MLE of μ is $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x^i$ and the MLE of Σ is $\hat{\Sigma} = \frac{1}{m}\sum_{i=1}^{m}(x^i-\hat{\mu})(x^i-\hat{\mu})^T$.

Bias of $\hat{\Sigma}$:

$\hat{\Sigma} = \frac{1}{m}\sum_{i=1}^{m}(x^i-\hat{\mu})(x^i-\hat{\mu})^T$ uses the sample mean $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x^i$. This estimate reduces the degrees of freedom since $\hat{\mu}$ is not independent of the samples x^i. As a result, the expectation $E[\hat{\Sigma}]$ is a scaled version of the true Σ, and so $\hat{\Sigma}$ is biased.

1.3 Exponential distribution

The probability density function of Exponential distribution is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

What is the maximum likelihood estimator of λ?

Answer: Assuming we have m i.i.d. samples $\{x_1, x_2, \ldots, x_m\}$ from this distribution, the likelihood function $L(\lambda)$ is the product of their PDFs,

$$L(\lambda) = \prod_{i=1}^{m} \lambda e^{-\lambda x_i},$$

and the log-likelihood is:

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

To find the MLE of λ, we take the derivative of $\ell(\lambda)$ with respect to λ and set it to zero:

$$\frac{d}{d\lambda} \ell(\lambda) = \frac{m}{\lambda} - \sum_{i=1}^{m} x_i = 0.$$

Solving for λ gives the MLE:

$$\hat{\lambda} = \frac{m}{\sum_{i=1}^{m} x_i}.$$

Thus, the MLE of λ in an Exponential distribution is the reciprocal of the sample mean.

2 k-means clustering

Given m data points $x_i (i = 1, \ldots, m)$, k-means clustering algorithm groups them into k clusters by minimizing the distortion function over $\{r^{ij}, \mu^j\}$:

$$J = \sum_{i=1}^{m} \sum_{j=1}^{k} r^{ij} \|x_i - \mu^j\|^2,$$

where $r^{ij} = 1$ if x_i belongs to cluster j, and $r^{ij} = 0$ otherwise.

2.1 Part (a)

Prove that using the squared Euclidean distance $\|x_i - \mu^j\|^2$ as the dissimilarity function and minimizing the distortion function, we will have

$$\mu^j = \frac{\sum_i r^{ij} x_i}{\sum_i r^{ij}}.$$

That is, μ^j is the center of the j-th cluster.

Answer: To minimize the distortion function J, we take its derivative w.r.t. μ^j and set it to zero:

$$
\begin{aligned}
J &= \sum_{i=1}^{m}\sum_{j=1}^{k} r^{ij}(x_i - \mu^j)\cdot(x_i - \mu^j) \\
\frac{\partial J}{\partial \mu^j} &= \sum_{i=1}^{m} r^{ij}\frac{\partial}{\partial \mu^j}[(x_i - \mu^j)\cdot(x_i - \mu^j)] && \text{(differentiate w.r.t. } \mu^j\text{)} \\
&= \sum_{i=1}^{m} r^{ij}[-2(x_i - \mu^j)] && \text{(simplify the derivative)} \\
0 &= \sum_{i=1}^{m} r^{ij}(\mu^j - x_i) && \text{(set to zero and simplify)} \\
\mu^j \sum_{i=1}^{m} r^{ij} &= \sum_{i=1}^{m} r^{ij} x_i && \text{(factor out } \mu^j\text{)} \\
\mu^j &= \frac{\sum_{i=1}^{m} r^{ij} x_i}{\sum_{i=1}^{m} r^{ij}} && \text{(solve for } \mu^j\text{)}
\end{aligned}
$$

This shows that μ^j is the centroid of the j-th cluster.

2.2 Part (b)

Suppose at each iteration, we need to find two clusers $\{x_1, x_2, \dots, x_p\}$ and $\{y_1, y_2, \dots, y_q\}$ with the minimum distance to merge. Some of the most commonly used distance metrics between two clusers are:

- Single linkage: the minimum distance between any pairs of points from the two clusters, i.e.
$$\min_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \|x_i - y_j\|$$
- Complete linkage: the maximum distance between any pairs of points from the two clusters, i.e.
$$\max_{\substack{i=1,\dots,p \\ j=1,\dots,q}} \|x_i - y_j\|$$
- Average linkage: the average distance between any pairs of points from the two clusters, i.e.
$$\frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q} \|x_i - y_j\|$$

Which of the three cluster distance metrics described above would most likely result in clusters most similar to those given by k-means? (Suppose k is a power of 2.)

Answer: k-means minimizes the within-cluster sum of squared distances:

$$J = \sum_{i=1}^{m}\sum_{j=1}^{k} r^{ij}\|x_i - \mu^j\|^2,$$

where μ^j is the centroid of cluster j, and r^{ij} indicates whether x_i is in cluster j.

Of the distance metrics provided, average linkage, which calculates the average of all pairwise distances between points in two clusters A and B:

$$D(A, B) = \frac{1}{|A||B|} \sum_{x \in A, y \in B} \|x - y\|,$$

is most similar to k-means, since both methods consider all points within clusters. Unlike single or complete linkage, which focus on extreme values, the average distances used by average linkage in determining cluster "closeness" during merging aligns more closely with the variance reduction of k-means.

Thus, average linkage is likely to produce clusters that most resemble those from k-means clustering.

2.3 Part (c)

For the two moons data, which of these three distance metrics (if any) would successfully separate the two moons?

Answer: The distance metric that would most likely successfully separate the two moons is single linkage:

$$\min_{\substack{i=1,\ldots,p \\ j=1,\ldots,q}} \|x_i - y_j\|.$$

This method calculates cluster distance as the minimum distance between any pair of points from each cluster. Since the closest points are likely to be within the same moon rather than across the two different moons, the single linkage distance metric will be able to successfully group the (close) points *within* each moon, and separate the (distant) points *between* the two moons.