COMP9417 – Machine Learning and Data Mining

Tutorial Week 3: Linear Regression - II

Arpit Kapoor
School of Mathematics and Statistics
arpit.kapoor@unsw.edu.au

March 3, 2023

Tutorial Slides

- Tutorial slides available at: `https://arpit-kapoor.com/COMP9417/`
- Tutorial 1 slides with solutions discussed in the class are now up
- Slides will be uploaded on Fridays EOD
- Disclaimer: These are supplementary slides and not endorsed by course admins.

Maximum Likelihood Estimate

Let's say our data samples are independently drawn from some distribution P with unknown parameters, ie.

$$X_1, X_2, \ldots X_n \overset{\text{iid}}{\sim} P \tag{1}$$

here, we do not have access to P, but we have sampled data ie. $X_1, X_2, \ldots, X_n$ Assuming that P belongs to a family of known parametric distributions (with parameters θ), can we learn the parameters for this distribution from the data (X_i) alone?

If we say that P belongs to a family of normal distributions, we are assuming the probability density function (pdf) of the form:

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

Maximum Likelihood Estimate

Likelihood: Measure of the probability of observing the data $X_1, \ldots, X_n$, given parameters θ,
ie.

$$\mathcal{L} = p(\mathbf{X} \mid \theta) = p_\theta(\mathbf{X}) \tag{2}$$

Now, given that our data is assumed to be independently sampled, the likelihood $\mathcal{L}$ is written as,

$$\begin{aligned} \mathcal{L}(\theta) &= p_\theta(X_1).p_\theta(X_2) \ldots p_\theta(X_n) \\ &= \prod_{i=1}^{n} p_\theta(X_i) \end{aligned} \tag{3}$$

Note: *Likelihood is a product of probability densities and not a probability. This means $\mathcal{L}$ can be greater than 1*

Maximum Likelihood Estimate

Maximum likelihood estimation (MLE) is the process of estimating the parameters of distribution $P(\theta)$ that maximize the likelihood,

$$\hat{\theta}_{MLE} := \underset{\theta \in \Theta}{\arg\max}\ \mathcal{L}(\theta) \tag{4}$$

where Θ is the parameter space.

Maximizing the $\log$ of likelihood is the same as maximizing the likelihood itself, as $\log$ is an *asymptotic* function. Therefore,

$$\hat{\theta}_{MLE} := \underset{\theta \in \Theta}{\arg\max}\ \log \mathcal{L}(\theta) \tag{5}$$

The $\log$ transformation makes computation easier by converting products to summations.

Maximum Likelihood Estimate

To compute the MLE estimate of θ, we compute the partial derivative of the log-likelihood with respect to each θ_i in θ and set it to zero. We then solve the equation for the value of θ_i.

$$\frac{\partial \log \mathcal{L}}{\partial \theta_i} \overset{set}{=} 0$$

Second Derivative Test

Let $f(x)$ be a twice-differentiable function, and let c be a critical point of $f(x)$. Then:

- If $f''(c) > 0$, then $f(c)$ is a local minimum of $f(x)$.
- If $f''(c) < 0$, then $f(c)$ is a local maximum of $f(x)$.
- If $f''(c) = 0$, then the test is inconclusive, requiring further investigation.

It is worth noting that the second derivative test only determines whether a critical point is a local maximum, a local minimum, or a saddle point. It does not provide any information about the global behaviour of the function.

Tutorial Question 1

(a) Assume that $X_1, X_2, \dots X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$, that is, we already know that the underlying distribution is Normal with a population variance of 1, but the population mean is unknown. Compute $\hat{\mu}_{MLE}$.

