10-315 Notes
Maximum Likelihood Estimation

Carnegie Mellon University
Machine Learning Department

Contents

1 Distributions

We'll focus on just a few key distributions: Bernoulli, categorical, Gaussian, and multivariate Gaussian. The first two are discrete distributions and the Gaussian distributions are continuous.

1.1 Bernoulli

The Bernoulli distribution is a discrete distribution with two possible outcomes, 0 and 1. The probability of 1 is denoted by ϕ and the probability of 0 is $1 - \phi$. (So, yeah, it's pretty boring, but of course, pretty useful.)

Canonical Example	Single flip of a (potentially biased) coin
Random Variable	Y, discrete, binary
Parameters	$\phi \in [0, 1]$
Probability, p.m.f.	$P(Y = 1 \mid \phi) = \phi$ $P(Y = 0 \mid \phi) = 1 - \phi$ $p(y \mid \phi) = \begin{cases} \phi & y = 1 \\ 1 - \phi & y = 0 \end{cases}$

1.2 Categorical

The categorical distribution is a discrete distribution with K possible values (often corresponding to K discrete events). This is not a very standard distribution from a statistical standpoint, but we use it *all the time* in machine learning to represent the probability of something belonging to one of K classes. In this course, we'll represent this distribution as a one-hot vector of K random binary random variables, $Y = [Y_1, Y_2, \ldots, Y_K]^\top$, where Y_k takes on the value 1 if the outcome is in class k and 0 otherwise. The probability of each Y_k being one is denoted by ϕ_k, and all of the ϕ_k's sum to 1 (e.g., because the event must belong to one of the K classes).

Canonical Example	Single roll of a (potentially weighted) K-sided die
Random Variable	Y, vector of K binary random variables Y_k
Parameters	$\boldsymbol{\phi} = [\phi_1, \phi_2, \ldots, \phi_K]^\top$, where $\phi_k \in [0, 1]$ and $\sum_{k=1}^{K} \phi_k = 1$
Probability, p.m.f.	$P(Y_k = 1 \mid \boldsymbol{\phi}) = \phi_k$ $p(y_k \mid \boldsymbol{\phi}) = \phi_k$

1.3 Gaussian

The Gaussian, or normal, distribution is a continuous distribution with two parameters: one for the mean, μ, and one for the variance, σ^2 (or equivalently, the standard deviation, σ). The Gaussian distribution is our go-to distribution for representing continuous random variables. It is known for its signature bell curve shape, where the likelihood of a value being close to the mean is much higher than the likelihood of a value being far from the mean. The Gaussian distribution is often denoted by $\mathcal{N}(\mu, \sigma^2)$.

Canonical Example	Height of a person
Random Variable	Y, continuous
Parameters	$\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$
Density, p.d.f.	$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)}$

1.4 Multivariate Gaussian

The multivariate Gaussian is a generalization of the Gaussian distribution for the case where the random variable has a dimension M rather than being restricted to 1-D. The parameters for the multivariate Gaussian similarly need to be in higher dimensions. The mean is now an M-dimensional vector, $\boldsymbol{\mu}$, and the variance is now an $M \times M$ covariance matrix, $\boldsymbol{\Sigma}$. Like the univariate Gaussian distribution, the multivariate Gaussian is also denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The classic bell curve p.d.f for dimension $M = 2$ now looks a bit like a hat.

Canonical Example	Height and weight of a person
Random Variable	M-dimensional vector $X \in \mathbb{R}^M$
Parameters	$\boldsymbol{\mu} = [\mu_1, \ldots, \mu_M]^\top \in \mathbb{R}^M$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{M \times M}$
Density, p.d.f.	$P(X = \mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ $p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}\lvert\boldsymbol{\Sigma}\rvert^{1/2}} e^{\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}$

2 Likelihood

Likelihood:

$$p(\mathcal{D} \mid \theta) = \prod_{i=1}^{N} p(y^{(i)} \mid \theta) \tag{1}$$

The likelihood is the probability (or density) of the data given the parameters of the model, $p(\mathcal{D} \mid \theta)$. We use θ generically represents the parameter(s) of the distribution that we are interested in, and we use $\mathcal{D}$ to represent the set of N observed values $\{y^{(i)}\}_{i=1}^N$ of the random variable Y.

By default, we assume that each $y^{(i)}$ is *i.i.d.* or *independent and identically distributed*, meaning that each $y^{(i)}$ is drawn from the same distribution and drawn independently from every other sample $y^{(n)}$, $n \neq i$. Because of this *i.i.d.* assumption, we can write the likelihood as:

$$\begin{aligned} p(\mathcal{D} \mid \theta) &= p(y^{(1)}, y^{(2)}, \ldots, y^{(N)} \mid \theta) && (2)\\ &= p(y^{(1)} \mid \theta)\ p(y^{(2)} \mid \theta)\ p(y^{(3)} \mid \theta) \ldots p(y^{(N)} \mid \theta) && (3)\\ &= \prod_{i=1}^{N} p(y^{(i)} \mid \theta) && (4) \end{aligned}$$

Note: If we didn't have this assumption, then we would have N different random variables, $Y^{(1)}, Y^{(2)}, \ldots, Y^{(N)}$ with N different parameters $\theta_1, \theta_2, \ldots, \theta_N$ (not identical) and we couldn't just multiply the individual probabilities/densities together, and we'd be stuck with each outcome depending on all previous

outcomes:

$$p(\mathcal{D} \mid \theta) = p(y^{(1)}, y^{(2)}, \ldots, y^{(N)} \mid \theta_1, \theta_2, \ldots, \theta_N) \tag{5}$$

$$= p(y^{(1)} \mid \theta_1)\ p(y^{(2)} \mid y^{(1)}, \theta_2)\ p(y^{(3)} \mid y^{(1)}, y^{(2)}, \theta_3) \ldots p(y^{(N)} \mid y^{(1)}, y^{(2)}, \ldots, y^{(N-1)}, \theta_N) \tag{6}$$

2.1 Examples

2.1.1 Bernoulli: Flipping a coin with sides heads and tails

Consider a Bernoulli random variable $Y \in \{0, 1\}$ with parameter ϕ where $Y = 1$ represents the value of a coin flip coming up heads and $Y = 0$ represents the value of a coin flip coming up tails.

