

# COMP9417 - Machine Learning

## Tutorial: Regression II

**Weekly Problem Set: Please submit questions 1a, 1b and 3 on Moodle by 11:55am Tuesday 1st March, 2022. Please only submit these requested questions and no others.**

### Question 1. Maximum Likelihood Estimation (MLE)

In this question we will first review and then work through a few examples of parameter estimation using the MLE technique. The following introduction can be skipped if you are comfortable with the MLE concept already.

The setting is as follows: we sample  $n$  observations (data), which we denote by  $X_1, X_2, \dots, X_n$ , and we assume that the data is independently drawn from some probability distribution  $P$ . The shorthand for this is:

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P,$$

where i.i.d. stands for independent and identically distributed. In practice, we never have access to  $P$ , we are just able to observe samples from  $P$  (namely  $X_1, \dots, X_n$ ), which we will use to learn something about  $P$ . In the simplest case, we assume that  $P$  belongs to a parametric family. For example, if we assume that  $P$  belongs to the family of normal distributions, then we are assuming that  $P$  has a probability density function (pdf) of the form

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad \theta = (\mu, \sigma^2), \quad \mu \in \mathbb{R}, \sigma > 0,$$

where we will usually refer to  $\mu$  as the mean, and  $\sigma^2$  as the variance, and we combine all unknown parameters into a single parameter vector  $\theta$  that lives in some parameter space  $\Theta$ . In this particular example,  $\Theta = \mathbb{R} \times [0, \infty)$ . Under this assumption, if we knew  $\theta$ , then we would know  $P$ , and so the learning problem reduces to learning the best possible parameter  $\theta^*$ , hence the name *parametric*.

Continuing with this example, we need a way of quantifying how good a particular choice of  $\theta$  is. To do this, we first recall the fact that for independent sets  $A, B, C$ , it holds that  $P(A \text{ and } B \text{ and } C) = P(A)P(B)P(C)$ . Therefore, we have:

$$\begin{aligned} \text{Prob of observing } X_1, \dots, X_n &= \text{Prob of observing } X_1 \times \dots \times \text{Prob of observing } X_n \\ &= p_\theta(X_1) \times \dots \times p_\theta(X_n) \\ &= \prod_{i=1}^n p_\theta(X_i) \\ &=: L(\theta). \end{aligned}$$

We call  $L(\theta)$  the *likelihood*, and it is a function of the parameter vector  $\theta$ . We interpret this quantity as the probability of observing the data when using a particular choice of parameter. Obviously, we want to choose the parameter  $\theta$  that gives us the highest possible likelihood, i.e. we wish to find the *maximum likelihood estimator*

$$\hat{\theta}_{\text{MLE}} := \arg \max_{\theta \in \Theta} L(\theta).$$

Since this is just an optimization problem, we can rely on what we know about calculus to solve for the MLE estimator.

- (a) Assume that  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, 1)$ , that is, we already know that the underlying distribution is Normal with a population variance of 1, but the population mean is unknown. Compute  $\hat{\mu}_{\text{MLE}}$ .

**Hint:** it is often much easier to work with the log-likelihood, i.e. to solve the optimisation:

$$\hat{\theta}_{\text{MLE}} := \arg \max_{\theta \in \Theta} \log L(\theta),$$

which gives exactly the same answer as solving the original problem (why?).

**Solution:**

The log-likelihood here is

$$\begin{aligned} \log L(m) &= \log \left( \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} (X_i - m)^2 \right) \right) \\ &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n (X_i - m)^2. \end{aligned}$$

Differentiating with respect to  $m$  and setting equal to zero yields:

$$\frac{\partial}{\partial m} \log L(m) = \sum_{i=1}^n (X_i - m) = 0 \implies \hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.$$