Solution. We know that the sampled data follows a normal distribution with μ mean and unit variance with pdf,

$$p_\mu(x) = \frac{1}{\sqrt{2\pi}} exp\Big(- \frac{(x-\mu)^2}{2}\Big) \tag{6}$$

The log-likelihood can thus be written as,

$$\log \mathcal{L}(m) \quad = \quad \log \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp\Big(- \frac{(X_i - m)^2}{2}\Big)\right)$$

$$
\begin{aligned}
\log \mathcal{L}(m) &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}} - \frac{(X_i - m)^2}{2} \\
&= -\frac{n}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{n} (X_i - m)^2 \qquad (7)
\end{aligned}
$$

Now taking the partial derivative with respect to m and setting to zero,

$$
\begin{aligned}
\frac{\partial \log \mathcal{L}}{\partial m} &= 0 \\
\sum_{i=1}^{n} (X_i - m) &= 0
\end{aligned}
$$

on solving for m,

$$
\boxed{\hat{\mu}_{MLE} = \overline{X}} \qquad (8)
$$

To check if this is indeed the maximum, we check if 2^{nd} derivative is ¡ 0,

$$\begin{aligned} \frac{\partial^2 \log \mathcal{L}(m)}{\partial m^2} &= \frac{\partial \sum_{i=1}^{n}(X_i - m)}{\partial m} \\ &= -n \end{aligned} \tag{9}$$

Thus,

$$\boxed{\frac{\partial^2 \log \mathcal{L}(m)}{\partial m^2} < 0}$$

Tutorial Question 1

(b) Assume that $X_1, X_2, \ldots X_n \overset{\text{iid}}{\sim} \textit{Bernoulli}(p)$, compute $\hat{p}_{MLE}$. Recall that the Bernoulli distribution is discrete and has a probability mass function:

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

(c) Assume that $X_1, X_2, \ldots X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Compute $(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE})$.

Bias and Variance

- **Bias** of an estimator represents the difference between the expected value of model prediction and the ground truth.
  If $\hat{\theta}$ is the estimate of true parameter θ, the bias of the estimator is given as,

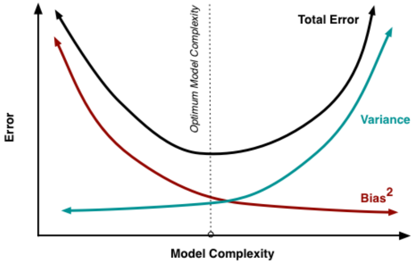
$$\textit{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta \tag{10}$$

- **Variance** represents the variability of model prediction for a given data point.

$$\textit{var}(\hat{\theta}) = \mathbb{E}[\hat{\theta} - \mathbb{E}(\hat{\theta})]^2 \tag{11}$$

- Theoretically, the expected value is like taking an average over infinite samples.

Bias-Variance Trade-off



- A model with high bias pays very little attention to the training data and oversimplifies the model which leads to a high error in training and test data.
- A model with high variance pays a lot of attention to training data and does not generalize well on unseen data leading to a high error in testing.
- A good estimator should have low bias and low variance.

Tutorial Question 2

(a) Find the bias and variance of $\hat{\mu}_{MLE}$ where $X_1, X_2, \ldots X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1)$

Solution. We konw that $\hat{\mu}_{MLE} = \overline{X}$
So,

$$\begin{aligned} bias(\hat{\mu}_{MLE}) &= bias(\overline{X}) \\ &= \mathbb{E}(\overline{X}) - \mu \\ &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) - \mu = \frac{1}{n}(n\mu) - \mu = 0 \end{aligned} \tag{12}$$

Therefore, we can say that $\overline{X}$ is an unbiased estimator of μ

Next,

$$
\begin{aligned}
var(\hat{\mu}_{MLE}) &= var(\overline{X}) \\
&= var\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} var(X_i) \qquad \text{[Using } var(cX) = c^2 var(X)\text{]} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} 1 = \frac{1}{n} \qquad (13)
\end{aligned}
$$

Tutorial Question 2

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

(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 θ, we define its MSE by:

$$MSE(\hat{\theta}) := \mathbb{E}(\hat{\theta} - \theta)^2 \tag{14}$$

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

$$MSE(\hat{\theta}) := bias(\hat{\theta})^2 + var(\hat{\theta}) \tag{15}$$

Solution.