Suppose we have the dataset $\mathcal{D} = \{1, 1, 0, 1\}$ corresponding to outcomes $\{heads, heads, tails, heads\}$.

Notational note: For Bernoulli and categorical random variables in machine learning, events often correspond to one and only one value, e.g. heads is 1 and tails is 0. This leads us to the notational shortcut of writing $P(Y = 1 \mid \phi)$ as just $p(heads \mid \phi)$.

We can write the likelihood as:

$$p(\mathcal{D} \mid \theta) = p\left(y^{(1)}, y^{(2)}, y^{(3)}, y^{(4)} \mid \phi\right) \tag{7}$$

$$= p(heads, heads, tails, heads \mid \phi) \tag{8}$$

$$= p(heads \mid \phi)\ p(heads \mid \phi)\ p(tails \mid \phi)\ p(heads \mid \phi) \tag{9}$$

$$= \phi\ \phi\ \ (1 - \phi)\ \ \phi \tag{10}$$

$$= \phi^3 (1 - \phi)^1 \tag{11}$$

For a fair coin, where $\phi = 1/2$, the likelihood is:

$$p(\mathcal{D} \mid \theta) = \phi\ \phi\ \ (1 - \phi)\ \ \phi \tag{12}$$

$$= \frac{1}{2}\ \frac{1}{2}\ \left(1 - \frac{1}{2}\right)\ \frac{1}{2} \tag{13}$$

$$= \frac{1}{8} \tag{14}$$

For an biased coin, where $\phi = 3/10$, the likelihood is:

$$p(\mathcal{D} \mid \theta) = \phi\ \phi\ \ (1 - \phi)\ \ \phi \tag{15}$$

$$= \frac{3}{10}\ \frac{3}{10}\ \left(1 - \frac{3}{10}\right)\ \frac{3}{10} \tag{16}$$

$$= \frac{189}{10000} \tag{17}$$

2.1.2 Categorical: Rolling a four-sided die with sides *red*, *green*, *blue*, and *magenta*

Consider a categorical random variable $Y = [Y_1, Y_2, Y_3, Y_4]^\top$ with parameters $\boldsymbol{\phi} = [\phi_1, \phi_2, \phi_3, \phi_4]^\top$ where $Y_1 = 1$ represents the value of a die roll coming up *red* and $Y_1 = 0$ represents the value of a die roll not coming up *red*. Similarly for Y_2, Y_3, Y_4 and *green*, *blue*, *magenta* respectively.

Notational note: Again, we'll use the notational shortcut that combines events with values for random variables. Specifically, we'll write our dataset and probabilities in terms of the colors of the die rolls rather than the one-hot encoded values of the random variable, e.g. *green* instead of $[0, 1, 0, 0]^\top$.

Suppose we have the dataset $\mathcal{D} = \{green, magenta, magenta, red, green\}$. We can write the likelihood as:

$$p(\mathcal{D} \mid \theta) = p\left(y^{(1)}, y^{(2)}, y^{(3)}, y^{(4)}, y^{(5)} \mid \phi\right) \tag{18}$$

$$= p(green, magenta, magenta, red, green \mid \boldsymbol{\phi}) \tag{19}$$

$$= p(green \mid \boldsymbol{\phi})\ p(magenta \mid \boldsymbol{\phi})\ p(magenta \mid \boldsymbol{\phi})\ p(red \mid \boldsymbol{\phi})\ p(green \mid \boldsymbol{\phi}) \tag{20}$$

$$= P\left(Y = \begin{bmatrix}0\\1\\0\\0\end{bmatrix} \mid \boldsymbol{\phi}\right) P\left(Y = \begin{bmatrix}0\\0\\0\\1\end{bmatrix} \mid \boldsymbol{\phi}\right) P\left(Y = \begin{bmatrix}0\\0\\0\\1\end{bmatrix} \mid \boldsymbol{\phi}\right) P\left(Y = \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \mid \boldsymbol{\phi}\right) P\left(Y = \begin{bmatrix}0\\1\\0\\0\end{bmatrix} \mid \boldsymbol{\phi}\right) \tag{21}$$

$$= P(Y_2 = 1 \mid \boldsymbol{\phi})\ P(Y_4 = 1 \mid \boldsymbol{\phi})\ P(Y_4 = 1 \mid \boldsymbol{\phi})\ P(Y_1 = 1 \mid \boldsymbol{\phi})\ P(Y_2 = 1 \mid \boldsymbol{\phi}) \tag{22}$$

$$= \phi_2\ \phi_4\ \phi_4\ \phi_1\ \phi_2 \tag{23}$$

$$= \phi_1\ \phi_2^2\ \phi_4^2 \tag{24}$$

For an evenly weighted die, where $\boldsymbol{\phi} = [1/4, 1/4, 1/4, 1/4]^\top$, the likelihood is:

$$p(\mathcal{D} \mid \theta) = \phi_2\ \phi_4\ \phi_4\ \phi_1\ \phi_2 \tag{25}$$

$$= \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4} \tag{26}$$

$$= \frac{1}{1024} \tag{27}$$

For a weighted die with $\boldsymbol{\phi} = [2/10, 1/10, 4/10, 3/10]^\top$, the likelihood is:

$$p(\mathcal{D} \mid \theta) = \phi_2\ \phi_4\ \phi_4\ \phi_1\ \phi_2 \tag{28}$$

$$= \frac{1}{10}\frac{3}{10}\frac{3}{10}\frac{2}{10}\frac{1}{10} \tag{29}$$

$$= \frac{18}{100000} \tag{30}$$

2.1.3 Gaussian: Grades on a test

Consider a Gaussian random variable Y with parameters μ and σ^2 where Y represents the grade on a test.

