Week 1: Tutorial

Distribution over discrete random variables

Let's take a toy example of discrete random variables.

This particular example is adapted from Section 2.3.1 of the book ``Probabilistic Machine Learning: An Introduction''.

Suppose you think you may have contracted COVID-19. You decide to take a diagnostic test, and you want to use its result to determine if you are infected or not.

Let $H = 1$ be the event that you are infected, and $H = 0$ the event that you are not infected. We have $Y = 1$ if the test is positive and $Y = 0$ that it is negative. We want to compute $p(H = h|Y = y)$.

The quantity obviously depends on how reliable the test is. There are two key parameters. The **sensitivity** (aka **true positive rate**) is defined as $p(Y = 1|H = 1)$. The **specificity** (aka **true negative rate**) is defined as $p(Y = 0|H = 0)$. Following Health Canada it seems reasonable for a PCR test to assume sensitivity 87.5% and specificity 97.5%.

Next we need to specify the prior. The quantity $p(H = 1)$ represents the **prelevance** of the disease in the area in which you live. We set this to $p(H = 1) = 0.1$.

Now we can easily compute the joint distribution of (Y, H). We have

$$\begin{aligned}
p(Y = 0, H = 0) &= 0.975 \cdot (1 - 0.1) \; = \; 0.8775 \\
p(Y = 0, H = 1) &= (1 - 0.875) \cdot 0.1 \; = \; 0.0125 \\
p(Y = 1, H = 0) &= (1 - 0.975) \cdot (1 - 0.1) \; = \; 0.0225 \\
p(Y = 1, H = 1) &= 0.875 \cdot 0.1 \; = \; 0.0875
\end{aligned}$$

Note that all four numbers are nonnegative and they sum to 1. The following table contains both the joint distribution of (Y, H) as well as both marginal distributions.

Y \ H	0	1	
0	0.8775	0.0125	0.89
1	0.0225	0.0875	0.11
	0.9	0.1	

Now suppose you test positive. We have

$$p(H=1|Y=1) \;=\; \frac{p(Y=1,H=1)}{p(Y=1)} \;=\; \frac{0.0875}{0.11} \;=\; 0.795$$

and so there is a 79.5% chance you are infected.

Now suppose you test negative. The probability you are infected is given by

$$p(H=1|Y=0) = \frac{p(Y=0,H=1)}{p(Y=0)} = \frac{0.0125}{0.89} \;=\; 0.014$$

and so there is just 1.4% chance you are infected.

Nowadays COVID-19 prevalence is much lower. Suppose we repeat these calculations using a base rate of 1%; now the posteriors reduce to 26% and 0.13% respectively.

The fact that you only have a 26% chance of being infected with COVID-19, even after a positive test is very counter-intuitive. The reason is that a single positive test is more likely to be false positive than due to the disease, since the disease is rare.

Note that we can think about the distribution of (Y,H) as a parametric model parametrized by three numbers: prevalence, sensitivity, and specificity.

Summary

- Given a joint distribution, we can compute both marginal and conditional distributions
- We'll consider distributions as equivalent to their parameters
- We can represent distributions by arrays of their parameters
- Operations like marginalizing and conditioning variables can be interpreted as operations on arrays of parameters.

MLE and Exponential Families

N-sample example: Multinomial distribution

A random variable $X \sim \mathrm{Multinomial}(\mathbf{q})$ where $\mathbf{q} \in \mathbb{R}^K$, which takes on $i \in [1,\ldots,K]$ discrete states each with probability q_i. That is $p(X=i) = q_i$, where $q_i \geq 0$ and $\sum_i q_i = 1$.
E.g. K-bit pixels, classes, unfair dice.

We observe N i.i.d. Multinomial($\mathbf{q}$), i.e. $\mathcal{D} = \{1,3,K,2,\ldots\}$ for N observations.

In the example from class, we had a single Bernoulli observation which we wrote as an exponential family. This time, we write the joint density of $\mathcal{D}$ as an exponential family.

