Prove That The Em Algorithm Monotonically Increases The Likelihood Of The Data At Each Iteration And

Prove That The Em Algorithm Monotonically Increases The Likelihood Of The Data At Each Iteration And is a fundamental concept in the field of statistical learning, particularly in the context of latent variable models and mixture models. The Expectation-Maximization (EM) algorithm is widely used for maximum likelihood estimation when data is incomplete or has hidden variables. One of its key properties is that it guarantees a non-decreasing sequence of the likelihood function values at each iteration, which ensures convergence to at least a local maximum. Understanding why this property holds requires a deep dive into the mathematical foundations of the EM algorithm, including the concepts of auxiliary functions, Jensen's inequality, and the structure of the iterative process.

Introduction to the EM Algorithm

The EM algorithm was introduced by Arthur Dempster, Nan Laird, and Donald Rubin in 1977 as a general method for maximum likelihood estimation in models with latent variables. Its core idea is to iteratively improve parameter estimates by alternating between an expectation step (E-step) and a maximization step (M-step).

What Are Latent Variables?

Latent variables are unobserved or hidden variables that influence the observed data. For example, in a mixture of Gaussians, the component labels for each data point are hidden. Since the likelihood function involving both observed and hidden variables can be complex, the EM algorithm simplifies the optimization process.

Steps of the EM Algorithm

The EM algorithm proceeds as follows:

1. **Initialization:** Start with initial parameter estimates, $\theta^{(0)}$.

2. **E-Step:** Compute the expected value of the complete-data log-likelihood with respect to the current estimate of the parameters, often denoted as $Q(\theta | \theta^{(t)})$.

3. **M-Step:** Maximize this expected log-likelihood to obtain new parameter estimates, $\theta^{(t+1)}$.

Why Does the EM Algorithm Monotonically Increase the Likelihood?

The key to understanding the monotonicity property lies in the mathematical relationship between the likelihood function, the auxiliary function $Q(\theta | \theta^{(t)})$, and Jensen's inequality.

The Likelihood Function and Complete-Data Log-Likelihood

Given observed data X, the likelihood function is:

$$
L(\theta) = p(X | \theta)
$$

The complete-data likelihood, involving both observed and hidden data Z, is:

$$
L_c(\theta) = p(X, Z | \theta)
$$

Since Z is unobserved, the EM algorithm leverages the expected complete-data log-likelihood.

The Role of the Auxiliary Function $Q(\theta | \theta^{(t)})$

At each iteration, the algorithm computes:

$$
Q(\theta | \theta^{(t)}) = \mathbb{E}_{Z | X, \theta^{(t)}}[\log p(X, Z | \theta)]
$$

This function acts as a lower bound to the observed-data log-likelihood, which can be shown as:

$$
\log p(X | \theta) \geq Q(\theta | \theta^{(t)}) + \text{constant}
$$

Mathematical Proof of Monotonicity

The proof hinges on two main inequalities:

- Jensen's inequality, which relates the expectation of a convex function to the function of the expectation.
- Properties of the maximization step, which guarantees that the new parameters maximize the auxiliary function.

The steps are as follows:
1. E-step: Compute $Q(\theta | \theta^{(t)})$. This is the expectation of the complete-data log-likelihood given the current parameters.
2. M-step: Find $\theta^{(t+1)}$ that maximizes $Q(\theta | \theta^{(t)})$:
$$
\theta^{(t+1)} = \arg \max_\theta Q(\theta | \theta^{(t)})
$$
3. It can be shown that:
$$
\log p(X | \theta^{(t+1)}) \geq Q(\theta^{(t+1)} | \theta^{(t)}) \geq Q(\theta^{(t)} | \theta^{(t)}) = \log p(X | \theta^{(t)})
$$
This chain of inequalities proves that the likelihood does not decrease at each iteration.