Suppose we have the dataset $\mathcal{D} = \{75, 95, 80\}$. We can write the likelihood as:

$$p(\mathcal{D} \mid \theta) = p\left(y^{(1)}, y^{(2)}, y^{(3)} \mid \mu, \sigma^2\right) \tag{31}$$

$$= p(75, 95, 80 \mid \mu, \sigma^2) \tag{32}$$

$$= p(75 \mid \mu, \sigma^2)\ p(80 \mid \mu, \sigma^2)\ p(95 \mid \mu, \sigma^2) \tag{33}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(75-\mu)^2}{2\sigma^2}}\ \ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(95-\mu)^2}{2\sigma^2}}\ \ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(80-\mu)^2}{2\sigma^2}} \tag{34}$$

For parameters $\mu = 80$ and $\sigma^2 = 5^2$, the likelihood is:

$$p(\mathcal{D} \mid \theta) = p(75 \mid \mu, \sigma^2)\ p(80 \mid \mu, \sigma^2)\ p(95 \mid \mu, \sigma^2) \tag{35}$$

$$= 0.0252834 \cdot 0.0000000157990 \cdot 0.150786 \tag{36}$$

$$= 6.02316\text{e}{-11} \tag{37}$$

For parameters $\mu = 85$ and $\sigma^2 = 10^2$, the likelihood is:

$$p(\mathcal{D} \mid \theta) = p(75 \mid \mu, \sigma^2)\ p(80 \mid \mu, \sigma^2)\ p(95 \mid \mu, \sigma^2) \tag{38}$$

$$= 0.000272286 \cdot 0.000272286 \cdot 0.0295651 \tag{39}$$

$$= 2.19194\text{e}{-9} \tag{40}$$

Important note: Notice that the likelihood values above are really small despite fairly reasonable parameters. These values will very quickly underflow and won't be able to compare likelihood values to figure out which parameters are best. We'll come back to this later, but this is one of the two main reasons why we use log-likelihoods.

3 Maximum Likelihood Estimation

The above examples for likelihood show that for a given set of parameters θ, we can compute the likelihood of the data $\mathcal{D}$. While it's definitely important to understand how to compute a likelihood value, we actually aren't terribly interested in the likelihood value itself.

For example, knowing that the likelihood $p(\mathcal{D} \mid \theta)$ is 6.02316e−11 for parameters $\mu = 80$, $\sigma^2 = 5^2$ doesn't tell us much. BUT... also knowing that $p(\mathcal{D} \mid \theta)$ is 2.19194e−9 for parameters $\mu = 85$, $\sigma^2 = 10^2$ tells us that the parameters $\mu = 85$, $\sigma^2 = 10^2$ are better than the parameters $\mu = 80$, $\sigma^2 = 5^2$ for this dataset!

Maximum likelihood estimation (MLE) is trying to find the best parameters for a specific dataset, $\mathcal{D}$. Specifically, we want to find the parameters $\hat{\theta}_{MLE}$ that maximize the likelihood for $\mathcal{D}$.

Caution: This may seem like we are starting to flip $p(\mathcal{D} \mid \theta)$ to $p(\theta \mid \mathcal{D})$ but we are *not*. The likelihood is still the probability (or density) of the data given the parameters. We are just searching for the parameter values that maximize this.

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\ p(\mathcal{D} \mid \theta) \tag{41}$$

3.1 Likelihood function

To emphasize that we are trying to optimize over the parameters, we introduce the **likelihood function**, $\mathcal{L}(\theta; \mathcal{D})$, also seen written as $\mathcal{L}(\theta \mid \mathcal{D})$ and $L(\theta; \mathcal{D})$.

The likelihood function is the SAME as the likelihood. We just write it differently because maximum likelihood estimation is optimizing a function of the parameters.

$$\mathcal{L}(\theta; \mathcal{D}) = p(\mathcal{D} \mid \theta) \tag{42}$$

$$\begin{aligned}
\hat{\theta}_{MLE} &= \underset{\theta}{\operatorname{argmax}}\ \mathcal{L}(\theta; \mathcal{D}) && (43)\\
&= \underset{\theta}{\operatorname{argmax}}\ p(\mathcal{D} \mid \theta) && (44)\\
&= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p\left(y^{(i)} \mid \theta\right) \quad \text{Assuming } \textit{i.i.d.} && (45)
\end{aligned}$$

Notational note: We'll often drop the $\mathcal{D}$ from the likelihood function and the parameter(s) θ from the probabilities/densities when it is (hopefully) clear that they are implied:

$$\begin{aligned}
\hat{\theta}_{MLE} &= \underset{\theta}{\operatorname{argmax}}\ \mathcal{L}(\theta) && (46)\\
&= \underset{\theta}{\operatorname{argmax}}\ p(\mathcal{D} \mid \theta) && (47)\\
&= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p\left(y^{(i)}\right) \quad \text{Assuming } \textit{i.i.d.} && (48)
\end{aligned}$$

3.2 Log-likelihood

When doing MLE, we typically use the log of the likelihood rather than the likelihood itself. We denote the log-likelihood function as:

$$\ell(\theta; \mathcal{D}) = \log \mathcal{L}(\theta; \mathcal{D}) = \log p(\mathcal{D} \mid \theta) \tag{49}$$

We do use the log-likelihood for two main reasons: to avoid underflow and to simplify the calculus for optimization.

