ST3009 – Statistics

Linear Regression

- **Line of Best Fit Guess:** $h_\theta(x) = \theta_0 + \theta_1 x$
- **Parameters:** θ_0 (Y-intercept), θ_1 (Slope)
- **Cost Function:** The cost function represents the distance of each actual data point (y_i) from the line of best fit ($h_\theta(x_i)$).

$$J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)^2$$

- **Goal:** Select a y-intercept (θ_0) and a slope (θ_1) for our line of best fit that minimises $J(\theta_0, \theta_1)$ i.e gives us the most accurate line of best fit.
- **Adding Noise:** Most of the time our sample data is affected by randomly distributed Gaussian noise M. This can be factored into our line of best fit simply by:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

$$Y = h_\theta(x) + M$$

Our training data can now be considered as:

$$\{(x_1, h_\theta(x_1) + M_1), \dots ((x_i, h_\theta(x_i) + M_i)\}$$

- **Gaussian RV:** A Gaussian random variable Z with mean μ and variance σ^2 has pdf:

$$f_z(Z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Z-\mu)^2}{2\sigma^2}}$$

Take for example an M with mean 0 and variance 1, we can then assume:

$$f_M(m) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{m^2}{2}}$$

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-h_\theta(x))^2}{2}}$$

- **Bayes Rule:** Using Bayes rule we can infer the posterior, likelihood and prior:

$$f_{\theta|D}(\theta|d) = \frac{f_{D|\theta}(d|\theta)f_{\theta}(\theta)}{f_D(d)}$$

Posterior: The probability of parameter given data
Likelihood: The probability of data given parameter
Prior: Probability of prior

- **Likelihood:** The likelihood - $f_{D|\theta}(d|\theta)$ – of the training data d is therefore:

$$f_{D|\theta}(d|\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - h_\theta(x_i))^2}{2}}$$

$$= \frac{1}{\left(\sqrt{2\pi}\right)^m} e^{-\sum_{i=1}^{m} \frac{(y_i - h_\theta(x_i))^2}{2}}$$

Taking the log of both sides we get:

$$\log f_{D|\theta}(d|\theta) = \log \frac{1}{\left(\sqrt{2\pi}\right)^m} - \sum_{i=1}^{m} \frac{(y_i - h_\theta(x_i))^2}{2}$$

- **Maximum Likelihood Estimate (ML):** The ML of θ maximises the likelihood. Equivalently, it maximises the log-likelihood. We can also drop the scaling factor $\frac{1}{\sqrt{2\pi}}$. This, therefore leaves us with a new equation to me maximised:

$$-\sum_{i=1}^{m} \frac{(y_i - h_\theta(x_i))^2}{2}$$

- **Maximum Posterior Estimate (MAP):** The MAP is an estimate for θ that maximises the posterior of Bayes rather than the likelihood.

Consider the following model:

$$Y = \sum_{i=1}^{m} \theta_i x_i + M$$

Where:

$\boldsymbol{M \sim N(0, 1)}$ i.e Gaussian distributed with mean 0 and variance 1
$\boldsymbol{\theta_i \sim N(0, \lambda)}$ i.e Gaussian distributed with mean 0 and variance λ

We already know that the likelihood is:

$$f_{D|\theta}(d|\theta) \propto e^{-\sum_{j=1}^{n}\frac{(y_j-\theta_i x_{i_j})^2}{2}}$$

From our model, we have prior as:

$$f_{\theta_i}(\theta) \propto e^{-\frac{\theta^2}{2\lambda}}$$

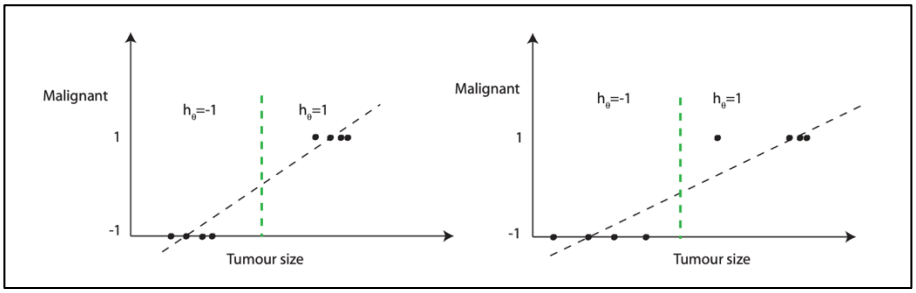
The evidence (denominator - $f_D(d)$) is a normalising constant, so area under PDF = 1. Combing these rules with Bayes we then have:

$$f_{D|\theta}(d|\theta) \propto e^{-\sum_{j=1}^{n}\frac{\left(y_j-\theta_i x_{i_j}\right)^2}{2}} * e^{-\frac{\theta^2}{2\lambda}}$$