Recall the natural form:

$$p(x|\eta) = h(x)\exp\{\eta^\top T(x) - A(\eta)\}$$

The model is $p(x^{(n)} = i|\mathbf{q}) = q_i$ with the constraint $\sum_i q_i = 1$. If $q_i > 0$ for all i then

$$\begin{aligned}
p(\mathcal{D};\mathbf{q}) &= \prod_{n=1}^{N} q_1^{1[x^{(n)}=1]} \cdots q_K^{1[x^{(n)}=K]} \\
&= \exp\log\prod_{n=1}^{N}\prod_{i=1}^{K} q_i^{1[x^{(n)}=i]} \\
&= \exp\sum_{n=1}^{N}\sum_{i=1}^{K} 1[x^{(n)} = i]\log q_i \\
&= \exp\sum_{i=1}^{K}\log q_i\sum_{n=1}^{N} 1[x^{(n)} = i] \\
&= \exp\sum_{i=1}^{K}\log q_i N_i
\end{aligned}$$

Therefore, the sufficient statistics for the multinomial distribution are the counts $N_i = \sum_{n=1}^{N} 1[x^{(n)} = i]$.

Aside (optional)
Note that in this form, we have that $\eta_i = \log q_i$, and if we use the formulas given in class for exponential families (the ones that link the derivative of $A(\eta)$ to the mean of the sufficient statistic), we get that the distribution has mean vector zero and variance matrix zero, which are not the mean and variance of the multinomial distribution. Note however that we are not allowed to use the formulas given in class here because the components of η that we have defined are not separately variable. Due to the constraint that $\sum_i q_i = 1$, we have the constraint that $\sum_i e^{\eta_i} = 1$, which complicates things.

There is an easy fix. Let's instead try to write the exponential family form to directly incorporate the constraint over $\mathbf{q}$. First, note that $q_K = 1 - \sum_{i=1}^{K-1} q_i$.

Substituting this in the formula above gives,

$$\begin{aligned}
p(\mathcal{D};\mathbf{q}) &= \exp\sum_i^K N_i\log q_i \\
&= \exp(\sum_i^{K-1} N_i\log q_i + N_K\log q_K) \\
&= \exp(\sum_i^{K-1} N_i\log q_i + (N-\sum_i^{K-1} N_i)\log(1-\sum_i^{K-1} q_i)) \\
&= \exp\left(\sum_i^{K-1} N_i\log\frac{q_i}{1-\sum_i^{K-1} q_i} + N\log(1-\sum_i^{K-1} q_i)\right)
\end{aligned}$$

We can read off the new parameterization now. For $i = 1,\dots,K-1$:

$$\begin{aligned}
T(x)_i &= N_i \\
\eta_i &= \log\frac{q_i}{1-\sum_i^{K-1} q_i} \\
h(x) &= 1 \\
A(\eta) &= -N\log(1-\sum_i^{K-1} q_i)
\end{aligned}$$

We would like to write $A(\eta)$ explicitly in terms of η instead of $\mathbf{q}$. First note that $\eta_i = \log\frac{q_i}{q_K} \implies e^{\eta_i} = q_i/q_K$ and $\sum_{i=1}^K q_i/q_K = 1/q_K = 1+\sum_{i=1}^{K-1} e^{\eta_i}$.
Then,

$$\begin{aligned}
A(\eta) &= N\log\frac{1}{(1-\sum_{i=1}^{K-1} q_i)} \\
&= N\log\frac{1}{q_K} \\
&= N\log\left(1+\sum_{i=1}^{K-1} e^{\eta_i}\right)
\end{aligned}$$

We can now use the formulas from class for exponential families to get the MLE. Alternatively we can use the initial likelihood we derived and solve for the MLE directly, as we show next.