Detailed Explanation of the Monotonic Increase

Let's examine the proof step-by-step:

Step 1: Establishing the Lower Bound

Using Jensen's inequality, we can write:
$$
\log p(X | \theta) = \log \left(\int p(X, Z | \theta) dZ \right) \geq \int p(Z | X, \theta^{(t)}) \log \frac{p(X, Z | \theta)}{p(Z | X, \theta^{(t)})} dZ
$$
which simplifies to:
$$
\log p(X | \theta) \geq Q(\theta | \theta^{(t)}) + \text{constant}
$$
This shows that $Q(\theta | \theta^{(t)})$ is a lower bound on the likelihood, and increasing $Q(\theta | \theta^{(t)})$ increases the likelihood.

Step 2: Maximizing the Auxiliary Function

During the M-step, the parameters are chosen to maximize $Q(\theta | \theta^{(t)})$. Since this maximization step is designed to find the global maximum of $Q(\theta | \theta^{(t)})$, it guarantees:
$$
Q(\theta^{(t+1)} | \theta^{(t)}) \geq Q(\theta^{(t)} | \theta^{(t)})
$$
which, in turn, implies:

$$
\log p(X | \theta^{(t+1)}) \geq \log p(X | \theta^{(t)})
$$
Therefore, the likelihood function is non-decreasing at each iteration.

Implications and Practical Significance

The monotonic property provides a theoretical guarantee that the EM algorithm will not decrease the likelihood, ensuring stability and convergence in practice. However, it does not guarantee convergence to the global maximum; it may only find a local maximum depending on initialization.

Convergence Behavior

- The likelihood sequence $\{L(\theta^{(t)})\}$ is non-decreasing.
- The sequence converges to a stationary point, which could be a local maximum or saddle point.
- Proper initialization and multiple runs can help in approaching the global maximum.

Conclusion

The proof that the EM algorithm monotonically increases the likelihood of the data at each iteration rests fundamentally on the properties of the auxiliary function $Q(\theta | \theta^{(t)})$ and Jensen's inequality. By carefully selecting the parameters to maximize this auxiliary function during the M-step, the algorithm ensures a non-decreasing likelihood sequence, providing both theoretical assurance and practical robustness. Understanding this proof not only solidifies confidence in the EM algorithm's effectiveness but also highlights the elegant interplay between probability theory and optimization techniques that underpin modern statistical inference.

Frequently Asked Questions

What is the main goal of the EM algorithm in statistical modeling?

The primary goal of the EM algorithm is to find maximum likelihood estimates of parameters in probabilistic models with latent variables by iteratively improving the likelihood of the observed data.

How does the EM algorithm ensure that the likelihood of the data increases with each iteration?

The EM algorithm guarantees that each iteration increases or maintains the likelihood because the E-step

computes the expected complete-data log-likelihood, and the M-step maximizes this expectation, ensuring non-decreasing likelihood values.

What is the role of the Jensen's inequality in proving the monotonicity of the EM algorithm?

Jensen's inequality is used to derive a lower bound on the observed data log-likelihood, and the EM algorithm improves this bound at each step, thereby ensuring that the likelihood does not decrease over iterations.

Can the EM algorithm decrease the likelihood at any iteration? Why or why not?

No, the EM algorithm cannot decrease the likelihood at any iteration because each step is designed to maximize the expected complete-data log-likelihood, which serves as a lower bound to the observed data likelihood, thus guaranteeing non-decreasing behavior.

What mathematical property of the M-step ensures the monotonic increase of the likelihood?

The M-step explicitly maximizes the expected complete-data log-likelihood function, leading to parameter updates that do not decrease the observed data likelihood, thereby ensuring monotonic increase.

How does the proof of monotonic likelihood increase relate to the concept of a fixed point in the EM algorithm?

The proof shows that each iteration moves the parameters closer to a fixed point where the likelihood cannot be increased further, indicating convergence; at this point, the parameters satisfy the stationary point condition of the likelihood function.

Why is the monotonic increase property important for the convergence of the EM algorithm?

The monotonic increase ensures that the algorithm steadily approaches at least a local maximum of the likelihood function, providing stability and convergence guarantees in parameter estimation.

