COMS20011 Maximum Likelihood and Classification

Laurence Aitchison

March 2022

The most exciting recent developments in "generative" AI are fundamentally based on maximum likelihood. That includes Stable Diffusion (for images) and GPT family models (for text). Taking GPT as an example, they fundamentally have a distribution over passages of text, x_i

$$\mathrm{P}\left(x_i|\theta\right) \tag{1}$$

parameterised by θ which is often e.g. neural network weights. The goal when training these models is to find the parameters, θ, which maximize the overall "likelihood" for all passages,

$$\mathrm{P}\left(x_1, \ldots, x_N|\theta\right) = \prod_{i=1}^{N} \mathrm{P}\left(x_i|\theta\right). \tag{2}$$

Once you have the right θ, you can use your neural net to sample passages that look like those in the training set!

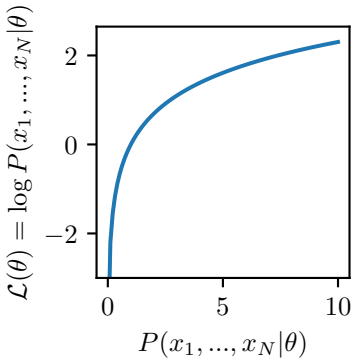
However, the big product in Eq. (2) is problematic for two reasons:

- It is very easy to get numerical over/underflow when we take big products. This is especially true in big datasets (Stable Diffusion is trained on $\sim$10 billion words).
- It is hard to do calculus on big products.

To avoid these issues, we instead consider the log-likelihood,

$$\mathcal{L}(\theta) = \log \mathrm{P}\left(x_1, \ldots, x_N|\theta\right) = \sum_i \log \mathrm{P}\left(x_i|\theta\right). \tag{3}$$

It turns out that the maximum of the log-likelihood is the same place as the maximum of the likelihood. This is intuitive, because log is an increasing function,



So if you've tweaked θ until you've reached the maximum value of $\mathrm{P}\left(x_1, \ldots, x_N|\theta\right)$, then you must also have reached the maximum value of $\mathcal{L}(\theta) = \log \mathrm{P}\left(x_1, \ldots, x_N|\theta\right)$.

1 Examples

1.1 Maximum likelihood for tossing a biased coin

That's all fine at a very high-level. But to ground everything, we need to start by working with a concrete distribution, $\mathrm{P}\left(x|\theta\right)$. Perhaps the simplest concrete distribution arises from tossing a biased coin. We're therefore going to use maximum likelihood to fit the probability, p. Once we have p, we can also sample new coin tosses that look like those in the data.

When tossing a biased coin, x can take on values of 1 or 0, with 1 corresponding to heads and 0 corresponding to tails. Because this is a biased coin, there is a parameter, p, which controls the probability of 1 = heads vs 0 = tails, Then, the probability is,

$$\mathrm{P}\left(x = 1|p\right) = p \tag{4}$$
$$\mathrm{P}\left(x = 0|p\right) = (1 - p). \tag{5}$$

The probability of heads *or* tails adds to 1,

$$\mathrm{P}\left(x = 1|p\right) + \mathrm{P}\left(x = 0|p\right) = p + (1 - p) = 1. \tag{6}$$

Overall, we sometimes write the probability as,

$$\mathrm{P}\left(x|p\right) = p^x(1 - p)^{1-x} \tag{7}$$

Just to check that this general form makes sense, we substitute $x = 1$ and $x = 0$ into the generic expression,

$$\mathrm{P}\left(x = 1|p\right) = p^1(1 - p)^0 = p \times 1 = p, \tag{8}$$
$$\mathrm{P}\left(x = 0|p\right) = p^0(1 - p)^1 = 1 \times (1 - p) = (1 - p), \tag{9}$$

which gets back to the original expressions.

Anyway: lets get back to our maximum likelihood problem. We have a dataset of tosses from a biased coin, $x_1, \ldots, x_N$. Our goal is to estimate the probability, p, that generated these coin tosses. To do that, we find the p that makes the observed data most probable; specifically we work with the log-probability of the data given p,

$$\mathcal{L}(p) = \log \mathrm{P}\left(x_1, \ldots, x_N | p\right) \tag{10}$$

$$\mathcal{L}(p) = \sum_{i=1}^{N} \log \mathrm{P}\left(x_i | p\right) \tag{11}$$

Substituting the full probability for tossing a biased coin (Eq. 7),

$$\mathcal{L}(p) = \sum_{i=1}^{N} \log \left(p^{x_i}(1-p)^{1-x_i}\right) \tag{12}$$

