

CSE 6740 Lecture 9

What Loss Function Should I Use? (Maximum Likelihood and Bayesian Inference)

Alexander Gray

agray@cc.gatech.edu

Georgia Institute of Technology

Today

1. Maximum likelihood estimation
2. Bayesian estimation

The starting point for making/picking a machine learning method is choosing the loss function to use. A major branching point here depends on your answer to the question “Should I be a Bayesian?”.

Maximum likelihood estimation

First, a few theoretical facts about maximum likelihood, or “What’s so special about maximum likelihood?”.

Recall: The Likelihood Function

Let X_1, \dots, X_N be IID with PDF $f(x|\theta)$ or $f(x; \theta)$. The *likelihood function* is defined by $L_N(\theta|x)$ or $L_N(\theta; x)$ or

$$\mathcal{L}_N(\theta) = \prod_{i=1}^N f(X_i; \theta), \quad (1)$$

or $L(\theta|x) = f(x|\theta)$.

The likelihood function is just the joint density of the data, except that we treat it as a function of the parameter θ , $L_N : \Theta \mapsto [0, \infty)$.

It is not a density function in general; it does not necessarily integrate to 1 with respect to θ .

Using/Interpreting the Likelihood

If X is discrete, then $L(\theta|x) = \mathbb{P}_\theta(X = x)$. If we compare the likelihood at two parameter values θ_1 and θ_2 and find that

$$\mathbb{P}_{\theta_1}(X = x) = L(\theta_1|x) > L(\theta_2|x) = \mathbb{P}_{\theta_2}(X = x) \quad (2)$$

then the sample we actually observed is more likely to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$. This can be interpreted as saying that θ_1 is a more plausible guess than θ_2 .

For continuous X , we have

$$\frac{\mathbb{P}_{\theta_1}(x - \epsilon < X < x + \epsilon)}{\mathbb{P}_{\theta_2}(x - \epsilon < X < x + \epsilon)} \approx \frac{L(\theta_1|x)}{L(\theta_2|x)}. \quad (3)$$

However, in general we don't interpret likelihoods as probabilities for θ .

Maximum Likelihood Estimator

The *maximum likelihood estimate* (MLE) $\hat{\theta}_N$, is the value of θ that maximizes $L_N(\theta)$.

The *log-likelihood function* is defined by $l_N(\theta) = \log L_N(\theta)$. Its maximum occurs at the same place as that of the likelihood function.

The same is true of the likelihood function times any constant. Thus we shall often drop constants in the likelihood function.

The log-likelihood is also sometimes called the *cross-entropy* or *deviance* in the context of classification.

Some Things We Know

The MLE is:

- consistent
- asymptotically normal
- asymptotically efficient or optimal

It is also:

- approximately the Bayes estimator

These are only true under some *regularity conditions*, which amount to smoothness conditions on $f(x; \theta)$, certain derivatives existing, etc.

Kullback-Leibler Divergence

If f_1 and f_2 are PDF's, we define the *Kullback-Liebler divergence*, (or “distance”) between f_1 and f_2 to be

$$D(f_1, f_2) = \int f_1(x) \log \left(\frac{f_1(x)}{f_2(x)} \right) dx. \quad (4)$$

It has the properties that $D(f_1, f_2) \geq 0$ and $D(f, f) = 0$, but $D(f_1, f_2) \neq D(f_2, f_1)$ in general.

We'll also write $D(\theta_1, \theta_2)$ to mean $D(f(x; \theta_1), f(x; \theta_2))$.

Identifiability

We'll say that the model \mathcal{F} is *identifiable* if $\theta_1 \neq \theta_2$ implies that $D(\theta_1, \theta_2) > 0$. This means that different values of the parameter correspond to different distributions. We'll assume henceforth that the model is identifiable.

MLE = Minimizing KL Divergence

Let θ^* denote the true value of θ . Maximizing $l_N(\theta)$ is equivalent to minimizing

$$M_N(\theta) = \frac{1}{N} (l_N(\theta) - l_N(\theta^*)) \quad (5)$$

$$= \frac{1}{N} \sum_i \log \frac{f(X_i; \theta)}{f(X_i; \theta^*)} \quad (6)$$

since $l_N(\theta^*)$ is a constant with respect to θ .

MLE = Minimizing KL Divergence

By the law of large numbers, $M_N(\theta)$ converges to

$$\mathbb{E}_{\theta^*} \left(\log \frac{f(X_i; \theta)}{f(X_i; \theta^*)} \right) = \int \log \left(\frac{f(x; \theta)}{f(x; \theta^*)} \right) f(x; \theta^*) dx \quad (7)$$

$$= - \int \log \left(\frac{f(x; \theta^*)}{f(x; \theta)} \right) f(x; \theta^*) dx \quad (8)$$

$$= -D(\theta^*, \theta). \quad (9)$$

So $M_N(\theta) \approx -D(\theta^*, \theta)$, which is maximized at θ^* since $-D(\theta^*, \theta^*) = 0$ and $-D(\theta^*, \theta) < 0$ for $\theta \neq \theta^*$.