To see that this is indeed a minimum, we should perform a second derivative test, which yields:

$$\frac{\partial^2}{\partial m^2} \log L(m) = -n < 0.$$

- (b) Assume that  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ , compute  $\hat{p}_{\text{MLE}}$ . Recall that the Bernoulli distribution is discrete and has probability mass function:

$$\mathbb{P}(X = k) = p^k (1 - p)^{1-k}, \quad k = 0, 1 \quad p \in [0, 1].$$

**Solution:**

Note here that  $\theta = p$  and the parameter space is  $\Theta = [0, 1]$ . We construct the log-likelihood in the usual way:

$$\begin{aligned}\log L(p) &= \log \left( \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i} \right) \\ &= n\bar{X} \log p + n(1-\bar{X}) \log(1-p)\end{aligned}$$

Then, differentiating and setting to zero yields:

$$\frac{\partial}{\partial p} \log L(p) = 0 \implies \hat{p}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}.$$

- (c) **optional:** Assume that  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ . Compute  $(\hat{\mu}_{\text{MLE}}, \hat{\sigma}_{\text{MLE}}^2)$ .

**Solution:**

The log-likelihood here is

$$\begin{aligned}\log L(m, s^2) &= \log \left( \prod_{i=1}^n \frac{1}{\sqrt{2\pi s^2}} \exp \left( -\frac{(X_i - m)^2}{2s^2} \right) \right) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(s^2) - \frac{1}{2s^2} \sum_{i=1}^n (X_i - m)^2.\end{aligned}$$

To solve for the two MLE estimates simultaneously, we need to differentiate the log-likelihood with respect to each of the two parameters and setting each to zero, which will yield two equations (i.e. a system of equations). Solving these equations simultaneously yields the correct solution. Differentiating with respect to  $m$  first and setting equal to zero yields:

$$\frac{\partial}{\partial m} \log L(m, s^2) = \frac{1}{s^2} \sum_{i=1}^n (X_i - m) = 0 \implies \hat{\mu}_{\text{MLE}} = \bar{X}.$$

Note that in this case, the first equation does not depend on the second parameter, so we can solve it directly (this is not always the case). Next, differentiating with respect to  $s^2$

$$\frac{\partial}{\partial s^2} \log L(m, s^2) = -\frac{n}{2s^2} - \frac{1}{2s^4} \sum_{i=1}^n (X_i - m)^2 = 0 \implies \hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - m)^2 = 0.$$

To solve this, we need to refer to the first of the two equations, which tells us that  $m = \bar{X}$  is optimal, and so

$$(\hat{\mu}_{\text{MLE}}, \hat{\sigma}_{\text{MLE}}^2) = \left( \bar{X}, \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right).$$

Note, in order to be completely rigorous when finding a local maximum  $(\hat{t}_1, \hat{t}_2)$  of a function  $F(t_1, t_2)$ , we need to check the following three conditions:

1. The first order partial derivatives at  $(\hat{t}_1, \hat{t}_2)$  are zero.
2. At least one second-order partial derivative is negative
3. The determinant of the Hessian matrix is positive.

We will not verify these conditions here, and this is beyond the scope of the course, but it is important to note that we are leaving out some details.

### Question 2. Bias and Variance of an Estimator

In the previous question, we discussed the MLE as a method of estimating a parameter. But there are an infinite number of ways to estimate a parameter. For example, one could choose to use the sample median instead of the MLE. It is useful to have a framework in which we can compare estimators in a systematic fashion, which brings us to two central concepts in machine learning: bias and variance. Assume that the true parameter is  $\theta$ , and we have an estimate  $\hat{\theta}$ . Note that an estimator is just a function of the observed (random) data (i.e. we can always write  $\hat{\theta} = \hat{\theta}(X)$ ) and so is itself a random variable! We can therefore define:

$$\begin{aligned}\text{bias}(\hat{\theta}) &= \mathbb{E}(\hat{\theta}) - \theta, \\ \text{var}(\hat{\theta}) &= \mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2.\end{aligned}$$

The lab this week explores these concepts as well, and you are encouraged to do the lab exercise as you complete this question to get a full picture. A short summary of the lab in words:

- bias: tells us how far the expected value of our estimator is from the truth. Recall that an estimator is a function of the data sample we observe. The expectation of an estimator can be thought of in the following manner: imagine instead of having a single data sample, we have an infinite number of data samples. We compute the same estimator on each sample, and then take an average. This is the expected value of the estimator.
- variance: how variable our estimator is. Again, if we have an infinite number of data samples, we would be able to compute the estimator an infinite number of times, and check the variation in the estimator across all samples.