The log turns the product into a sum,

$$\mathcal{L}(p) = \sum_{i=1}^{N} \left(\log\left(p^{x_i}\right) + \log\left((1-p)^{1-x_i}\right)\right) \tag{13}$$

And the log turns powers into products,

$$\mathcal{L}(p) = \sum_{i=1}^{N} \left(x_i \log p + (1-x_i)\log(1-p)\right) \tag{14}$$

Applying the sum to each term separately, and noting that p is independent of i,

$$\mathcal{L}(p) = \left(\sum_{i=1}^{N} x_i\right) \log p + \left(\sum_{i=1}^{N} (1-x_i)\right) \log(1-p) \tag{15}$$

And finally, separating out the two terms in the second sum,

$$\mathcal{L}(p) = \left(\sum_{i=1}^{N} x_i\right) \log p + \left(N - \sum_{i=1}^{N} x_i\right) \log(1-p) \tag{16}$$

Now, we find the maximum likelihood value of p by solving for where the gradient is zero,

$$0 = \frac{\partial \mathcal{L}(p)}{\partial p} \tag{17}$$

$$0 = \left(\sum_{i=1}^{N} x_i\right) \frac{\partial \log p}{\partial p} + \left(N - \sum_{i=1}^{N} x_i\right) \frac{\partial \log(1-p)}{\partial p} \tag{18}$$

It is a standard result that,

$$\frac{\partial \log p}{\partial p} = \frac{1}{p}. \tag{19}$$

But to compute the other derivative, we need to apply the chain rule. Specifically, we use,

$$u = (1-p) \tag{20}$$
$$y = \log(1-p) = \log u. \tag{21}$$

Thus,

$$\frac{\partial y}{\partial p} = \frac{\partial y}{\partial u}\frac{\partial u}{\partial p} \tag{22}$$
$$= \frac{\partial \log u}{\partial u}\frac{\partial}{\partial p}[1-p] \tag{23}$$
$$= \frac{1}{u} \times (-1) \tag{24}$$
$$= -\frac{1}{1-p} \tag{25}$$

Substituting in the derivatives,

$$0 = \left(\sum_{i=1}^{N} x_i\right)\frac{1}{p} - \left(N - \sum_{i=1}^{N} x_i\right)\frac{1}{1-p} \tag{26}$$

To get the p's out of the denominators, we multiply both sides by $p(1-p)$,

$$0 = \left(\sum_{i=1}^{N} x_i\right)\frac{p(1-p)}{p} - \left(N - \sum_{i=1}^{N} x_i\right)\frac{p(1-p)}{1-p} \tag{27}$$
$$0 = \left(\sum_{i=1}^{N} x_i\right)(1-p) - \left(N - \sum_{i=1}^{N} x_i\right)p \tag{28}$$

Now, we separate out all the terms,

$$0 = \left(\sum_{i=1}^{N} x_i\right) - \left(\sum_{i=1}^{N} x_i\right)p - Np + \left(\sum_{i=1}^{N} x_i\right)p. \tag{29}$$

And cancel the second and fourth term,

$$0 = \left(\sum_{i=1}^{N} x_i\right) - Np. \tag{30}$$

Now, we add Np to both sides,

$$Np = \sum_{i=1}^{N} x_i \tag{31}$$

And divide both sides by N,

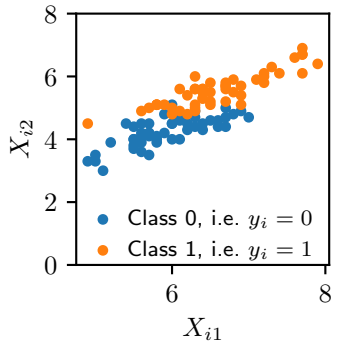
$$p = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{32}$$

2 Maximum likelihood for classification

Our data is:

- N inputs, $\mathbf{x}_1, \ldots, \mathbf{x}_N$. Could be anything, but is most often a vector of continuous features.
- N binary, $y_1, \ldots, y_N$, targets.

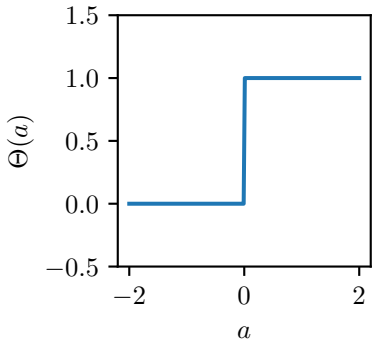
For instance:



If we forget about maximum likelihood for a second, the goal is to learn a mapping from $\mathbf{x}$'s, which could be anything but are often a vector of continuous numbers, to y, which is a class-label (usually 0 or 1). For instance, we could use,

$$f_{\mathbf{w}}(\mathbf{x}) = \Theta(\mathbf{w}^T \mathbf{x}) \tag{33}$$

where $\Theta(a)$ is the Heaviside step function, which returns 0 when the input argument is negative ($a < 0$), and returns 1 otherwise.