For the MLE for this distribution, we consider the log-likelihood:

$$\ell(\mathbf{q};\mathcal{D}) = \sum_i \log q_i N_i$$

We can't simply take its derivative and set it equal to zero as it requires enforcing the constraint that $\sum_k q_k = 1$. We write the Lagrangian

$$L(\mathbf{q},\lambda) = \sum_i \log q_i N_i + \lambda(1-\sum_i q_i)$$

and take its derivative with respect to each q_i, set it equal to zero and solve for them.

$$\frac{dL}{dq_i} = \frac{N_i}{q_i} - \lambda = 0 \implies N_i/\lambda = q_i$$

Since $\sum_i N_i/\lambda = \sum_i q_i = 1$, we have $\lambda = \sum_i N_i = N$. Substituting into expression above, $q_i = \frac{N_i}{N}$.

Therefore, the maximum likelihood estimate for the class parameters in a multivariate distribution are the normalized counts for each class.

Example: Sufficient Statistics and MLE for Univariate Normal

We assume that the data is N i.i.d. samples $\{x^{(i)}\}_1^N \in \mathbb{R}$ from the Gaussian distribution, i.e.

$$\begin{aligned} x^{(i)} \sim p(x|\theta) &= \mathcal{N}(x|\mu, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} \end{aligned}$$

Gaussian distribution is a member of the exponential family, so we can put it into a natural form

$$\begin{aligned} p(x|\theta) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}\} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}\mu^2\} \exp\{\begin{bmatrix} \frac{\mu}{\sigma^2} & \frac{1}{\sigma^2} \end{bmatrix} \begin{bmatrix} x \\ -\frac{x^2}{2} \end{bmatrix}\} \end{aligned}$$

from here, it is clear that the natural parameters and the sufficient statistics are

- $\eta = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ \frac{1}{\sigma^2} \end{bmatrix}$
- $T(x) = \begin{bmatrix} x \\ -\frac{x^2}{2} \end{bmatrix}$

and so $\mu = \frac{\eta_1}{\eta_2}$, $\sigma^2 = \frac{1}{\eta_2}$, $-\frac{1}{2\sigma^2}\mu^2 = -\frac{1}{2}\frac{\eta_1^2}{\eta_2}$.

Note that given the sufficient statistics as defined, we could not determine anything more about η if we had access to any other information about the dataset. In this way the sufficient statistics are the minimum required statistics of the data. The normalization factor is read off once $\eta, T(x)$, and $h(x)$ are determined.

Re-writing the likelihood in terms of η

$$p(x|\eta) = \sqrt{\frac{\eta_2}{2\pi}} \cdot \exp\{-\frac{\eta_1^2}{2\eta_2}\} \cdot \exp\{\eta^T T(x)\}$$

noting that

- $h(x) = (2\pi)^{-\frac{1}{2}}$
- $A(\eta) = -\frac{1}{2}\log(\eta_2) + \frac{\eta_1^2}{2\eta_2}$

At this point we can use the general result from the exponential family, or take the derivatives in this form to find the MLE for those natural statistics η.

However, we often prefer to work with the parameterization by $\theta = [\mu, \sigma^2]$, so let's see this instead. We write the log-likelihood:

$$\begin{aligned}\ell(\theta;\mathcal{D}) = \log p(\mathcal{D}|\theta) &= \log \prod_n p(x^{(n)}|\theta) \\ &= \log \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x^{(n)} - \mu)^2\} \\ &= \sum_n -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x^{(n)} - \mu)^2 \\ &= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_n \frac{(x^{(n)} - \mu)^2}{\sigma^2}\end{aligned}$$

Solving for the derivatives of θ:

$$\frac{\partial \ell}{\partial \mu} = 0 + \frac{1}{\sigma^2}\sum_n x^{(n)} - \mu = 0 \Rightarrow \mu_{\text{MLE}} = \frac{1}{N}\sum_n x^{(n)}$$

So the MLE for the mean of a Gaussian is the mean of the data, intuitive!

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_n ((x^{(n)})^2 - \mu_{\text{MLE}}^2) = 0 \Rightarrow \sigma_{\text{MLE}}^2 = \frac{1}{N}\sum_n ((x^{(n)})^2 - \mu_{\text{MLE}}^2)$$

Also the MLE for the variance looks like the variance of the data.