A good estimator should have low bias and low variance.

- (a) Find the bias and variance of  $\hat{\mu}_{\text{MLE}}$  where  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, 1)$ .

**Solution:**

We have already found that  $\hat{\mu}_{\text{MLE}} = \bar{X}$ . Therefore

$$\begin{aligned}\text{bias}(\hat{\mu}_{\text{MLE}}) &= \text{bias}(\bar{X}) \\ &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) - \mu \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) - \mu \\ &= \frac{1}{n} \sum_{i=1}^n \mu - \mu \\ &= \mu - \mu \\ &= 0,\end{aligned}$$

and we say  $\bar{X}$  is an unbiased estimator for  $\mu$ . Next, we have

$$\begin{aligned}\text{var}(\hat{\mu}_{\text{MLE}}) &= \text{var}(\bar{X}) \\ &= \text{var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i) \\ &= \frac{1}{n},\end{aligned}$$

where in the third equality we have used the independence of the  $X_i$ 's, and in the final equality we have used the fact that  $X_i \sim N(\mu, 1)$ .

- (b) Find the bias and variance of  $\hat{p}_{\text{MLE}}$  where  $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ .

**Solution:**

We have already found that  $\hat{p}_{\text{MLE}} = \bar{X}$ , and it can easily be shown that  $\text{var}(X_i) = p(1 - p)$ . We therefore get that

$$\text{bias}(\hat{p}_{\text{MLE}}) = 0, \quad \text{var}(\hat{p}_{\text{MLE}}) = \frac{p(1 - p)}{n}.$$

- (c) The mean squared error (MSE) is a metric that is widely used in statistics and machine learning. For an estimator  $\hat{\theta}$  of the true parameter  $\theta$ , we define its MSE by:

$$\text{MSE}(\hat{\theta}) := \mathbb{E}(\hat{\theta} - \theta)^2.$$

Show that the MSE obeys a bias-variance decomposition, i.e. we can write

$$\text{MSE}(\hat{\theta}) := \text{bias}(\hat{\theta})^2 + \text{var}(\hat{\theta}).$$

**Solution:**

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \mathbb{E}(\hat{\theta} - \theta)^2 \\ &= \mathbb{E}[\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta]^2 \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 + 2(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta) + (\mathbb{E}(\hat{\theta}) - \theta)^2] \\ &= \mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 + 2\mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)] + \mathbb{E}[\mathbb{E}(\hat{\theta}) - \theta]^2 \\ &= \mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 + 2(\mathbb{E}(\hat{\theta}) - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta) + [\mathbb{E}(\hat{\theta}) - \theta]^2 \\ &= \mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 + 0 + [\mathbb{E}(\hat{\theta}) - \theta]^2 \\ &= \text{var}(\hat{\theta}) + \text{bias}(\hat{\theta})^2.\end{aligned}$$

### Question 3. Probabilistic View of Least-Squares regression

In the tutorial last week, we viewed the least-squares problem purely from an optimisation point of view. We specified the model we wanted to fit, namely:

$$\hat{y} = w^T x$$

as well as a loss function (MSE), and simply found the weight vector  $w$  that minimized the loss. We proved that when using MSE, the best possible weight vector was given by

$$\hat{w} = (X^T X)^{-1} X^T y.$$

In this question, we will explore a different point of view, which we can call the statistical view. At the heart of the statistical view is the data generating process (DGP), which assumes that there is some true underlying function that generates the data, which we call  $f$ , but we only have access to noisy observations of  $f$ . That is, we observe