We could then tweak $\mathbf{w}$ so that this function generally returns similar class-labels to those in the data.

As usual, to tweak $\mathbf{w}$ systematically, we need a loss function. The obvious choice is classification error:

$$\mathcal{L}(\mathbf{w}) = \text{number of wrong classifications} \tag{34}$$

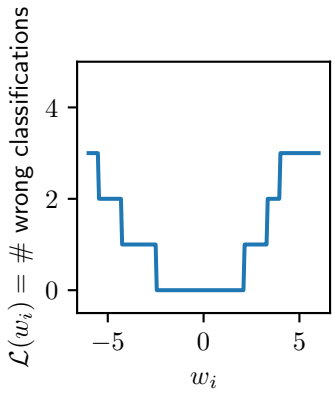
$$= \sum_{i=1}^{N} \text{different}(y_i, f_{\mathbf{w}}(\mathbf{x}_i)) \tag{35}$$

where different returns 0 if the data class-label, y_i and the prediction, $f_{\mathbf{w}}(\mathbf{x}_i)$ are the same, and returns 1 if they're different,

$$0 = \text{different}(0, 0) = \text{different}(1, 1) \tag{36}$$

$$1 = \text{different}(0, 1) = \text{different}(1, 0) \tag{37}$$

The problem is that the number of wrong classifications is always an integer (0, 1, 2 etc.)



Therefore, we can't use the usual trick of differentiating and finding where the

gradient is zero. FYI: we can't even do gradient descent (i.e. differentiating and following the gradient down hill).

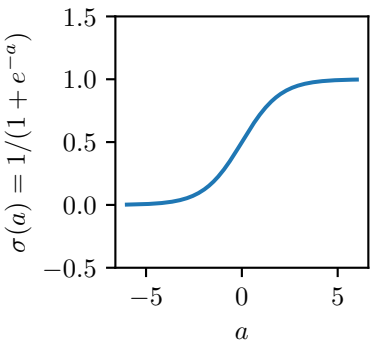
Instead, we need a differentiable loss. How do we get a differentiable loss? Remember the coin-flipping example. The data was again 0 or 1 (coin flips), but we had a differentiable objective (the log-likelihood). So can we do something similar here? Yes! The basic idea is that instead of learning a function, $f_{\mathbf{w}}(x)$, which just predicts a class, we instead predict a probability,

$$\mathrm{P}\left(y = 1|\mathbf{x}\right) =? \tag{38}$$

Key question: probabilities must be between 0 and 1. How do we design the function $p_{\mathbf{w}}(\mathbf{x})$ so that its outputs are always between 0 and 1? Answer: the sigmoid function!

$$\sigma(a) = \frac{1}{1 + e^{-a}} \tag{39}$$

Takes any a (from $-\infty$ to ∞) and returns a number from 0 to 1:



As such, we can set the probability of class 1 using the sigmoid,

$$\mathrm{P}\left(y_i = 1|\mathbf{x}_i\right) = \sigma(\mathbf{w}^T\mathbf{x}_i) = \sigma(\sum_{j=1}^{D} w_j X_{ij}) \tag{40}$$

$$\mathrm{P}\left(y_i = 0|\mathbf{x}_i\right) = 1 - \mathrm{P}\left(y_i = 1|\mathbf{x}_i\right). \tag{41}$$

It turns out we still can't analytically find the best $\mathbf{w}$. FYI: but we can now use gradient ascent (i.e. differentiating and following the gradient up hill), because the loss is smooth and differentiable.