The MAP estimation for this given model is the value of θ that maximises $f_{D|\theta}(d|\theta)$

Logistic Regression

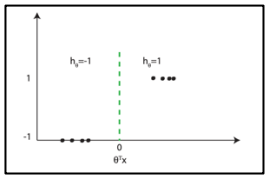
- **Classification:** Logistic regression serves the purpose of classification (email – spam or not spam?). Y values now only take values {-1, 1} whereas before Y was real-valued. We want to build a classifier that predicts the label of a new object (whether spam or not).



Line of best fit no longer works here, can result in misclassification. A more suitable model would be to predict output 1 (malignant) when $\theta^T x \geq 0$ and output -1 (not malignant) when $\theta^T x < 0$

$$h_\theta(x) = sign(\theta^T x)$$

- **Plane Fitting:** Instead of trying to choose a line of best fit between data points, logistic regression aims to choose a plane that separates Y=1 data from the Y=0 data.



- **Logistic Regression Cost Function:** Similar to how in linear regression we used a cost function such as least squares to find the line of best fit, in logistic regression we use:

$$\frac{1}{m}\sum_{i=1}^{m}\frac{\log\left(1+e^{-y_i\theta^T x_i}\right)}{\log(2)}$$

Scaling by log(2) here is optional, but it makes the loss 1 when $y_i\theta^T x_i = 0$.

- **Maximum Likelihood Estimate (ML):** Label Y only takes values of -1 and 1. Assume:

$$P(Y = y|\theta, x) = \frac{1}{1+e^{-y\theta^T x}}$$

The likelihood $P(d|\theta)$ of the training data d is therefore:

$$P(d|\theta) = \prod_{i=1}^{m}\frac{1}{1+e^{-y\theta^T x}}$$

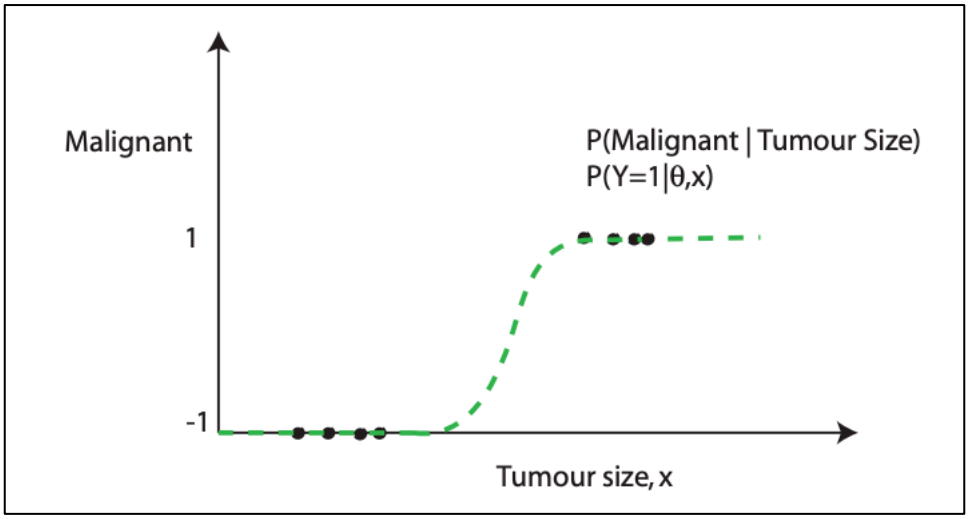
Taking logs, this gives us:

$$\log P(d|\theta) = \sum_{i=1}^{m}\log\frac{1}{1+e^{-y\theta^T x}}$$

The ML can be considered the value of θ that maximises the above equation. However, in order to simplify this, we can use the log rule $\log\left(\frac{1}{z}\right) = -\log(z)$ and then the ML becomes the value of θ that **minimises**:

$$-\sum_{i=1}^{m}\log\frac{1}{1+e^{-y\theta^T x}} = \sum_{i=1}^{m}\log(1+e^{-y\theta^T x})$$

- **Example:** Using the above examples, and the hypothesis $h_\theta(x) = sign(\theta^T x)$, we now have an estimate for our confidence in the prediction, namely: $\frac{1}{1+e^{-y\theta^T x}}$



We can see that when $\frac{1}{1+e^{-y\theta^T x}}$ is close to 1 then we are confident in our prediction, but when $\frac{1}{1+e^{-y\theta^T x}}$ is small then we are less confident in our prediction.