Therefore, we expect that the maximizer will tend to θ^* . This plus some technicalities gives us consistency of MLE:

$$\hat{\theta}_N \xrightarrow{p} \theta^*.$$

Fisher Information

The *score function* is defined to be

$$s(X; \theta) = \frac{\delta \log f(X; \theta)}{\delta \theta}. \quad (10)$$

The *Fisher information* is defined to be

$$I_N(\theta) = \mathbb{V}_\theta \left(\sum_{i=1}^N s(X_i; \theta) \right) \quad (11)$$

$$= \sum_{i=1}^N \mathbb{V}_\theta (s(X_i; \theta)). \quad (12)$$

Fisher Information

For $N = 1$ we'll write $I(\theta)$ instead of $I_1(\theta)$. It turns out $I_N(\theta) = NI(\theta)$. Also,

$$I(\theta) = \mathbb{E}_\theta \left(\frac{\delta^2 \log f(x; \theta)}{\delta \theta^2} \right) \quad (13)$$

$$= - \int \left(\frac{\delta^2 \log f(x; \theta)}{\delta \theta^2} \right) f(x; \theta) dx. \quad (14)$$

When there are multiple parameters, we have a Fisher information matrix.

Asymptotic Normality of the MLE

Let $\text{se} = \sqrt{\mathbb{V}(\hat{\theta}_N)}$. Under appropriate regularity conditions,

$$\text{se} \approx 1/\sqrt{I_N(\theta)} \quad \text{and} \quad \frac{(\hat{\theta}_N - \theta)}{\text{se}} \rightsquigarrow N(0, 1). \quad (15)$$

In other words, $\hat{\theta}_N \approx N(\theta, \text{se})$.

This is still true if we replace se by $\widehat{\text{se}}$, the estimated standard error.

Thus we have an approximate confidence interval for any maximum likelihood estimator. And asymptotic unbiasedness.

Relative Efficiency

Suppose that $X_1, \dots, X_N \sim N(\theta, \sigma^2)$. The MLE is $\hat{\theta}_N = \bar{X}_N$. Another reasonable estimator of θ is the sample median $\tilde{\theta}_N$. The MLE satisfies

$$\sqrt{N}(\hat{\theta}_N - \theta) \rightsquigarrow N(0, \sigma^2). \quad (16)$$

It can be shown that the median satisfies

$$\sqrt{N}(\tilde{\theta}_N - \theta) \rightsquigarrow N\left(0, \sigma^2 \frac{\pi}{2}\right). \quad (17)$$

They both converge to the right value but the median has a higher variance.

Relative Efficiency

In general, suppose that for two estimators T_N and U_N

$$\sqrt{N}(T_N - \theta) \rightsquigarrow \mathcal{N}(0, t^2) \quad (18)$$

and

$$\sqrt{N}(U_N - \theta) \rightsquigarrow \mathcal{N}(0, u^2). \quad (19)$$

We define the *asymptotic relative efficiency* (ARE) of U to T by $\text{ARE}(U, T) = t^2 / u^2$.

In the Normal example, $\text{ARE}(\tilde{\theta}_N, \hat{\theta}_N) = 2/\pi = .63$. The interpretation is that if you use the median, you are effectively using only a fraction of the data.

Optimality of the MLE

If $\hat{\theta}_N$ is the MLE and $\tilde{\theta}_N$ is any other estimator then

$$\text{ARE}(\tilde{\theta}_N, \hat{\theta}_N) \leq 1. \quad (20)$$

Thus, the MLE has the smallest (asymptotic) variance and we say that the MLE is *efficient* or *asymptotically optimal*.

Note that all of the results we have presented are predicated on the model class being correct.

Bayesian estimation

What is the Bayesian viewpoint and methodology? Is that what I should be using?

Frequentist Viewpoint

The perspective we have taken so far is that of *frequentist* (or *classical*) statistics. It is based on these tenets:

- Probabilities refer to limiting relative frequencies. They are objective properties of the real world.

Frequentist Viewpoint

The perspective we have taken so far is that of *frequentist* (or *classical*) statistics. It is based on these tenets:

- Probabilities refer to limiting relative frequencies. They are objective properties of the real world.
- Parameters are fixed, unknown constants. Because they are not fluctuating, we do not make probability statements about parameters.

Frequentist Viewpoint

The perspective we have taken so far is that of *frequentist* (or *classical*) statistics. It is based on these tenets:

- Probabilities refer to limiting relative frequencies. They are objective properties of the real world.
- Parameters are fixed, unknown constants. Because they are not fluctuating, we do not make probability statements about parameters.
- Statistical procedures should have well-defined long-run frequency properties. For example, a 95% confidence interval should trap the true value of the parameter with limiting frequency at least 95%.

Bayesian Viewpoint

Bayesian inference takes a very different stance:

- Probability describes degree of subjective belief, not limiting frequency. Thus we can make probability statements about things other than data that can recur from some source.

Bayesian Viewpoint

Bayesian inference takes a very different stance:

- Probability describes degree of subjective belief, not limiting frequency. Thus we can make probability statements about things other than data that can recur from some source.
- We can make probability statements about parameters, even though they are fixed constants.

Bayesian Viewpoint

Bayesian inference takes a very different stance:

- Probability describes degree of subjective belief, not limiting frequency. Thus we can make probability statements about things other than data that can recur from some source.
- We can make probability statements about parameters, even though they are fixed constants.
- We make inferences about a parameter by producing a probability distribution for it. Point estimates and interval estimates may then be extracted from this distribution.