The maximum likelihood estimator (MLE) is the hypothesis \( h \) that maximizes the likelihood, \( p(\mathcal{D}|h) \), of a dataset \( \mathcal{D} \) over a set of possible hypotheses \( \mathcal{H} \),

\[
h_{\text{MLE}} = \arg \max_{h \in \mathcal{H}} p(\mathcal{D}|h).
\]

In supervised learning, we define \( \mathcal{D} \) as a set of ordered pairs \( \{(x_i, y_i)\}_{i=1}^{N} \), and we’ll use \( \mathcal{D}_i \) to denote ordered pair number \( i \), \( (x_i, y_i) \). If we assume these ordered pairs are independent and identically distributed (IID), we have

\[
h_{\text{MLE}} = \arg \max_{h \in \mathcal{H}} \prod_{i=1}^{N} p(\mathcal{D}_i|h).
\]

Generative classifiers model the probability \( p(y_i, x_i|h) \),

\[
p(\mathcal{D}_i|h) = p(y_i, x_i|h) = p(x_i|y_i, h)p(y_i|h).
\]

In naive Bayes, we assume that the variables \( x^1, x^2, \ldots, x^D \) are mutually conditionally independent given \( y \), which gives us

\[
p(x_i|y_i, h) = p(x^1_i|x^2:D_i, y_i, h)p(x^2:D_i|y_i, h) \\
= p(x^1_i|y_i, h)p(x^2:D_i|y_i, h) \\
= p(x^1_i|y_i, h)p(x^2|x^3:D_i, y_i, h)p(x^3:D_i|y_i, h) \\
= p(x^1_i|y_i, h)p(x^2|y_i, h)p(x^3:D_i|y_i, h) \\
= \prod_{j=1}^{D} p(x^j_i|y_i, h)
\]

We need to choose how we will define \( p(y_i|h) \) and \( p(x^j_i|y_i, h) \). If \( y \) is binary \( \{0, 1\} \), then it makes sense to use Bernoulli distributions (If we toss a coin that lands ‘heads’ with probability \( \theta \), we say that the distribution of \{heads,tails\} follows a Bernoulli distribution with parameter \( \theta \)).

We’ll use \( \theta \) as the parameter of the Bernoulli distribution for \( y \), so that \( y_i \) is distributed according to a Bernoulli random variable with parameter \( \theta \) (so we’ll have \( \theta \in h \), and \( h \) will also include the parameters of the other distributions we’ll use in the model), which we write as

\[
y_i \sim \text{Ber} (\theta),
\]

From the definition of a Bernoulli random variable, we have under this assumption that

\[
p(y_i|h) = p(y_i|\theta) = \theta^{I(y_i=1)}(1-\theta)^{I(y_i=0)}.
\]
If the $x^j_i$ are also binary, it still makes sense to use a Bernoulli distribution but we will have a different Bernoulli distribution depending on the value of the corresponding $y_i$. So for each variable $j$ we will have two parameters $\theta^j_1$ and $\theta^j_0$, and the value of $y_i$ decides which one we use,

$$x^j_i | y_i \sim \text{Ber}(\theta^j_{y_i}),$$

$$p(x^j_i | y_i, h) = p(x^j_i | y_i, \theta^j_{y_i}) = (\theta^j_{y_i})^{f(x^j_i = 1)} (1 - \theta^j_{y_i})^{f(x^j_i = 0)}.$$