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}(\hat{\theta} - \theta)^2 \\ &= \mathbb{E}(\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta)^2 \end{aligned} \tag{16}$$

Considering,

$$\begin{aligned} a &= \hat{\theta} - \mathbb{E}(\hat{\theta}) \\ b &= \mathbb{E}(\hat{\theta}) - \theta \end{aligned}$$

And applying,

$$(a+b)^2 = a^2 + b^2 + 2ab$$

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}\Big[(\hat{\theta} - \mathbb{E}(\hat{\theta})^2 + 2(\hat{\theta} - \mathbb{E}(\hat{\theta})(\mathbb{E}(\hat{\theta}) - \theta) + (\mathbb{E}(\hat{\theta}) - \theta)^2\Big] \\ &= \mathbb{E}[\hat{\theta} - \mathbb{E}(\hat{\theta})]^2 + 2[\mathbb{E}(\hat{\theta}) - \mathbb{E}(\hat{\theta})][\mathbb{E}(\hat{\theta}) - \mathbb{E}(\theta)] + (\mathbb{E}(\hat{\theta}) - \theta)^2 \\ &= var(\hat{\theta}) + bias(\hat{\theta})^2 \end{aligned} \tag{17}$$

MLE of Least Squares Regression

We derived that the least-squares estimate for linear regression problem $\hat{y} = \theta^T x$ is,

$$\hat{\theta} = (X^T X)^{-1} X^T y \tag{18}$$

From a statistical view, we assume that our data is generated by some underlying function $f(.)$, but we only have access to noisy observations of f:

$$y = f(x) + \epsilon, \qquad \epsilon \text{ is random noise} \tag{19}$$

With linear regression, we make two assumptions,

- Noise in observations is normally distributed with zero mean and some variance, ie. $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- Generating function f is linear with some true parameters θ^*, ie. $f(x) = x^T \theta^*$

MLE of Least Squares Regression

With these assumptions, we have

$$y = x^T\theta^* + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

and therefore,

$$y \mid x \sim \mathcal{N}(x^T\theta^*, \sigma^2)$$

We can now 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.

Tutorial Question 3

Q. You are given a dataset $D = \{(X_1, y_1), \ldots, (X_n, y_n)\}$ and you make the assumption that $y_i \mid x_i = x_i^T \beta^* + \epsilon$ for some unknown β^* and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, where all the 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 $\hat{\beta}_{MLE}$.

Solution. We can say,

$$y_i \mid x_i \sim \mathcal{N}(x_i^T \theta^*, \sigma^2)$$

or in matrix notation,

$$y \mid X \sim \mathcal{N}(X\theta^*, \sigma^2) \tag{20}$$

The likelihood of data can be written as,

$$\mathcal{L}(\beta) = p(y \mid X, \beta) \tag{21}$$

We want to find the best estimate for β that maximizes $\mathcal{L}(\beta)$ where the log-likelihood is written as,

$$\begin{aligned}
\log \mathcal{L}(\beta) &= \log p(y \mid X, \beta) \\
&= \log \prod_{i=1}^{n} p(y_i \mid x_i, \beta) = \sum_{i=1}^{n} \log p(y_i \mid x_i, \beta) \\
&= \sum_{i=1}^{n} \log \Big(\frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(- \frac{(y_i - x_i^T \beta)^2}{2\sigma^2} \Big) \Big) \\
&= \frac{-n}{2} \log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(y_i - x_i^T \beta)^2}{2\sigma^2} \\
&= \frac{-n}{2} \log(2\pi\sigma^2) - \frac{\|y - X\beta\|_2^2}{2\sigma^2} \qquad (22)
\end{aligned}$$

MLE estimate of β is,

$$\begin{aligned}\hat{\beta}_{MLE} &= \underset{\beta}{\arg\max}\ \log \mathcal{L}(\beta) \\ &= \underset{\beta}{\arg\max} \left\{ \frac{-n}{2}\log(2\pi\sigma^2) - \frac{\|y - X\beta\|_2^2}{2\sigma^2} \right\} \\ &= \underset{\beta}{\arg\min} \left\{ \frac{n}{2}\log(2\pi\sigma^2) + \frac{\|y - X\beta\|_2^2}{2\sigma^2} \right\} \\ &= \underset{\beta}{\arg\min} \|y - X\beta\|_2^2 \qquad (23)\end{aligned}$$

Therefore,

$$\boxed{\hat{\beta}_{MLE} = (X^T X)^{-1} X^T y = \hat{\beta}_{LS}} \qquad (24)$$

Note that this result is only true in this case because of the assumptions that we have made about the distribution and the generating function. This result can be seen as the probabilistic justification of doing least-squares estimation.