Once we have a $\mathbf{w}$ what can we do? We can do two things. The most obvious thing we can do is to compute the probability of a class-label,

$$\mathrm{P}\left(y_{\text{test}} = 1|\mathbf{x}_{\text{test}}\right) = \sigma(\mathbf{w}^T\mathbf{x}_{\text{test}}) \tag{42}$$

which is particularly useful if you want to know how sure (certain) your classifier is. But what if you just want a best-guess? Then you can guess class 1 if the

probability of class 1 (i.e. $\mathrm{P}\left(y_{\text{test}} = 1|\mathbf{x}_{\text{test}}\right)$) is more than 50%. Alternatively, if the probabability of class 1 is less than 50%, then the probability of class 0 must be greater than 50%, and we can guess class 0.

$$f_{\mathbf{w}}(\mathbf{x}_{\text{test}}) = \begin{cases} 1 & \text{if } 0.5 < \mathrm{P}\left(y_{\text{test}} = 1|\mathbf{x}_{\text{test}}\right) \\ 0 & \text{otherwise.} \end{cases} \tag{43}$$

Which turns out to be the same as,

$$f_{\mathbf{w}}(\mathbf{x}_{\text{test}}) = \begin{cases} 1 & \text{if } 0 < \mathbf{w}^T\mathbf{x}_{\text{test}} \\ 0 & \text{otherwise.} \end{cases} \tag{44}$$

because $\mathrm{P}\left(y_{\text{test}} = 1|\mathbf{x}_{\text{test}}\right) = \sigma(\mathbf{w}^T\mathbf{x}_{\text{test}})$, and

$$\sigma(0) = \frac{1}{1+e^0} = \frac{1}{1+1} = \frac{1}{2}. \tag{45}$$

2.1 Maximum likelihood for a Gaussian

The Gaussian probability density is,

$$\mathrm{P}\left(x|m,v\right) = \frac{1}{\sqrt{2\pi v}} \exp\left(-\tfrac{1}{2v}(x-m)^2\right) \tag{46}$$

where m is the mean and v is the variance. Note that we're working with the variance, $v = \sigma^2$, rather than the standard deviation, σ, because it makes the calculus slightly easier. We're going to be using the log-probability,

$$\log \mathrm{P}\left(x|m,v\right) = \log\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2v}(x-m)^2\right)\right) \tag{47}$$

As log converts a product into a sum,

$$= \log \tfrac{1}{\sqrt{2\pi}} + \log \tfrac{1}{\sqrt{v}} + \log\left(\exp\left(-\tfrac{1}{2v}(x-m)^2\right)\right) \tag{48}$$

As log is the inverse of exp, and as $\log 1/\sqrt{x} = \log x^{-1/2} = -\frac{1}{2}\log x$,

$$= -\tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}\log(v) - \tfrac{1}{2v}(x-m)^2. \tag{49}$$

$$= -\tfrac{1}{2}\left(\log(2\pi) + \log(v) + \tfrac{1}{v}(x-m)^2\right). \tag{50}$$

Our goal is to find the maximum likelihood values of the mean, m, and variance, v. The likelihood for the full dataset is given by,

$$\mathrm{P}\left(x_1, x_2, \ldots, x_N|m,v\right) = \prod_{i=1}^{N} \mathrm{P}\left(x_i|m,v\right) \tag{51}$$

The objective is the log-likelihood,

$$\mathcal{L}(m, v) = \log \mathrm{P}\left(x_1, x_2, \ldots, x_N | m, v\right) \tag{52}$$

Note that there are two parameters, the mean, m and variance, v, not one, as we've seen in previous examples. It turns out that the maximum is usually at a location where the gradient wrt both parameters, m and v, is zero,

$$\frac{\partial \mathcal{L}(m, v)}{\partial m} = 0 \qquad\qquad \frac{\partial \mathcal{L}(m, v)}{\partial v} = 0. \tag{53}$$

Now, we separate $\mathcal{L}(m, v)$ into terms for each datapoint,

$$\mathcal{L}(m, v) = \sum_{i=1}^{N} \log \mathrm{P}\left(x_i | m, v\right) \tag{54}$$

Substitute the form for the Gaussian log-probability density,

$$\mathcal{L}(m, v) = -\tfrac{1}{2} \sum_{i=1}^{N} \left(\log(2\pi) + \log(v) + \tfrac{1}{v}(x_i - m)^2\right). \tag{55}$$

As 2π and v does not depend on the datapoint,

$$\mathcal{L}(m, v) = -\tfrac{N}{2} \left(\log(2\pi) + \log v\right) - \tfrac{1}{2v} \sum_{i=1}^{N} (x_i - m)^2. \tag{56}$$

Now, we solve for where the gradient wrt m is zero,

$$0 = \frac{\partial \mathcal{L}(m, v)}{\partial m} \tag{57}$$

$$0 = \frac{\partial}{\partial m} \left[-\tfrac{N}{2} \left(\log(2\pi) + \log v\right) - \tfrac{1}{2v} \sum_{i=1}^{N} (x_i - m)^2 \right]. \tag{58}$$

As m only appears in the final term,

$$0 = -\frac{1}{2v} \frac{\partial}{\partial m} \left[\sum_{i=1}^{N} (x_i - m)^2 \right]. \tag{59}$$

Assuming v is neither zero nor infinity, we can multiply both sides by $-2v$,

$$0 = \frac{\partial}{\partial m} \left[\sum_{i=1}^{N} (x_i - m)^2 \right]. \tag{60}$$

We have seen before that this gradient is zero when,

$$m = \tfrac{1}{N} \sum_{i=1}^{N} x_i, \tag{61}$$

i.e. when m is set to the sample mean.