Avoiding underflow: As we saw with some of the numerical examples in the previous section, with just a few data points, the likelihood values can become extremely small very quickly, for example, 6.02316e−11 for just three points in a Gaussian model. These will just keep getting smaller and smaller as we add more data points. Mathematically, this isn't an issue, but when working with computers, these values will underflow.

Simplifying calculus: As we'll soon see, to find the parameters that minimize the likelihood, we'll take the gradient with respect to the parameters and either set it equal to zero or use some form of gradient descent. Because we have many *i.i.d.* data points, the likelihood is product of N factors. Products are a pain with calculus, but applying a log nicely converts products to sums:

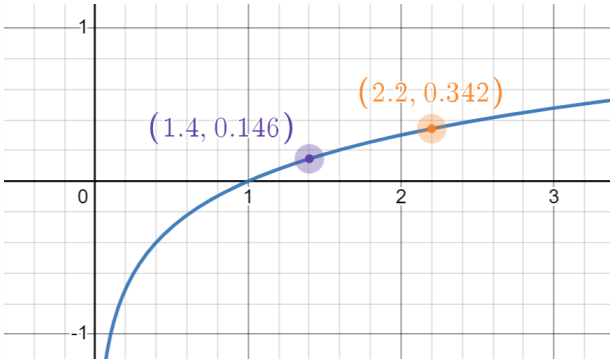
$$\ell(\theta; \mathcal{D}) = \log p(\mathcal{D} \mid \theta) \tag{50}$$

$$= \log \prod_{i=1}^{N} p\left(y^{(i)} \mid \theta\right) \quad \text{Assuming } i.i.d. \tag{51}$$

$$= \sum_{i=1}^{N} \log p\left(y^{(i)} \mid \theta\right) \tag{52}$$

The log can also simplify exponential family distributions such as the Gaussian distribution, effectively removing the exponential. We'll definitely see this more later.

We can can get away with taking the log during MLE because the log function is monotonically increasing, so if $x_2 > x_1$ then $\log(x_2) > \log(x_1)$:



The log function is monotonically increasing, see interactiveDesmos demo

Specifically, if $p(\mathcal{D} \mid \theta_B) > p(\mathcal{D} \mid \theta_A)$ then $\log p(\mathcal{D} \mid \theta_B) > \log p(\mathcal{D} \mid \theta_A)$ and:

$$\hat{\theta}_{MLE} = \operatorname*{argmax}_{\theta} p(\mathcal{D} \mid \theta) \tag{53}$$

$$= \operatorname*{argmax}_{\theta} \log p(\mathcal{D} \mid \theta) \tag{54}$$

(even though $p(\mathcal{D} \mid \theta)$ and $\log p(\mathcal{D} \mid \theta)$ are not equal).

3.3 Negative log-likelihood

Just so we can continue to use gradient *descent* rather than changing to gradient *ascent*, we negate the log-likelihood and then flip the argmax to an argmin:

$$\hat{\theta}_{MLE} = \underset{\theta}{\text{argmax}}\ \ell(\theta; \mathcal{D}) \tag{55}$$

$$= \underset{\theta}{\text{argmin}}\ -\ell(\theta; \mathcal{D}) \tag{56}$$

3.4 Optimization for MLE

Generally speaking, the recipe for doing MLE is to:

1. Formulate the likelihood, $p(\mathcal{D} \mid \theta)$
2. Set the objective $J(\theta)$ equal to the negative log-likelihood, $J(\theta) = -\log p(\mathcal{D} \mid \theta)$
3. Compute the (partial) derivative $\partial J / \partial \theta$
4. Find $\hat{\theta}_{MLE}$ by either
   (a) Setting derivative equal to zero and solve for θ or
   (b) Using a form of gradient descent to step towards better θ values

Note: Steps 3 and 4 are exactly what we've been doing for linear regression, logistic regression, and neural networks.

3.5 Examples

3.5.1 Notation

Before we get started with examples, let's go over notational tricks that we'll be using with the likelihood for Bernoulli and categorical distributions.

Bernoulli:

In our earlier examples in the likelihood section, it was easy to convert the likelihood factors, $p(y^{(i)} \mid \phi)$, into either ϕ or $1 - \phi$ because we had specific values for each $y^{(i)}$. When we just have $Y = y^{(i)}$ rather than $Y = 1$ or $Y = 0$, we need a notational trick to select the correct factor. For this trick, we'll use a combination of zero and one exponents and the indicator function.

For the i-th factor corresponding to the i-th data point, we can write:

$$P\left(Y = y^{(i)} \mid \phi\right) = \phi^{\mathbb{I}\left(y^{(i)}=1\right)}(1-\phi)^{\mathbb{I}\left(y^{(i)}=0\right)} \tag{57}$$

This will then result in $P\left(Y = y^{(i)} \mid \phi\right) = \phi$ if $y^{(i)} = 1$ and $P\left(Y = y^{(i)} \mid \phi\right) = 1 - \phi$ if $y^{(i)} = 0$.

