Chapter 4

Maximum Likelihood

In the first part of this course, we introduced the idea of designing ML models around decision boundaries. Our goal was to find some way to separate our data points into different classes, either using an instance-based approach (i.e., k-NNs) or through a learned parametric decision boundary (i.e., perceptrons). However, building ML models solely based on decision boundaries has its limitations. Most prominently, decision boundaries cannot provide more nuanced notions of how *likely* it is that a point belongs to a certain class. Models like perceptrons can classify points, but they are unable to give probabilistic predictions and do not naturally incorporate uncertainty.

In this next part of the book, we will focus on another foundational approach to designing ML models: likelihood-based optimization. Likelihood-based approaches have probabilistic foundations, and are perhaps the most dominant methodology in ML. Instead of searching for decision boundaries, likelihood-based approaches learn by maximizing the probability (or the likelihood) of the training data under certain statistical assumptions. As we will see, these approaches still implicitly learn decision boundaries, but their probabilistic formulation opens up many design choices that are not accessible from a pure decision-boundary perspective.

In this chapter, we will begin with a basic overview of the idea of maximizing likelihood. The goal for this chapter is simply to optimize some model parameters to match the statistics of a simple dataset. We will ignore many details around supervised learning (e.g., having separate feature and target values) in this chapter, and focus on simple datasets involving univariate (i.e., one dimensional) data points. In the following chapters, we will build upon the ideas introduced here to design powerful, supervised learning models.

4.1 Learning about Probabilities

Estimating the probabilities of different events is at the core of statistical and machine learning. For example, suppose that we are trying to design a spam detection algorithm that learns from data. At the core of this task is the problem of estimating how *likely* an email is to be spam.

In this section, we will consider a simplified version of spam classification as a motivating problem: our goal is simply to estimate how likely a particular sender is to send out spam email. Assume that we have a collection of 10 emails from this sender, and we know that 7 of these emails were marked as spam by the receiver. Assuming no major changes in behaviour, how likely do we think that any given email from this sender is likely to be spam in the future?

The obvious answer in this example is that we expect a 70% chance (i.e., a probability of 0.7) that future emails from this sender are spam. After all, 70% of the emails in the past were spam. This kind of estimation is, in fact, the most basic form of likelihood-based optimization.

4.2 Maximum Likelihood for the Bernoulli

In the spam example from the previous section, we estimated a 70% chance that the sender's future emails would be spam, based on the fact that 70% of their previous emails were spam. How can we justify this estimation formally? In formal terms, we computed a *maximum likelihood* estimate for the probability of a *Bernoulli random variable.* We now describe these ideas in detail.

Defining the Bernoulli likelihood

First off, we assume that the event of the sender sending spam can be represented as a binary (i.e., $\{0, 1\}$) outcome, and we assume that we have a dataset of 10 previous emails, 7 of which were spam, i.e., a dataset

$$\mathcal{D} = \{1, 1, 1, 1, 1, 1, 1, 0, 0, 0\}$$

.

We will use y to denote a binary random variable, and binary outcomes of this kind are generally modelled via the Bernoulli distribution, which assumes that

$$P(y = 1) = \theta \tag{4.1}$$

for some $\theta \in [0, 1]$. In other words, the parameter θ encodes the probability that the random variable will be one. Note that we will often use the subscript P_θ to denote that the distribution is determined by the parameter θ.