To compute the MLE, for numerical reasons we typically work in the log-domain (taking the logarithm doesn’t change the argmax), and plugging in everything above we get

$$h_{\text{MLE}} = \arg \max_{h \in \mathcal{H}} p(D | h) \quad \text{(definition of MLE)}$$

$$= \arg \max_{h \in \mathcal{H}} \prod_{i=1}^N p(D_i | h) \quad \text{(IID assumption)}$$

$$= \arg \max_{h \in \mathcal{H}} \log \prod_{i=1}^N p(D_i | h) \quad \text{(log does not change optimal value)}$$

$$= \arg \max_{h \in \mathcal{H}} \sum_{i=1}^N \log p(D_i | h) \quad \text{(log turns multiplication in addition)}$$

$$= \arg \max_{h \in \mathcal{H}} \sum_{i=1}^N \log p(y_i, x_i | h) \quad \text{(definition of $D_i$)}$$

$$= \arg \max_{h \in \mathcal{H}} \sum_{i=1}^N \log(p(y_i | h) p(x_i | y_i, h)) \quad \text{(product rule)}$$

$$= \arg \max_{h \in \mathcal{H}} \sum_{i=1}^N \log p(y_i | h) + \log p(x_i | y_i, h) \quad \text{(log turns multiplication into addition)}$$

$$= \arg \max_{h \in \mathcal{H}} \sum_{i=1}^N \left[ \log p(y_i | h) + \sum_{j=1}^D \log p(x^j_i | y_i, h) \right] \quad \text{(naive Bayes assumption)}$$

$$= \arg \max_{\theta \in [0,1], \theta^j \in [0,1], \forall i \in \{1, 2, \ldots, D\}} \sum_{i=1}^N \left[ \log p(y_i | \theta) + \sum_{j=1}^D \log p(x^j_i | y_i, \theta^j_{y_i}) \right] \quad \text{(Bernoulli parameterization)}$$

Each term in this sum only depends on either $\theta$ or a single value of $\theta^j_1$ or $\theta^j_0$. This means we can solve for these parameters independently (in optimization, this is called a separable function). Let’s just concentrate
on the terms that depend on \( \theta \):

\[
\theta_{MLE} = \arg \max_{\theta \in [0, 1]} \sum_{i=1}^{N} \log p(y_i | \theta)
\]

\[
= \arg \max_{\theta \in [0, 1]} \sum_{i=1}^{N} \log(\theta I(y_i = 1)(1 - \theta)I(y_i = 0))
\]

\[
= \arg \max_{\theta \in [0, 1]} \sum_{i=1}^{N} [I(y_i = 1) \log(\theta) + I(y_i = 0) \log(1 - \theta)]
\]

\[
= \arg \max_{\theta \in [0, 1]} \log(\theta) N_{1} + \log(1 - \theta) N_{0},
\]

where \( N_{1} \) is the number of times \( y_i = 1 \) in the training data and \( N_{0} \) is the number of times \( y_i = 0 \). We will be able to prove this with tools we develop later, but right now I will claim that there is one stationary point of the log-likelihood in terms of \( \theta \) in the interval \([0, 1]\) and that this is a maximizer. To find this stationary point, take the derivative and set it to 0,

\[
0 = \frac{N_{1}}{\theta} - \frac{N_{0}}{1 - \theta}.
\]

Re-arrange this to get

\[
\frac{\theta}{1 - \theta} = \frac{N_{1}}{N_{0}} = \frac{N_{1}/N}{N_{0}/N}.
\]

The solution to this (within \([0, 1]\)) is

\[
\theta = \frac{N_{1}}{N_{1} + N_{0}} = \frac{N_{1}}{N},
\]

to see this observe that \( 1 - \theta = 1 - N_{1}/N = N_{0}/N \).

This is an overly complicated way to say that if you flip a coin 100 times and it lands heads 40 times, then if you have no prior knowledge your most likely guess for the probability that it will land heads is 40/100.

The general solution when \( y_i \in \{1, 2, \ldots, C\} \) and we have parameters \( \{\theta_1, \theta_2, \ldots, \theta_C\} \) is

\[
\theta_{c} = \frac{N_{c}}{N}.
\]

For a binary \( x_i \) conditioned on these \( y_i \), you get

\[
\theta_{c}^{j} = \frac{N_{c1}^{j}}{N_{c}},
\]

where \( N_{c1}^{j} \) is the number of times variable \( x_i^{j} = 1 \) and \( y_i = c \).