If we assume m is set to the sample mean, we can find the optimal value for the variance by solving for where the gradient wrt v is zero,

$$0 = \frac{\partial \mathcal{L}(m, v)}{\partial v}. \tag{62}$$

Substitute for $\mathcal{L}(m, v)$,

$$0 = \frac{\partial}{\partial v}\left[-\tfrac{N}{2}\log(2\pi) - \tfrac{N}{2}\log v - \tfrac{1}{2v}\sum_{i=1}^{N}(x_i - m)^2\right]. \tag{63}$$

$$0 = -\frac{\partial}{\partial v}\left[\tfrac{N}{2}\log(2\pi)\right] - \tfrac{N}{2}\frac{\partial \log v}{\partial v} - \left(\tfrac{1}{2}\sum_{i=1}^{N}(x_i - m)^2\right)\frac{\partial v^{-1}}{\partial v}. \tag{64}$$

$$0 = -\frac{N}{2v} + \frac{1}{2v^2}\sum_{i=1}^{N}(x_i - m)^2. \tag{65}$$

Multiply both sides by $2v^2$,

$$0 = -Nv + \sum_{i=1}^{N}(x_i - m)^2. \tag{66}$$

Add Nv to both sides,

$$Nv = \sum_{i=1}^{N}(x_i - m)^2. \tag{67}$$

Divide both sides by N,

$$v = \tfrac{1}{N}\sum_{i=1}^{N}(x_i - m)^2. \tag{68}$$

This is one of the standard forms for the variance.

3 Exercises

Exercise 1. *This question is based on the following dataset:*

X_{i1}	X_{i2}	y_i
-2.1	*-3.2*	*0*
-3.4	*-1.2*	*0*
1.2	*3.6*	*1*
-0.1	*0.8*	*1*

Compute the log-probability, $\log \mathrm{P}\left(y_1, \ldots, y_4 | \mathbf{x}_1, \ldots, \mathbf{x}_4\right)$ *of the class-labels for the model,*

$$\mathrm{P}\left(y_i = 1 | \mathbf{x}_i\right) = \sigma(X_{i1} + X_{i2}) \tag{69}$$

Exercise 2. *You were consulted by a Physics student who is trying to estimate a parameter, the voltage (V), given measurements of the current* (I_i) *when they set a particular level of the resistance* (R_i) *information. The student informs you that the physical model is*

$$I_i = \frac{V}{R_i} + \epsilon_i$$

subject to a measurement error $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$*. Assuming i.i.d observations, the likelihood is* $p(I_i | R_i, V)$*,*

$$p(I_i | V, R_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\left(I_i - \frac{V}{R_i}\right)^2}$$

where I_i *is ith observed current obtained when we set the resistance to* R_i*. Show that the Maximum Likelihood value the parameter* V *(which is the same for all observations) is* $V = \sum \frac{I_i}{R_i} / \sum \frac{1}{R_i^2}$ *(differentiate* $\log p(D|V)$ *wrt to the parameter,* V*, then find the maximum by solving for the setting of* V *for which the gradient is zero)*

Exercise 3. *Suppose that that the data,* x_i*, are discrete with the following probability mass function, where* $0 \leq \theta \leq 1$ *is a parameter.*

x_i	*0*	*1*	*2*	*3*
$\mathrm{P}\left(x_i\right)$	$\frac{2\theta}{3}$	$\frac{\theta}{3}$	$\frac{2(1-\theta)}{3}$	$\frac{(1-\theta)}{3}$

First, verify that for $0 \leq \theta \leq 1$ *this is a valid distribution (i.e. the probabilities are between* 0 *and* 1 *and sum to* 1*).*

The following 10 independent observations were taken from this distribution:

