

CSE 6740 Lecture 10

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

Alexander Gray

agray@cc.gatech.edu

Georgia Institute of Technology

Today

1. Bayesian estimation
2. Should I be a Bayesian or not?

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?”.

Comments

There are some confusing aspects of Bayesianism.

There are many flavors of Bayesianism.

Some people get very emotional about all of this.

The Bayesian Method

Bayesian inference is usually done like this:

1. Choose a probability density $f(\theta)$, called the *prior* distribution, which expresses our beliefs about a parameter θ before we see any data.

The Bayesian Method

Bayesian inference is usually done like this:

1. Choose a probability density $f(\theta)$, called the *prior* distribution, which expresses our beliefs about a parameter θ before we see any data.
2. Choose a model $f(x|\theta)$ (not $f(x;\theta)$) that reflects our beliefs about x given θ .

The Bayesian Method

Bayesian inference is usually done like this:

1. Choose a probability density $f(\theta)$, called the *prior* distribution, which expresses our beliefs about a parameter θ before we see any data.
2. Choose a model $f(x|\theta)$ (not $f(x;\theta)$) that reflects our beliefs about x given θ .
3. After observing data X_1, \dots, X_N , we update our beliefs and calculate the *posterior* distribution $f(\theta|X_1, \dots, X_N)$.

The Bayesian Method

Bayesian inference is usually done like this:

1. Choose a probability density $f(\theta)$, called the *prior* distribution, which expresses our beliefs about a parameter θ before we see any data.
2. Choose a model $f(x|\theta)$ (not $f(x;\theta)$) that reflects our beliefs about x given θ .
3. After observing data X_1, \dots, X_N , we update our beliefs and calculate the *posterior* distribution $f(\theta|X_1, \dots, X_N)$.
4. Obtain point and interval estimates from the posterior distribution.

Recall Bayes' Rule

Also called *Bayes' Theorem*:

$$f(x|y) = \frac{f(y|x)f(x)}{\int f(y|x)f(x)dx}. \quad (1)$$

Note that Bayes just came up with Bayes' rule. Just a statement about conditional distributions.

Now let's combine distributions over data and parameters using Bayes' rule:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}. \quad (2)$$

Recall the Likelihood

If we have N IID observations X_1, \dots, X_N , or $\{X\}^N$, yielding a dataset x_1, \dots, x_N , or $\{x\}^N$, we can replace $f(x|\theta)$ with

$$f(\{x\}^N|\theta) = f(x_1, \dots, x_N|\theta) = \prod_{i=1}^N f(x_i|\theta) = \mathcal{L}_N(\theta). \quad (3)$$

Now the likelihood really is to be interpreted as giving the probability of a parameter.

The Posterior

Now

$$f(\theta|\{x\}^N) = \frac{f(\{x\}^N|\theta)f(\theta)}{\int f(\{x\}^N|\theta)f(\theta)d\theta} = \frac{\mathcal{L}_N(\theta)f(\theta)}{c_N} \propto \mathcal{L}_N(\theta)f(\theta) \quad (4)$$

where $c_N = \int f(\{x\}^N|\theta)f(\theta)d\theta$ is called the *normalizing constant*. Often ignored because we are interested in comparing different values of θ . So

$$f(\theta|\{x\}^N) \propto \mathcal{L}_N(\theta)f(\theta), \quad (5)$$

i.e. the *posterior* is proportional to the *likelihood* times the *prior*.

Bayes Risk

The risk of an estimator $\hat{\theta}$

$$R(\theta, \hat{\theta}) = \mathbb{E}_{\theta} \left(L(\theta, \hat{\theta}) \right) = \int L(\theta, \hat{\theta}(x)) f(x; \theta) dx \quad (6)$$

can be regarded as a function of θ .

The *Bayes risk* is

$$r(f, \hat{\theta}) = \int R(\theta, \hat{\theta}) f(\theta) d\theta \quad (7)$$

where $f(\theta)$ is a prior for θ .