Additional Resources

The EM Algorithm: Ensuring Monotonic Likelihood Improvement at Every Step

---

In the realm of statistical modeling and machine learning, the Expectation-Maximization (EM) algorithm stands out as a cornerstone technique for maximum likelihood estimation in the presence of incomplete or hidden data. Its elegant iterative process allows practitioners to find parameter estimates that maximize the likelihood function, even when direct optimization is infeasible. One of the fundamental properties that make EM both reliable and theoretically appealing is its guarantee of monotonically increasing likelihood with each iteration. This feature ensures that, as the algorithm proceeds, it continually moves toward better models, avoiding the pitfall of regressions or oscillations that can plague other optimization routines.

In this comprehensive exploration, we will dissect the mathematical foundations that underpin this monotonic behavior, clarify how the EM algorithm guarantees it, and demonstrate its significance in practical applications. We'll navigate through the core concepts, formal proofs, and intuitive explanations, making the discussion accessible yet rigorous.

Understanding the Foundations: Likelihood, Complete and Incomplete Data

Before delving into the proof of monotonicity, it's essential to clarify the key concepts involved:

- Likelihood Function: A function that measures how well a set of parameters explains the observed data. Formally, for observed data Y and parameters θ, the likelihood is $L(\theta) = p(Y | \theta)$.

- Complete Data: An augmented dataset that includes both observed data Y and hidden or latent variables Z. The joint likelihood of complete data is $p(Y, Z | \theta)$.

- Incomplete Data: The observed data Y alone, with the latent variables Z unobserved or missing.

The EM algorithm leverages the relationship between these two data types to iteratively improve parameter estimates. The core idea is to alternate between estimating the distribution of hidden variables given current parameters (E-step) and maximizing the expected complete-data log-likelihood (M-step).

The EM Algorithm: An Iterative Procedure

The EM algorithm proceeds through the following steps:

1. Initialization: Start with an initial parameter guess $\theta^{(0)}$.

2. E-step (Expectation): Compute the expected value of the complete-data log-likelihood with respect to the current estimate $\theta^{(t)}$:

$$
Q(\theta | \theta^{(t)}) = \mathbb{E}_{Z|Y, \theta^{(t)}} [\log p(Y, Z | \theta)]
$$

This step involves calculating the posterior distribution of the hidden data given observed data and current parameters.

3. M-step (Maximization): Update parameters by maximizing $Q(\theta | \theta^{(t)})$:

$$
\theta^{(t+1)} = \arg \max_\theta Q(\theta | \theta^{(t)})
$$

The new estimate $\theta^{(t+1)}$ is chosen to increase the expected complete-data log-likelihood.

The process repeats until convergence, with each iteration intended to improve or at least not worsen the likelihood of the observed data.

Monotonicity of the EM Algorithm: The Core Theorem

The critical property of EM is that it guarantees:

$$
L(\theta^{(t+1)}) \geq L(\theta^{(t)})
$$

for each iteration t. In other words, the likelihood of the observed data does not decrease as the algorithm progresses. This is a powerful assurance, especially in complex models where direct maximization of $L(\theta)$ is challenging.

Formal Proof of Monotonic Increase

The proof hinges on the properties of the Kullback-Leibler (KL) divergence and the Jensen's inequality.

Step 1: Expressing the Likelihood and Complete-Data Log-Likelihood

Recall that the observed data likelihood can be written as:

$$
L(\theta) = p(Y | \theta) = \int p(Y, Z | \theta) dZ
$$

Similarly, the complete-data likelihood is $p(Y, Z | \theta)$, which, when maximized, is often easier to handle.

Step 2: Introducing the Jensen's Inequality

The key insight is to relate the likelihood at $\theta^{(t+1)}$ to the current parameters $\theta^{(t)}$. Starting from the log-likelihood:

$$
\log p(Y | \theta) = \log \int p(Y, Z | \theta) dZ
$$

Applying Jensen's inequality with respect to a probability distribution $q(Z)$:

$$
\log p(Y | \theta) \geq \int q(Z) \log \frac{p(Y, Z | \theta)}{q(Z)} dZ
$$

Choosing $q(Z) = p(Z | Y, \theta^{(t)})$ (the posterior distribution of the hidden variables given current parameters), we get:

$$
\log p(Y | \theta) \geq \int p(Z | Y, \theta^{(t)}) \log \frac{p(Y, Z | \theta)}{p(Z | Y, \theta^{(t)})} dZ
$$