We can define the *likelihood* of a set of datapoints $\mathcal{D}$ under a Bernoulli distribution with a particular θ as

$$\begin{aligned} \mathcal{L}(\theta, \mathcal{D}) &= \prod_{y_i \in \mathcal{D}} P_\theta(y_i) && (4.2) \\ &= \prod_{y_i \in \mathcal{D}} \theta^{y_i} (1 - \theta)^{(1 - y_i)}. && (4.3) \end{aligned}$$

The likelihood essentially tells us how probable or likely a set of points is, assuming a particular distribution (i.e., θ value). Note that we used a notational trick in Equation 4.2 that will often come up in the course, where we use exponents to capture different cases for a binary value:

$$
\begin{aligned}
P(y) &= \begin{cases} \theta & \text{if } y = 1 \\ 1-\theta & \text{if } y = 0 \end{cases} && (4.4) \\
&= \theta^{y}(1-\theta)^{(1-y)}. && (4.5)
\end{aligned}
$$

The important thing to note is that the exponents will either be one or zero, and raising something to the power of zero simply sets that term to one (i.e., ignores it). This trick is often used to allow us to write expressions in a single line that would otherwise involve tedious separation between cases.

Maximizing the likelihood

Now, our goal is to find a θ value that maximizes the likelihood of our dataset. We can do this by finding the maximum of the likelihood function using calculus. To find the maximum, we differentiate with respect to θ, set to 0, and solve.[1] However, one issue with directly differentiating Equation 4.2 is that it would involve repeated and cumbersome applications of the product rule. To get around this, it is standard practice to take the logarithm of the likelihood and differentiate this instead:

$$
\begin{aligned}
\frac{\partial \log \mathcal{L}(\theta, \mathcal{D})}{\partial \theta} &= \frac{\partial}{\partial \theta} \log \left(\prod_{y_i \in \mathcal{D}} \theta^{y_i}(1-\theta)^{(1-y_i)} \right) && (4.6) \\
&= \frac{\partial}{\partial \theta} \left(\sum_{y_i \in \mathcal{D}} y_i \log(\theta) + (1-y_i)\log(1-\theta) \right) && (4.7) \\
&= \sum_{y_i \in \mathcal{D}} \frac{y_i}{\theta} - \frac{1-y_i}{1-\theta} && (4.8) \\
&= \frac{Y}{\theta} + \frac{(Y-N)}{(1-\theta)} \qquad \text{where } Y = \sum_{y_i \in \mathcal{D}} y_i \text{ and } N = |\mathcal{D}|. && (4.9)
\end{aligned}
$$

Note that we can replace the likelihood with the log-likelihood because the logarithm is a *monotonic* function, which means that a θ maximizes the likelihood if and only if it also maximizes the log-likelihood. Generally, log-likelihoods are easier to work with because they involve summations, rather than products of probabilities.

[1] Technically, this only gives us an extremum of the function, which may be a maximum or minimum. You can verify for yourself using the second derivative test or other techniques that the extremum is indeed a maximum.

Finally, setting the log-likelihood to zero and solving, we get

$$\frac{Y}{\theta} + \frac{(Y-N)}{(1-\theta)} = 0 \tag{4.10}$$

$$(1-\theta)Y + \theta(Y-N) = 0 \tag{4.11}$$

$$-\theta N = -Y \tag{4.12}$$

$$\theta = \frac{Y}{N} \tag{4.13}$$

Thus, we have formally derived that the maximum likelihood estimate for θ is given by the proportion of positive points in our dataset. Again, this is intuitive. If 70% of emails were spam in the past, then a reasonable estimate is that there is a 70% chance of the next email being spam. However, the theory of maximum likelihood allows to to derive this fact in a formal fashion.

4.3 Maximum Likelihood for the Gaussian

Maximizing likelihood is a general idea and is not only relevant for simple binary processes. For example, we can also apply maximum likelihood to continuous-valued random variables. In this section, we discuss maximum likelihood for one of the simplest continuous-valued distributions: the Gaussian with fixed variance.

As a motivation, consider a variation of the email example, where instead of predicting whether an email is spam, we are instead predicting the size of the emails that a user sends (in terms of megabytes). Again, suppose we assume that we've received 10 emails from this user in the past, and we know the size of these emails:

$$\mathcal{D} = \{0.2, 0.5, 0.7, 1.2, 0.1, 0.4, 0.7, 0.8, 0.5, 1.2\}$$

What is our reasonable guess for how big the next email will be? An obvious guess would simply be the average size of the emails that this person has sent before, i.e.,

$$\frac{\sum_{y_i \in \mathcal{D}} y_i}{|\mathcal{D}|} = 0.63.$$

Just as with the Bernoulli case, we can justify this guess based on maximum likelihood.