Then, inserting this into the Bernoulli likelihood, we have:

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^{N} p(y^{(i)} \mid \phi) \tag{58}$$

$$= \prod_{i=1}^{N} \phi^{\mathbb{I}\left(y^{(i)}=1\right)}(1-\phi)^{\mathbb{I}\left(y^{(i)}=0\right)} \tag{59}$$

Categorical:

The trick for categorical is very similar. The main difference is that instead of ϕ and $1-\phi$, we just have a ϕ_k for each of the K classes and then $P(Y_k = 1) = \phi_k$. Also, recall that each data point is a one-hot vector of length K, $\mathbf{y}^{(i)}$.

For the i-th factor corresponding to the i-th data point, we can write:

$$p\left(\mathbf{y}^{(i)} \mid \boldsymbol{\phi}\right) = \prod_{k=1}^{K} \phi_k^{\mathbb{I}\left(y_k^{(i)}=1\right)} \tag{60}$$

This will then result in $p\left(\mathbf{y}^{(i)} \mid \boldsymbol{\phi}\right) = \phi_k$ when the i-th data point belongs to the k-th class.

Then, inserting this into the categorical likelihood, we have:

$$\mathcal{L}(\theta; \mathcal{D} = \prod_{i=1}^{N} p\left(\mathbf{y}^{(i)} \mid \boldsymbol{\phi}\right) \tag{61}$$

$$= \prod_{i=1}^{N} \prod_{k=1}^{K} \phi_k^{\mathbb{I}\left(y_k^{(i)}=1\right)} \tag{62}$$

Note: Before continuing, it's worth taking a few minutes to connect these notation tricks to the numerical likelihood examples that we had earlier.

For example, make sure to understand the two nested products in the context of the categorical example for the four-sided die and dataset $\mathcal{D} = \{green, magenta, magenta, red, green\}$:

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^{N} p\left(\mathbf{y}^{(i)} \mid \boldsymbol{\phi}\right) \tag{63}$$

$$= \prod_{i=1}^{N} \prod_{k=1}^{K} \phi_k^{\mathbb{I}\left(y_k^{(i)}=1\right)} \tag{64}$$

$$= \prod_{i=1}^{N} \left(\phi_1^{\mathbb{I}\left(y_1^{(i)}=1\right)} \phi_2^{\mathbb{I}\left(y_2^{(i)}=1\right)} \phi_3^{\mathbb{I}\left(y_3^{(i)}=1\right)} \phi_4^{\mathbb{I}\left(y_4^{(i)}=1\right)} \right) \tag{65}$$

$$= \left(\phi_1^0\ \phi_2^1\ \phi_3^0\ \phi_4^0\right) \left(\phi_1^0\ \phi_2^0\ \phi_3^0\ \phi_4^1\right) \left(\phi_1^0\ \phi_2^0\ \phi_3^0\ \phi_4^1\right) \left(\phi_1^1\ \phi_2^0\ \phi_3^0\ \phi_4^0\right) \left(\phi_1^0\ \phi_2^1\ \phi_3^0\ \phi_4^0\right) \tag{66}$$

$$= \phi_2\ \phi_4\ \phi_4\ \phi_1\ \phi_2 \tag{67}$$

3.5.2 Bernoulli

Suppose we are flipping a biased coin. We can consider each flip to be a Bernoulli(ϕ) random variable, Y, where $Y = 1$ represents the coin coming up heads and $Y = 0$ represents tails, and $P(Y = 1) = \phi$ and $P(Y = 0) = 1 - \phi$

Suppose our dataset $\mathcal{D}$ contains N_H heads and N_T tails.

Formulating the likelihood and the log-likelihood:

$$\mathcal{L}(\phi; \mathcal{D}) = \prod_{i=1}^{N} \phi^{\mathbb{I}\left(y^{(i)}=1\right)} (1-\phi)^{\mathbb{I}\left(y^{(i)}=0\right)} \tag{68}$$

$$\ell(\phi; \mathcal{D}) = \log \prod_{i=1}^{N} \phi^{\mathbb{I}\left(y^{(i)}=1\right)} (1-\phi)^{\mathbb{I}\left(y^{(i)}=0\right)} \tag{69}$$

$$= \sum_{i=1}^{N} \left[\log \phi^{\mathbb{I}\left(y^{(i)}=1\right)} + \log(1-\phi)^{\mathbb{I}\left(y^{(i)}=0\right)} \right] \tag{70}$$

$$= \sum_{i=1}^{N} \left[\mathbb{I}(y^{(i)}=1) \log(\phi) + \mathbb{I}(y^{(i)}=0) \log(1-\phi) \right] \tag{71}$$

$$= \log(\phi) \sum_{i=1}^{N} \mathbb{I}(y^{(i)}=1) + \log(1-\phi) \sum_{i=1}^{N} \mathbb{I}(y^{(i)}=0) \tag{72}$$

$$= N_H \log(\phi) + N_T \log(1-\phi) \tag{73}$$

Now that we have the log-likelihood function, our $\hat{\phi}_{MLE}$ is the value that maximizes the log-likelihood. For Bernoulli, we can take the derivative of our objective function and set it to 0. (Note that in this closed-form case, it doesn't really matter if negate the log-likelihood or not as we're just setting it to zero.)

$$\frac{d\ell}{d\phi} = N_H \frac{1}{\phi} - N_T \frac{1}{1-\phi} \tag{74}$$

$$= \frac{(1-\phi)N_H - \phi N_T}{\phi(1-\phi)} \tag{75}$$

$$\frac{(1-\phi)N_H - \phi N_T}{\phi(1-\phi)} = 0 \tag{76}$$

$$(1-\phi)N_H - \phi N_T = 0 \tag{77}$$

$$N_H - \phi(N_H + N_T) = 0 \tag{78}$$

$$\phi = \frac{N_H}{N_H + N_T} \tag{79}$$

Therefore, $\hat{\phi}_{MLE} = \frac{N_H}{N_H+N_T} = \frac{N_H}{N}$, which should match our intuitive estimate of the fraction of times heads appears in the data.