3	*0*	*2*	*1*	*3*	*2*	*1*	*0*	*2*	*1*

Find the Maximum Likelihood estimate of θ*.*

4 Answers

Answer 1. *To begin, each class-label depends only on the corresponding input,*

$$\mathrm{P}\left(y_1, \ldots, y_4 | \mathbf{x}_1, \ldots, \mathbf{x}_4\right) = \mathrm{P}\left(y_1 | \mathbf{x}_1\right) \mathrm{P}\left(y_2 | \mathbf{x}_2\right) \mathrm{P}\left(y_3 | \mathbf{x}_3\right) \mathrm{P}\left(y_4 | \mathbf{x}_4\right). \tag{70}$$

And the log turns products into sums,

$$\begin{aligned} \log \mathrm{P}\left(y_1, \ldots, y_4 | \mathbf{x}_1, \ldots, \mathbf{x}_4\right) = \\ \log \mathrm{P}\left(y_1 | \mathbf{x}_1\right) + \log \mathrm{P}\left(y_2 | \mathbf{x}_2\right) + \log \mathrm{P}\left(y_3 | \mathbf{x}_3\right) + \log \mathrm{P}\left(y_4 | \mathbf{x}_4\right). \end{aligned} \tag{71}$$

We therefore need to compute each of these individual $\log \mathrm{P}\left(y_i|\mathbf{x}_i\right)$ *terms.*

First, we consider $\log Py_3|\mathbf{x}_3$ *and* $\log \mathrm{P}\left(y_4|\mathbf{x}_4\right)$*, as these datapoints are in class* 1 *(i.e.* $y_3 = 1$ *and* $y_4 = 1$*) so they turn out to be slightly easier to deal with.*

$$\mathrm{P}\left(y_3 = 1|\mathbf{x}_3\right) = \sigma(X_{31} + X_{32}) \tag{72}$$

Substitute in the first, $X_{31} = 1.2$*, and second,* $X_{32} = 3.6$*, feature for this datapoint,*

$$\mathrm{P}\left(y_3 = 1|\mathbf{x}_3\right) = \sigma(1.2 + 3.6) = \sigma(4.8). \tag{73}$$

Calculate the sigmoid,

$$\mathrm{P}\left(y_3 = 1|\mathbf{x}_3\right) = \sigma(4.8) = \frac{1}{1 + e^{-4.8}} = \frac{1}{1 + 0.0082} = 0.9918 \tag{74}$$

So far, this is indicative of a good classifier. The real datapoint is $y_3 = 1$*, and the probability given by the classifier that* $y_3 = 1$ *is large (*0.9981*). Finally, we compute the log,*

$$\log \mathrm{P}\left(y_3 = 1|\mathbf{x}_3\right) = \log 0.9918 = -0.0082. \tag{75}$$

We can do the same thing with the fourth datapoint,

$$\mathrm{P}\left(y_4 = 1|\mathbf{x}_4\right) = \sigma(X_{41} + X_{42}) \tag{76}$$

$$\mathrm{P}\left(y_4 = 1|\mathbf{x}_4\right) = \sigma(-0.1 + 0.8) = \sigma(0.7) \tag{77}$$

Calculate the sigmoid,

$$\mathrm{P}\left(y_4 = 1|\mathbf{x}_4\right) = \sigma(0.7) = \frac{1}{1 + e^{-0.7}} = \frac{1}{1 + 0.497} = 0.668. \tag{78}$$

Calculate the log,

$$\log \mathrm{P}\left(y_4 = 1|\mathbf{x}_4\right) = \log 0.668 = -0.403. \tag{79}$$

Now, the first and second datapoint are class 0*, so things are a bit different for them. The basic strategy is to first compute the probability for class* 1*, (i.e.* $\mathrm{P}\left(y_i = 1|\mathbf{x}_i\right)$*), then compute the probability we want (i.e. for class* 0*) using* $\mathrm{P}\left(y_i = 0|\mathbf{x}_i\right) = 1 - \mathrm{P}\left(y_i = 1|\mathbf{x}_i\right)$*. For the first datapoint,*

$$\mathrm{P}\left(y_1 = 1|\mathbf{x}_1\right) = \sigma(X_{11} + X_{12}) \tag{80}$$

$$\mathrm{P}\left(y_1 = 1|\mathbf{x}_1\right) = \sigma((-2.1) + (-3.2)) = \sigma(-5.3) \tag{81}$$

Calculate the sigmoid,

$$\mathrm{P}\left(y_1 = 1|\mathbf{x}_1\right) = \sigma(-5.3) = \frac{1}{1 + e^{-(-5.3)}} = \frac{1}{1 + e^{5.3}} = \frac{1}{1 + 200} = 0.00497 \tag{82}$$