Bayes Point Estimates

A *decision rule* is another name for an estimator. A decision rule which minimizes the Bayes risk is called a *Bayes rule* or *Bayes estimator*, i.e. $\hat{\theta}_f$ is a Bayes rule with respect to the prior f if

$$r(f, \hat{\theta}_f) = \inf_{\tilde{\theta}} r(f, \tilde{\theta}) \quad (8)$$

where the infimum is over all estimators $\tilde{\theta}$.

Bayesian Point Estimates

The *posterior mean*

$$\bar{\theta}_N = \int \theta f(\theta|\{x\}^N) d\theta = \frac{\int \theta \mathcal{L}_N(\theta) f(\theta) d\theta}{\int \mathcal{L}_N(\theta) f(\theta) d\theta}. \quad (9)$$

is the Bayes estimator usually considered.

Bayesian Point Estimates

Different loss functions give different Bayes rules. If $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ then the Bayes rule is the posterior mean

$$\hat{\theta}_f(x) = \int \theta f(\theta|x) d\theta = \mathbb{E}(\theta|X = x). \quad (10)$$

If $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ then the Bayes rule is the median of the posterior.

If $L(\theta, \hat{\theta})$ is 0-1 loss then the Bayes rule is the mode of the posterior. This is called the *maximum a posteriori* (MAP) estimate.

Bayesian Interval Estimates

To obtain a Bayesian interval estimate, we find a and b such that $\int_{-\infty}^a f(\theta|\{x\}^N)d\theta = \int_b^{\infty} f(\theta|\{x\}^N)d\theta = \alpha/2$. Then

$$\mathbb{P}(\theta \in (a, b)|\{x\}^N) = \int_a^b f(\theta|\{x\}^N)d\theta = 1 - \alpha \quad (11)$$

and (a, b) is called a $1 - \alpha$ *posterior interval*.

Example: Bernoulli I

Let $X_1, \dots, X_N \sim \text{Bernoulli}(p)$. Suppose we take the uniform distribution $f(p) = 1$ as a prior. The posterior has the form

$$f(p|\{x\}^N) \propto f(p)\mathcal{L}_N(p) = p^s(1-p)^{N-s} \quad (12)$$

where s is the number of successes.

Example: Bernoulli I

To get the posterior mean, we must compute

$$\bar{\theta}_N = \int \theta f(\theta|\{x\}^N) d\theta = \frac{\int \theta \mathcal{L}_N(\theta) f(\theta) d\theta}{\int \mathcal{L}_N(\theta) f(\theta) d\theta} \quad (13)$$

This example can be done analytically. The posterior is a special case of a general kind of distribution called a Beta distribution with parameters $\alpha = s + 1$ and $\beta = N - s + 1$, which has mean $\alpha/(\alpha + \beta)$. The denominator is 1. So the estimate is

$$\bar{p} = \frac{s + 1}{N + 2}. \quad (14)$$

Example: Bernoulli I

Note that the maximum likelihood estimate of p is $\hat{p} = s/N$, which is unbiased. It turns out that

$$\bar{p} = \lambda_N \hat{p} + (1 - \lambda_N) \tilde{p} \quad (15)$$

where \tilde{p} is the prior mean and $\lambda_N = N/(N + 2) \approx 1$.

Example: Bernoulli II

Now suppose that instead of a uniform prior, we use the prior $p \sim \text{Beta}(\alpha, \beta)$.

It turns out that the posterior then has the form of a beta with parameters $\alpha + s$ and $\beta + N - s$. We can again calculate the posterior mean analytically, this time as $(\alpha + s)/(\alpha + \beta + N)$.

When the prior and posterior are in the same model family, we say the prior is *conjugate* with respect to the model.

Simulation

We can obtain point and interval estimates without analytical calculations by using *simulation* (sampling).