3.5.3 Categorical

Suppose we are rolling a (weighted) four-sided die with sides *red*, *green*, *blue*, and *magenta*. We can consider each roll to be a categorical random variable $Y = [Y_1, Y_2, Y_3, Y_4]^\top$ with parameters $\boldsymbol{\phi} = [\phi_1, \phi_2, \phi_3, \phi_4]^\top$ corresponding to *red*, *green*, *blue*, *magenta*, respectively.

Suppose our dataset $\mathcal{D}$ contains N_1 red, N_2 green, N_3 blue, and N_4 magenta values and the total number of points is $N = N_1 + N_2 + N_3 + N_4$.

Formulating the likelihood and the log-likelihood:

$$\mathcal{L}(\boldsymbol{\phi}; \mathcal{D}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \phi_k^{\mathbb{I}\left(y_k^{(i)}=1\right)} \tag{80}$$

$$\ell(\phi; \mathcal{D}) = \log \prod_{i=1}^{N} \prod_{k=1}^{K} \phi_k^{\mathbb{I}\left(y_k^{(i)}=1\right)} \tag{81}$$

$$= \sum_{i=1}^{N} \log \prod_{k=1}^{K} \phi_k^{\mathbb{I}\left(y_k^{(i)}=1\right)} \tag{82}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \log \phi_k^{\mathbb{I}\left(y_k^{(i)}=1\right)} \tag{83}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}\left(y_k^{(i)} = 1\right) \log \phi_k \tag{84}$$

$$= \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbb{I}\left(y_k^{(i)} = 1\right) \log \phi_k \tag{85}$$

$$= \sum_{k=1}^{K} \left[\sum_{i=1}^{N} \mathbb{I}\left(y_k^{(i)} = 1\right) \right] \log \phi_k \tag{86}$$

$$= \sum_{k=1}^{K} N_k \log \phi_k \tag{87}$$

$$= N_1 \log \phi_1 + N_2 \log \phi_2 + N_3 \log \phi_3 + N_4 \log \phi_4 \tag{88}$$

We're going to hold off on solving for the categorical MLE as it involves one more trick related to the fact that $\sum_k \phi_k = 1$. But that trick does lead us to $\hat{\phi}_{k,MLE} = \frac{N_k}{N}$, which should match our intuitive estimate of the fraction of times class k appears in the data.

3.5.4 Gaussian

Suppose we observe N samples from a Gaussian distribution, $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$, where $x^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma)$. Recall, that for a 1-D Gaussian distribution, the pdf is:

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The likelihood and log-likelihood are:

$$\mathcal{L}(\mu, \sigma; \mathcal{D}) = \prod_{i=1}^{N} p(x \mid \mu, \sigma) \tag{89}$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x^{(i)} - \mu)^2}{2\sigma^2}\right) \tag{90}$$

$$\ell(\mu, \sigma; \mathcal{D}) = \log \prod_{i=1}^{N} p(x \mid \mu, \sigma) \tag{91}$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x^{(i)} - \mu)^2}{2\sigma^2}\right) \tag{92}$$

$$= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x^{(i)} - \mu)^2}{2\sigma^2}\right) \tag{93}$$

$$= \sum_{i=1}^{N} -\log(\sqrt{2\pi}) - \log(\sigma) - \frac{(x^{(i)} - \mu)^2}{2\sigma^2} \tag{94}$$

$$= -N \log(\sqrt{2\pi}) - N \log(\sigma) - \sum_{i=1}^{N} \frac{(x^{(i)} - \mu)^2}{2\sigma^2} \tag{95}$$

We can then use the log-likelihood to determine the MLE for μ:

$$\ell(\mu, \sigma; \mathcal{D}) = -N \log(\sqrt{2\pi}) - N \log(\sigma) - \sum_{i=1}^{N} \frac{(x^{(i)} - \mu)^2}{2\sigma^2} \tag{96}$$

$$\frac{\partial \ell}{\partial \mu} = \frac{\partial}{\partial \mu} \sum_{i=1}^{N} -\frac{(x^{(i)} - \mu)^2}{2\sigma^2} \tag{97}$$

$$= \sum_{i=1}^{N} \frac{1}{2\sigma^2} 2(x^{(i)} - \mu) \tag{98}$$

$$\sum_{i=1}^{N} \frac{1}{2\sigma^2} 2(x^{(i)} - \mu) = 0 \tag{99}$$

$$\sum_{i=1}^{N} x^{(i)} - \mu = 0 \tag{100}$$

$$\sum_{i=1}^{N} \mu = \sum_{i=1}^{N} x^{(i)} \tag{101}$$

$$N\mu = \sum_{i=1}^{N} x^{(i)} \tag{102}$$

$$\mu = \frac{\sum_{i=1}^{N} x^{(i)}}{N} \tag{103}$$

Hence $\hat{\mu} = \frac{\sum_{i=1}^{N} x^{(i)}}{N}$, which is the average as we might expect.

4 MLE for Logistic Regression

So far, we have been working with the likelihood of data related to a random variable given its parameters, $p(\mathcal{D} \mid \theta)$, e.g. $p(x^{(1)}, ..., x^{(N)} \mid \theta)$ or $p(y^{(1)}, ..., y^{(N)} \mid \theta)$. However, this was just a starting point. Let's use what we've learned to help us create a probabilistic model for predicting an output value given an input value. Specifically, we would like to determine a probability model

with parameters θ that can give us the probability (or density) of a random variable Y taking on the value y given the value x or random variable X,

$$P(Y = y \mid X = x, \theta) = p(y \mid x, \theta) \tag{104}$$

Once we have an estimate for this distribution, we can then use it to take a new input value x and predict the output value y.