Defining the Gaussian likelihood

A common assumption for continuous random variables is that they come from a Gaussian (or normal) distribution, defined by the density function

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}. \tag{4.14}$$

To keep things simple, we will consider the case where we assume that our distribution has a fixed variance $\sigma^2 = 1$. In other words, we assume that the mean μ is unknown but we assume that the distribution has unit variance. Under the assumption, we can write the likelihood of a dataset as

$$\mathcal{L}(\mu, \mathcal{D}) = \prod_{y_i \in \mathcal{D}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - \mu)^2}{2}} \tag{4.15}$$

with the log-likelihood given by

$$\log \mathcal{L}(\mu, \mathcal{D}) = \sum_{y_i \in \mathcal{D}} -\frac{(y_i - \mu)^2}{2}, \tag{4.16}$$

where we have omitted any constant terms that do not depend on μ. Just as in the Bernoulli case, the likelihood tells us how probable the data is, assuming a particular mean μ.[2]

Maximizing the Gaussian likelihood

Like the Bernoulli likelihood, the Gaussian likelihood can also be maximized by differentiating, setting to 0, and solving for μ:

$$\frac{\partial \log \mathcal{L}(\mu, \mathcal{D})}{\partial \mu} = 0 \tag{4.17}$$

$$\frac{\partial}{\partial \mu} \sum_{y_i \in \mathcal{D}} -\frac{(y_i - \mu)^2}{2} = 0 \tag{4.18}$$

$$\sum_{y_i \in \mathcal{D}} y_i - \mu = 0 \tag{4.19}$$

$$\Rightarrow$$

$$\mu = \frac{\sum_{y_i \in \mathcal{D}} y_i}{|\mathcal{D}|}. \tag{4.20}$$

And, again, we retrieve the expected answer: the maximum likelihood estimate for the mean is equal to the empirical average of the values in our dataset. Note that it is common practice to denote this estimated value with a caret superscript to indicate that it is an estimate, rather than the underlying true μ of the unknown distribution, i.e., we say

$$\hat{\mu} = \frac{\sum_{y_i \in \mathcal{D}} y_i}{|\mathcal{D}|}. \tag{4.21}$$

[2] However, note that unlike the Bernoulli case, we cannot interpret the Gaussian likelihood as a proper probability, since it is based on a continuous density function.

Relaxing the constant variance assumption

If we want to derive a more expressive estimate, we can relax the assumption that the variance of the Gaussian is a known constant. In this case, the log-likelihood is given by

$$\log \mathcal{L}(\mu, \sigma, \mathcal{D}) = \sum_{y_i \in \mathcal{D}} -\frac{(y_i - \mu)^2}{2\sigma^2} - \log\left(\sqrt{2\pi}\sigma\right). \tag{4.22}$$

We leave it as an exercise to verify that differentiating and solving this expression for its maximum with respect to μ gives the same mean estimate as Equation 4.21. Maximizing for σ, on the other hand, gives the uncorrected sample variance as the maximum likelihood estimate:

$$\hat{\sigma} = \sqrt{\frac{\sum_{y_i \in \mathcal{D}} (y_i - \hat{\mu})^2}{|\mathcal{D}|}}. \tag{4.23}$$

Note that—for reasons outside the scope of this course—people often use the bias-corrected estimate of the sample variance, rather than the raw maximum likelihood estimate:

$$\hat{\sigma} = \sqrt{\frac{\sum_{y_i \in \mathcal{D}} (y_i - \hat{\mu})^2}{|\mathcal{D}| - 1}}. \tag{4.24}$$