Suppose we draw $\theta_1, \dots, \theta_M \sim f(\theta|\{x\}^N)$. We can:

- Approximate the posterior mean $\bar{\theta}$ by $\frac{1}{M} \sum_{j=1}^M \theta_j$.
- Approximate the posterior $1 - \alpha$ interval by $(\theta_{\alpha/2}, \theta_{1-\alpha/2})$ where $\theta_{\alpha/2}$ is the $\alpha/2$ sample quantile of $\theta_1, \dots, \theta_M$.
- Empirically analyze the density by approximating it nonparametrically.

Note that simulation may be very difficult.

MLE and Bayes

Let $\hat{\theta}_N$ be the maximum likelihood estimate of θ and $\widehat{\text{se}} = 1/\sqrt{NI(\hat{\theta}_N)}$.

Under appropriate regularity conditions, the posterior mean

$$\bar{\theta}_N \approx N(\hat{\theta}_N, \widehat{\text{se}}), \quad (16)$$

i.e. $\bar{\theta}_N \approx \hat{\theta}_N$.

Also, if $C_N = (\hat{\theta}_N - z_{\alpha/2}\widehat{\text{se}}, \hat{\theta}_N + z_{\alpha/2}\widehat{\text{se}})$ is the asymptotic frequentist $1 - \alpha$ confidence interval, then C_N is also an approximate $1 - \alpha$ Bayesian posterior interval.

Where Do We Get the Prior From?

We may have a subjective opinion about θ before data are observed, or true prior knowledge.

However, often, because we don't really have prior knowledge and/or because we are a kind of Bayesian who wants to be objective, we choose a *noninformative prior*.

Or we may be a different kind of Bayesian who believes in estimating the prior from data. This is called *empirical Bayes*, sometimes called *type II maximum likelihood*.

Flat Priors

Consider a *flat prior*

$$f(\theta) \propto c \quad (17)$$

where $c > 0$ is a constant.

Note that $\int f(\theta)d\theta = \infty$ so this is not a pdf. We call such a prior an *improper prior*.

In general, improper priors are not a problem as long as the resulting posterior is a well-defined pdf.

Flat priors are sometimes considered ill-defined because a flat prior on a parameter does not imply a flat prior on a transformed version of the parameter.

Universal Priors

One popular idea is that of a universal prior, or a default prior distribution that can/should be used in all situations. It is often derived from the likelihood distribution.

Examples include the minimum description length (MDL) and Jeffrey's prior.

These are usually fairly uninformative.

Should I Be a Bayesian or not?

How to choose your religion.

Arguments for Bayesianism

- **Honesty:** “You always have prior beliefs - this forces you to make them explicit.” “There is no such thing as an objective statistical inference procedure.”
- **Elegance:** Allows you to make probability statements about parameters. Takes the likelihood function to its logical conclusion. Removes the distinction between data and parameters.
- **Grand unification:** There is one unified Bayesian method (“just turn the crank”). No case-by-case analytic derivations needed to get estimators and intervals.

Arguments for Bayesianism

- **Provides more structure:** Good for small-sample situations.
- **Prior information:** Gives you a convenient way to put in prior information.
- **Historical: Correctness (now specious):** The Bayesian approach corrects certain paradoxes in maximum likelihood.

Arguments Against Bayesianism

- **No confidence:** The posterior interval is not a true confidence interval, *i.e.* we cannot make statements about the true parameter with these intervals. Generally biased as estimators.
- **Parametric-centric:** Likelihood is brittle in many nonparametric cases.
- **Computationally intensive:** Requires possibly intractable integrals/simulations or destructive approximations.

Arguments Against Bayesianism

- **Unnecessary complication:** Requires putting in prior functions even when there is no true prior information. This is its own art.
- **Prior information:** You can always put in prior information using a hierarchical model.
- **Testing:** Bayesian hypothesis testing is very sensitive to the choice of prior.

Main Things You Should Know

- Three properties of MLE
- Relative efficiency
- The Bayesian method
- What the posterior and posterior mean are
- What the Bayes risk is
- The qualitative effect of the prior on an estimator
- Frequentism vs. Bayesianism arguments