Our new task is to:

1. Collect supervised dataset of N *i.i.d.* samples $\mathcal{D} = \left\{(x^{(i)}, y^{(i)}\right\}_{i=1}^{N}$
2. Use the relationship between input data x and output data y to define a probability distribution for $p(y \mid x, \theta)$
3. Estimate the value(s) for parameter(s) θ that maximizes $\prod_{i=1}^{N} p\left(y^{(i)} \mid x^{(i)}, \theta\right)$

4.1 Conditional likelihood and M(C)LE

We're going to use all of the principles from maximum likelihood estimation but first, we need to point out a subtle difference that can cause some confusion both here and when we get to more complicated probabilistic models later.

The **likelihood** is the probability or density of data given parameters can be $p(x^{(1)}, ..., x^{(N)} \mid \theta)$ or $p(y^{(1)}, ..., y^{(N)} \mid \theta)$ or even $p(x^{(1)}, y^{(1)}, ..., x^{(N)}, y^{(N)} \mid \theta)$ if we are considering data from both X and Y.

When we are working with $p\left(y^{(1)}, ..., y^{(N)} \mid x^{(1)}, ..., x^{(N)}, \theta\right)$, this is the **conditional likelihood**.

The confusing part is that in machine learning we use this conditional likelihood so often, we often resort to referring to it as just the "likelihood," dropping the "conditional." We'll even write things like this when we are working with the conditional likelihood:

$$p(\mathcal{D} \mid \theta) = p\left(y^{(1)}, ..., y^{(N)} \mid x^{(1)}, ..., x^{(N)}, \theta\right) \tag{105}$$

which is very confusing notation when we are trying to unpack $\mathcal{D} = \left\{(x^{(i)}, y^{(i)}\right\}_{i=1}^{N}$ and fit it into this.

We do the same thing with **maximum likelihood estimation** and **maximum conditional likelihood estimation**. Despite estimating the parameters that maximize the conditional likelihood, we never really use the term "MCLE" and just resort to calling it "MLE" or "maximum likelihood estimation," again dropping the "conditional."

So again, pardon the reuse of $p(\mathcal{D} \mid \theta)$, L, and ℓ as we define the following in the context of conditional likelihood and maximum conditional likelihood estimation:

$$\begin{aligned}
\mathcal{L}(\theta; \mathcal{D}) &= p(\mathcal{D} \mid \theta) && (106)\\
&= p\left(y^{(1)}, ..., y^{(N)} \mid x^{(1)}, ..., x^{(N)}, \theta\right) && (107)\\
&= \prod_{i=1}^{N} p\left(y^{(i)} \mid x^{(i)}, \theta\right) \quad \text{Assuming } \textit{i.i.d.} && (108)
\end{aligned}$$

Notational note: As always, we'll often drop the $\mathcal{D}$ from the likelihood function and the parameter(s)

θ from the probabilities/densities when it is (hopefully) clear that they are implied:

$$\mathcal{L}(\theta) = p(\mathcal{D} \mid \theta) \tag{109}$$
$$= p\left(y^{(1)}, ..., y^{(N)} \mid x^{(1)}, ..., x^{(N)}\right) \tag{110}$$
$$= \prod_{i=1}^{N} p\left(y^{(i)} \mid x^{(i)}\right) \quad \text{Assuming } i.i.d. \tag{111}$$

4.2 Logistic regression

4.2.1 Binary logistic regression

Let's start with logistic regression because we've actually done most of the work already. Specifically, we already decided that we wanted to model binary logistic regression for input $\mathbf{x} \in \mathbb{R}^M$ and binary output y as:

$$p(y \mid x) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}}} \quad \text{for } y = 1 \tag{112}$$
$$p(y \mid x) = 1 - \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}}} \quad \text{for } y = 0 \tag{113}$$

where $\boldsymbol{\theta} \in \mathbb{R}^M$ is the vector of logistic regression parameters.

Rather than defining a cross-entropy loss and optimizing an empirical risk objective, we can just use M(C)LE.

To simplify notation, let us define temporary variable $\phi^{(i)}$ as follows:

$$\phi^{(i)} = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}^{(i)}}} \tag{114}$$

The likelihood is:

$$\mathcal{L}(\theta; \mathcal{D}) = p(\mathcal{D} \mid \theta) \tag{115}$$
$$= p\left(y^{(1)}, ..., y^{(N)} \mid x^{(1)}, ..., x^{(N)}, \theta\right) \tag{116}$$
$$= \prod_{i=1}^{N} p\left(y^{(i)} \mid x^{(i)}, \theta\right) \tag{117}$$
$$= \prod_{i=1}^{N} {\phi^{(i)}}^{\mathbb{I}\left(y^{(i)}=1\right)} \left(1 - \phi^{(i)}\right)^{\mathbb{I}\left(y^{(i)}=0\right)} \tag{118}$$

(Note the similarities with a basic Bernoulli likelihood!)