$$y = f(x) + \epsilon, \quad \epsilon \text{ is some random noise.}$$

For example, assume your  $y$ 's represent the daily temperature in Kensington. Any thermometer - even the most expensive - is prone to measurement error, and so what we actually observe is the true temperature ( $f(x)$ ) plus some random noise  $\epsilon$ . Most commonly, we will assume that the noise is normally distributed with zero mean, and variance  $\sigma^2$ . Now, consider the (strong) assumption that  $f(x)$  is linear, which means that there is some true  $\beta^*$  such that  $f(x) = x^T \beta^*$ . Therefore, we have that

$$y = x^T \beta^* + \epsilon, \quad \epsilon \sim N(0, \sigma^2),$$

and therefore,

$$y|x \sim N(x^T \beta^*, \sigma^2).$$

What this says is that our response (conditional on knowing the feature value  $x$ ) follows a normal distribution with mean  $x^T \beta^*$  and variance  $\sigma^2$ . We can therefore think of our data as a random sample of observations coming from this distribution, which in turn allows us to estimate unknown parameters via maximum likelihood, just as we did in the previous questions.

- (a) You are given a dataset  $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$  and you make the assumption that  $y_i|x_i = x_i^T \beta^* + \epsilon_i$  for some unknown  $\beta^*$  and  $\epsilon_i \sim N(0, \sigma^2)$ , where all the  $\epsilon_i$ 's are independent of each other. Write down the log-likelihood for this problem as well as the maximum likelihood estimation objective and solve for the MLE estimator  $\hat{\beta}_{\text{MLE}}$ .

**Solution:**

Under this assumption, we have that  $y_i|x_i \sim N(x_i^T \beta^*, \sigma^2)$ ,  $i = 1, \dots, n$ , or we can write this in matrix notation as:

$$y|X \sim N(X\beta^*, \sigma^2 I).$$

We can compute the likelihood of the data as

$$L(\beta) = P(y|X, \beta).$$

It is important to interpret this probability properly: it is the probability of seeing the responses  $y$  given that we have features  $X$  and we are assuming the underlying vector is  $\beta$ .  $L(\beta)$  will give us a different value of the likelihood for different choices of  $\beta$ , and we want to choose the best  $\beta$ , i.e. the one that maximizes  $L(\beta)$ . Now, let's write out what  $\log L(\beta)$  is in detail:

$$\begin{aligned} \log L(\beta) &= \log P(y|X, \beta) \\ &= \log \left( \prod_{i=1}^n P(y_i|x_i, \beta) \right) \\ &= \sum_{i=1}^n \log P(y_i|x_i, \beta) \\ &= \sum_{i=1}^n \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - x_i^T \beta)^2}{2\sigma^2} \right) \right) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x_i^T \beta)^2 \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|y - X\beta\|_2^2. \end{aligned}$$

We therefore get that the MLE estimator of  $\beta$  is

$$\begin{aligned} \hat{\beta}_{\text{MLE}} &= \arg \max_{\beta} \log L(\beta) \\ &= \arg \max_{\beta} \left\{ -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|y - X\beta\|_2^2 \right\} \\ &= \arg \min_{\beta} \left\{ \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \|y - X\beta\|_2^2 \right\} \\ &= \arg \min_{\beta} \|y - X\beta\|_2^2 \\ &= (X^T X)^{-1} X^T y \\ &= \hat{\beta}_{\text{LS}}. \end{aligned}$$

The second equality holds because maximising an objective is the same as minimizing the negative of the same objective, and the third equality holds since the minimizer is unaffected by the first term. Note that we have shown that the MLE in this case is exactly identical to the least squares estimator, this is not true in general. For example, had we used a different loss function (not MSE), or a different assumption about the noise, or a different distribution (other than the normal) then we would not get equivalence between LS and MLE. We can think of this as a probabilistic justification for doing least squares.