Step 3: Deriving the Monotonicity Condition

Define:

$$
Q(\theta | \theta^{(t)}) = \mathbb{E}_{Z|Y, \theta^{(t)}} [\log p(Y, Z | \theta)]
$$

and note that:

$$
\mathcal{L}(\theta, \theta^{(t)}) = \int p(Z | Y, \theta^{(t)}) \log p(Y, Z | \theta) dZ
$$

which is the auxiliary function used in the EM algorithm.

The difference in log-likelihoods between successive iterations is:

$$
\log p(Y | \theta^{(t+1)}) - \log p(Y | \theta^{(t)}) \geq \mathcal{L}(\theta^{(t+1)}, \theta^{(t)}) - \mathcal{L}(\theta^{(t)}, \theta^{(t)})
$$

Since the M-step chooses $\theta^{(t+1)}$ to maximize $Q(\theta | \theta^{(t)})$, it ensures:

$$
Q(\theta^{(t+1)} | \theta^{(t)}) \geq Q(\theta^{(t)} | \theta^{(t)})
$$

and because the Jensen's inequality and the definitions involved guarantee that:

$$
\log p(Y | \theta^{(t+1)}) \geq \mathcal{L}(\theta^{(t+1)}, \theta^{(t)}) \geq \mathcal{L}(\theta^{(t)}, \theta^{(t)}) = \log p(Y | \theta^{(t)})
$$

we conclude that:

$$
\boxed{

\text{Likelihood never decreases:} \quad L(\theta^{(t+1)}) \geq L(\theta^{(t)})
}
\]

This chain of inequalities confirms the monotonic increase of the likelihood at each EM iteration.

Intuitive Explanation: Why Does EM Guarantee Monotonicity?

While the formal proof relies on inequalities and the properties of KL divergence, an intuitive understanding helps solidify why EM behaves reliably:

- E-step: Given the current parameters, estimate the expected contribution of the hidden data to the likelihood. Think of this as "filling in the missing pieces" based on current knowledge.

- M-step: Use this "filled-in" data to find new parameters that improve the model fit, effectively making the model better aligned with the observed data.

Because each M-step maximizes the expected complete-data likelihood conditioned on the current estimate, it cannot reduce the overall likelihood of the observed data. The process acts like climbing a hill where each step is designed to go uphill or stay level, but never downhill.

Implications for Practical Model Fitting

The monotonicity property of EM has profound implications for its applicability:

- Reliability: Practitioners can be confident that the algorithm won't regress, avoiding the risk of oscillating or diverging.

- Convergence: Although EM guarantees non-decreasing likelihoods, convergence to a local maximum (not necessarily the global maximum) is typical. Multiple initializations can help find better solutions.

- Monitoring: The likelihood value can be monitored at each iteration to assess progress, and the process can be stopped once improvements fall below a threshold.

- Robustness: EM’s property makes it particularly suitable for models with latent variables, such as Gaussian

mixtures, hidden Markov models, and factor analysis.

Extensions and Variations

While the classic EM algorithm is designed for monotonic likelihood increases, various extensions and modifications aim to improve convergence speed or escape local maxima:

- Generalized EM (GEM): Ensures non-decreasing likelihood even when the M-step only partially maximizes \(Q(\theta | \theta^{

Prove That The Em Algorithm Monotonically Increases The Likelihood Of The Data At Each Iteration And

Prove That The Em Algorithm Monotonically Increases The Likelihood Of The Data At Each Iteration And

Related Articles

- The Angle Below Is The Image Formed When A Figure Containing /KLM Isdilated By A Scale Factor Of 1/2
- Suppose That Starbucks Reduces The Price Of Its Premium Coffee From $2.20 To $1.80 Per Cup, And As A
- Suppose That An Individual Has A Body Fat Percentage Of 16.8% And Weighs 152 Pounds. How Many Pounds

Back to Home