Stochastic Gradient Search and Stochastic Approximation for MLE Approximation: Handout

By Henry Bourne

In this document I provide some preliminary material presented in an easily digestible way to help the listener better follow the lecture. We begin by describing various properties of the maximum likelihood estimator (MLE).

1 The MLE

In the lecture we say that the MLE is a consistent estimator and that no other consistent estimator has lower asymptotic error than the MLE (assuming the correct distribution has been picked). But what is a consistent estimator?

Definition 1.1 ***Consistent estimator:***
An estimator that converges (in probability) to the true value of the parameter being estimated as the sample size increases. Let $\{Y_1, ..., Y_n\}$ be a sequence of observations and let $\hat{\theta}_n$ be the estimator found using $\{Y_1, ..., Y_n\}$. Then $\hat{\theta}_n$ is consistent if for any $\epsilon > 0$,

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) \to 0 \tag{1}$$

as $n \to \infty$.

This means that the MLE will become more accurate as more data is collected. The Cramer-Rao Lower bound states:

Theorem 1.1 ***Cramer-Rao Lower Bound:***
Describes a lower bound on the variance of any estimator, $\hat{\theta}$, of the deterministic parameter θ. That is,

$$Var(\hat{\theta}) \geq \frac{(\frac{\partial}{\partial\theta}\mathbb{E}(\hat{\theta}))^2}{I(\theta)} \tag{2}$$

where $I(\theta)$ is the Fischer information matrix

The MLE achieves the Cramer-Rao lower bound and as it also is a consistent estimator is thus an **efficient estimator**.

I also mention in the lecture that the MLE is approximately normal,

Theorem 1.2 ***The MLE is asymptotically normal:***
If sample size n is sufficiently large and and the true value of the parameters lies in the interior of the parameter space, then the MLE is normally distributed (given certain conditions) with mean equal to the true value of the parameter and variance equal to the inverse of the Fisher information. Let $\{Y_1, ..., Y_n\}$ be a sequence of iid observations where,

$$Y_k \sim f_\theta(y) \tag{3}$$

Let $\hat{\theta}$ be the MLE of θ, then,

$$\sqrt{n}(\hat{\theta} - \theta) \to N(0, I(\theta)^{-1}) \tag{4}$$

Note that the MlE isn't always approximately normal, for more information on the conditions under which it is normal and for a proof of the above theorem please refer to chapter 7 in [1].

2 One-hot encoding

One-hot encoding is a method for encoding data, it works as follows:

Definition 2.1 ***One-hot Encoding:***
Is where we encode using a vector with all entries equal to zero bar one which is set to one. In the context of classification one-hot encoding is used to encode the class labels, where to encode a label representing the i-th class we create a vector with all entries equal to zero bar the i-th entry, ie, let y be the label we want to encode and let's say we want the label to represent the i-th class then y is encoded as follows,

$$y = e_i \tag{5}$$

where e_i is a vector of zeros bar a 1 at the i-th entry.

Advantages of using one-hot encoding include allowing us to encode every class label as a fixed-length vector (where the length is equal to the number of class labels) and allowing each class to be represented as a separate dimension in the vector space, which can make it easier to learn relationships between datapoints.

3 Bias

In the lecture we also mention bias, specifically unbiasedness. We can define bias as follows:

Definition 3.1 ***Bias:***
Suppose we want to estimate the true parameter θ^* and we have obtained a statistic $\hat{\theta}$ which is an estimator of θ based on some observed data, D. Then we define the bias of $\hat{\theta}$ relative to θ as,

$$Bias(\hat{\theta}, \theta^*) = \mathbb{E}_{D|\theta}(\hat{\theta} - \theta^*) \tag{6}$$

We then say an estimator is **unbiased** if the bias is equal to zero.

3.1 Notation

Finally I will note here some of the notation used in the lecture.

Notation	**Meaning**
D	A dataset
θ	Model parameters
$\hat{\theta}$	The MLE
$\hat{\theta}_t$	The t-th guess of the MLE
$x^{(i)}$	The i-th element of a vector x
$N_x(\mu, \sigma)$	Pdf of a normal distribution with mean μ and standard deviation σ evaluated at x
x_i	The i-th observation
z_i	The latent variable corresponding to the i-th observation
$L(\theta)$	The likelihood evaluated for parameters θ
$x^{(i)}$	The i-th element of a vector x
α	The step-size
$q_t(z)$	The t-th guess of the pdf of the latent variable
F	The lower bound function for the EM algorithm

Table 1: Notation

References

[1] Erich Leo Lehmann. *Elements of large-sample theory.* Springer, 1999.