This again is indicative of a good classifier. Remember that the real class-label is $y_1 = 0$, so the probability of $y_1 = 1$ should be small, and it is. Now we need to compute the probability for the actual label in the data (i.e. $y_1 = 0$),

$$\mathrm{P}\left(y_1 = 0|\mathbf{x}_1\right) = 1 - \mathrm{P}\left(y_1 = 1|\mathbf{x}_1\right) = 1 - 0.00497 = 0.995. \tag{83}$$

And finally, we compute the log-probability,

$$\log \mathrm{P}\left(y_1 = 0|\mathbf{x}_1\right) = \log 0.995 = -0.00498. \tag{84}$$

And for the second datapoint,

$$\mathrm{P}\left(y_2 = 1|\mathbf{x}_2\right) = \sigma(X_{21} + X_{22}) \tag{85}$$
$$\mathrm{P}\left(y_2 = 1|\mathbf{x}_2\right) = \sigma((-3.4) + (-1.2)) = \sigma(-4.6) \tag{86}$$

Calculate the sigmoid,

$$\mathrm{P}\left(y_2 = 1|\mathbf{x}_2\right) = \sigma(-4.6) = \frac{1}{1 + e^{-(-4.6)}} = \frac{1}{1 + e^{4.6}} = \frac{1}{1 + 99.4} = 0.00995 \tag{87}$$
$$\mathrm{P}\left(y_2 = 0|\mathbf{x}_2\right) = 1 - \mathrm{P}\left(y_2 = 1|\mathbf{x}_2\right) = 1 - 0.00995 = 0.990. \tag{88}$$

And finally, we compute the log-probability,

$$\log \mathrm{P}\left(y_2 = 0|\mathbf{x}_2\right) = \log 0.990 = -0.010 \tag{89}$$

Finally, we combine the log-probabilities,

$$\log \mathrm{P}\left(y_1, \ldots, y_4|\mathbf{x}_1, \ldots, \mathbf{x}_4\right) \tag{90}$$
$$= \log \mathrm{P}\left(y_1 = 0|\mathbf{x}_1\right) + \log \mathrm{P}\left(y_2 = 0|\mathbf{x}_2\right) + \log \mathrm{P}\left(y_3 = 1|\mathbf{x}_3\right) + \log \mathrm{P}\left(y_4 = 1|\mathbf{x}_4\right).$$
$$= -0.00498 - 0.010 - 0.0082 - 0.403 \tag{91}$$
$$= -0.426 \tag{92}$$

Answer 2. *The objective is the log-likelihood,*

$$\mathcal{L}(V) = \log \prod_{i=1}^{N} \mathrm{P}\left(I_i|V, R_i\right) \tag{93}$$
$$\mathcal{L}(V) = \log \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\left(I_i - \frac{V}{R_i}\right)^2} \right) \tag{94}$$

The logarithm converts the outer product to a sum,

$$\mathcal{L}(V) = \sum_{i=1}^{N} \log \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\left(I_i - \frac{V}{R_i}\right)^2} \right). \tag{95}$$

The logarithm converts the inner product to a sum,

$$\mathcal{L}(V) = \sum_{i=1}^{N} \left(-\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \left(I_i - \tfrac{V}{R_i} \right)^2 \right). \tag{96}$$

Applying the sum separately to each term, and observing that $\sqrt{2\pi}\sigma$ *is independent of* i*,*

$$\mathcal{L}(V) = N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left((I_i - \tfrac{V}{R_i} \right)^2. \tag{97}$$

Now, we take the derivative wrt V*, and solve for the value of* V *where that gradient is zero,*

$$0 = \frac{\partial \mathcal{L}(V)}{\partial V}. \tag{98}$$

Substituting for $\mathcal{L}(V)$*,*

$$0 = \frac{\partial}{\partial V} \left[N \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left((I_i - \tfrac{V}{R_i} \right)^2 \right]. \tag{99}$$

The gradient of the first term is zero,

$$0 = \frac{\partial}{\partial V} \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left((I_i - \tfrac{V}{R_i} \right)^2 \right]. \tag{100}$$

Putting the gradient inside the sum,

$$0 = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \frac{\partial}{\partial V} \left((I_i - \tfrac{V}{R_i} \right)^2. \tag{101}$$

Expanding the bracket,

$$0 = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \frac{\partial}{\partial V} \left(I_i^2 - 2\tfrac{I_i}{R_i} V + \tfrac{1}{R_i^2} V^2 \right) \tag{102}$$

Computing the derivatives,

$$0 = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(-2\tfrac{I_i}{R_i} + \tfrac{2}{R_i^2} V \right) \tag{103}$$