The log-likelihood is:

$$\ell(\theta; \mathcal{D}) = \log \prod_{i=1}^{N} {\phi^{(i)}}^{\mathbb{I}\left(y^{(i)}=1\right)} \left(1 - \phi^{(i)}\right)^{\mathbb{I}\left(y^{(i)}=0\right)} \tag{119}$$
$$= \sum_{i=1}^{N} \log {\phi^{(i)}}^{\mathbb{I}\left(y^{(i)}=1\right)} + \log \left(1 - \phi^{(i)}\right)^{\mathbb{I}\left(y^{(i)}=0\right)} \tag{120}$$
$$= \sum_{i=1}^{N} \mathbb{I}\left(y^{(i)} = 1\right) \log \phi^{(i)} + \mathbb{I}\left(y^{(i)} = 0\right) \log \left(1 - \phi^{(i)}\right) \tag{121}$$

The MLE is:

$$\hat{\boldsymbol{\theta}}_{MLE} = \underset{\boldsymbol{\theta}}{\text{argmin}} \; -\ell(\theta; \mathcal{D}) \tag{122}$$

$$= \underset{\boldsymbol{\theta}}{\text{argmin}} \; -\sum_{i=1}^{N} \mathbb{I}\left(y^{(i)} = 1\right) \log \phi^{(i)} + \mathbb{I}\left(y^{(i)} = 0\right) \log\left(1 - \phi^{(i)}\right) \tag{123}$$

But wait, watch what happens after we make the following trivial substitutions: $y^{(i)} = \mathbb{I}\left(y^{(i)} = 1\right)$; $\left(1 - y^{(i)}\right) = \mathbb{I}\left(y^{(i)} = 0\right)$; and $\hat{y}^{(i)} = \phi^{(i)}$:

$$\hat{\boldsymbol{\theta}}_{MLE} = \underset{\boldsymbol{\theta}}{\text{argmin}} \; -\ell(\theta; \mathcal{D}) \tag{124}$$

$$= \underset{\boldsymbol{\theta}}{\text{argmin}} \; -\sum_{i=1}^{N} \mathbb{I}\left(y^{(i)} = 1\right) \log \phi^{(i)} + \mathbb{I}\left(y^{(i)} = 0\right) \log\left(1 - \phi^{(i)}\right) \tag{125}$$

$$= \underset{\boldsymbol{\theta}}{\text{argmin}} \; -\sum_{i=1}^{N} y^{(i)} \log \hat{y}^{(i)} + \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right) \tag{126}$$

This is the exact same optimization as minimizing the empirical risk for binary logistic regression with cross-entropy loss!!

4.2.2 Multi-class logistic regression

Again, we've already done most of the work by defining the output $\mathbf{y}$ to be a one-hot vector in $\mathbb{R}^K$ where $y_k = 1$ when the data point belongs to the k-th of K classes and defining the model as:

$$P(Y_k = 1 \mid \mathbf{x}; \Theta) = \frac{\exp\left(\boldsymbol{\theta}_k^\top \mathbf{x}\right)}{\sum_{l=1}^{K} \exp\left(\boldsymbol{\theta}_l^\top \mathbf{x}\right)} \tag{127}$$

Similar to binary logistic regression above, let's define $\phi_k^{(i)}$ to be:

$$\phi_k^{(i)} = \frac{\exp\left(\boldsymbol{\theta}_k^\top \mathbf{x}^{(i)}\right)}{\sum_{l=1}^{K} \exp\left(\boldsymbol{\theta}_l^\top \mathbf{x}^{(i)}\right)} \tag{128}$$

Just like for the categorical distribution, for the i-th factor corresponding to the i-th data point, we can write:

$$p\left(\mathbf{y}^{(i)} \mid \Theta\right) = \prod_{k=1}^{K} {\phi_k^{(i)}}^{\mathbb{I}\left(y_k^{(i)} = 1\right)} \tag{129}$$

The likelihood is:

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^{N} p\left(\mathbf{y}^{(i)} \mid \Theta\right) \tag{130}$$

$$= \prod_{i=1}^{N} \prod_{k=1}^{K} {\phi_k^{(i)}}^{\mathbb{I}\left(y_k^{(i)} = 1\right)} \tag{131}$$

The log-likelihood is:

$$\ell(\Theta; \mathcal{D}) = \log \prod_{i=1}^{N} \prod_{k=1}^{K} {\phi_k^{(i)}}^{\mathbb{I}\left(y_k^{(i)}=1\right)} \tag{132}$$

$$= \sum_{i=1}^{N} \log \prod_{k=1}^{K} {\phi_k^{(i)}}^{\mathbb{I}\left(y_k^{(i)}=1\right)} \tag{133}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \log {\phi_k^{(i)}}^{\mathbb{I}\left(y_k^{(i)}=1\right)} \tag{134}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}\left(y_k^{(i)} = 1\right) \log \phi_k^{(i)} \tag{135}$$

The MLE is:

$$\hat{\boldsymbol{\Theta}}_{MLE} = \underset{\boldsymbol{\theta}}{\text{argmin}} \; -\ell(\Theta; \mathcal{D}) \tag{136}$$

$$= \underset{\boldsymbol{\theta}}{\text{argmin}} \; -\sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}\left(y_k^{(i)} = 1\right) \log \phi_k^{(i)} \tag{137}$$

And, like the binary case, watch what happens after we make the following trivial substitutions: $y_k^{(i)} = \mathbb{I}\left(y_k^{(i)} = 1\right)$ and $\hat{y}_k^{(i)} = \phi_k^{(i)}$:

$$\hat{\boldsymbol{\Theta}}_{MLE} = \underset{\boldsymbol{\theta}}{\text{argmin}} \; -\ell(\Theta; \mathcal{D}) \tag{138}$$

$$= \underset{\boldsymbol{\theta}}{\text{argmin}} \; -\sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}\left(y_k^{(i)} = 1\right) \log \phi_k^{(i)} \tag{139}$$

$$= \underset{\boldsymbol{\theta}}{\text{argmin}} \; -\sum_{i=1}^{N} \sum_{k=1}^{K} y_k^{(i)} \log \hat{y}_k^{(i)} \tag{140}$$

This is the exact same optimization as minimizing the empirical risk for multi-class logistic regression with cross-entropy loss!!