Putting the sum inside the bracket,

$$0 = -\frac{1}{2\sigma^2} \left(-2 \sum_{i=1}^{N} \tfrac{I_i}{R_i} + 2 \left(\sum_{i=1}^{N} \tfrac{1}{R_i^2} \right) V \right) \tag{104}$$

Cancel the 2's, and multiply both sides by σ^2,

$$0 = \sum_{i=1}^{N} \frac{I_i}{R_i} - \left(\sum_{i=1}^{N} \frac{1}{R_i^2} \right) V \tag{105}$$

Add $\left(\sum_{i=1}^{N} \frac{1}{R_i^2} \right) V$ to both sides,

$$\left(\sum_{i=1}^{N} \frac{1}{R_i^2} \right) V = \sum_{i=1}^{N} \frac{I_i}{R_i} \tag{106}$$

Divide both sides by $\left(\sum_{i=1}^{N} \frac{1}{R_i^2} \right)$,

$$V = \frac{\sum_{i=1}^{N} \frac{I_i}{R_i}}{\sum_{i=1}^{N} \frac{1}{R_i^2}}. \tag{107}$$

Answer 3. *We start by finding the probability of the data,*

$$\mathrm{P}\left(x_1, \ldots, x_{10} | \theta\right) = \prod_{i=1}^{10} \mathrm{P}\left(x_i | \theta\right)$$

As there are two 0's, three 1's, three 2's and two 3's,

$$\mathrm{P}\left(x_1, \ldots, x_{10} | \theta\right) = \left(\frac{2\theta}{3} \right)^2 \left(\frac{\theta}{3} \right)^3 \left(\frac{2(1-\theta)}{3} \right)^3 \left(\frac{1-\theta}{3} \right)^2$$

$$\mathrm{P}\left(x_1, \ldots, x_{10} | \theta\right) = \frac{2^5}{3^{10}} \theta^5 \left(1-\theta\right)^5$$

Now, we take the logarithm,

$$\mathcal{L}(\theta) = \log \mathrm{P}\left(x_1, \ldots, x_{10} | \theta\right) \tag{108}$$

$$\mathcal{L}(\theta) = 5 \log 2 - 10 \log 3 + 5 \log \theta + 5 \log(1-\theta). \tag{109}$$

And solve for where the gradient is zero,

$$0 = \frac{\partial \mathcal{L}(\theta)}{\partial \theta} \tag{110}$$

Substitute for $\mathcal{L}(\theta)$,

$$0 = \frac{\partial}{\partial \theta} \left[5 \log 2 - 10 \log 3 + 5 \log \theta + 5 \log(1-\theta) \right]. \tag{111}$$

The first two terms are constants, so their gradients are zero,

$$0 = 5 \frac{\partial}{\partial \theta} \log \theta + 5 \frac{\partial}{\partial \theta} \log(1-\theta). \tag{112}$$

Now, we separately compute the two derivatives. The first derivative is a standard result,

$$\frac{\partial}{\partial\theta}\log\theta = \frac{1}{\theta}. \tag{113}$$

For the second derivative, we need the chain rule,

$$u = 1-\theta, \tag{114}$$
$$y = \log(1-\theta) = \log u. \tag{115}$$

Thus,

$$\begin{aligned}\frac{\partial}{\partial\theta}\log(1-\theta) &= \frac{\partial y}{\partial\theta} && (116)\\ &= \frac{\partial y}{\partial u}\frac{\partial u}{\partial\theta} && (117)\\ &= \frac{\partial\log u}{\partial u}\frac{\partial(1-\theta)}{\partial\theta} && (118)\\ &= \frac{1}{u}(-1) && (119)\\ &= -\frac{1}{1-\theta}. && (120)\end{aligned}$$

Substituting these derivatives,

$$0 = 5\tfrac{1}{\theta} - 5\tfrac{1}{1-\theta}. \tag{121}$$

Multiply both sides by $\theta(1-\theta)$,

$$0 = 5\tfrac{1}{\theta}\theta(1-\theta) - 5\tfrac{1}{1-\theta}\theta(1-\theta). \tag{122}$$

Cancel terms in the numerator and denominator,

$$0 = 5(1-\theta) - 5\theta. \tag{123}$$

Expand the bracket,

$$0 = 5 - 5\theta - 5\theta. \tag{124}$$

Combine terms,

$$0 = 5 - 10\theta. \tag{125}$$

Add 10θ to both sides,

$$10\theta = 5. \tag{126}$$

Divide both sides by 10

$$\theta = \tfrac{5}{10} = \tfrac{1}{2}. \tag{127}$$