One hot encoding and the Multinomial

EE-209 - Eléments de Statistiques pour les Data Sciences

One hot encoding

When working with a *nominal* or *ordinal variable* X taking values in $\{1, \ldots, K\}$, it is always more convenient in statistics and machine learning to work with its **indicator vector** representation often called **one hot encoding**:

$$Z = (Z_1, \ldots, Z_K) \quad \text{with} \quad Z_1 = 1_{\{X=1\}}, \quad \ldots, \quad Z_K = 1_{\{X=K\}}.$$

This called **one hot encoding** because Z_k takes values in $\{0, 1\}$ and $Z_1 + \ldots + Z_K = 1$.

Example: For X taking values in $\{1, \ldots, 4\}$ with $P_X(k) = \pi_k$.

$P_X(x)$	x	z
π_1	1	$(1, 0, 0, 0)$
π_2	2	$(0, 1, 0, 0)$
π_3	3	$(0, 0, 1, 0)$
π_4	4	$(0, 0, 0, 1)$

Note that we have:

$$\mathbb{E}[Z_k] = \mathbb{P}(Z_k = 1) = \mathbb{P}(X = k) = \pi_k$$

Continuing with the same example on the next slide...

Counts from sampling a discrete r.v. and the Multinomial

Sampling $n = 17$ independent values from X:

x_i	$P_X(x_i)$	z_{i1}	z_{i2}	z_{i3}	z_{i4}
1	π_1	1	0	0	0
4	π_4	0	0	0	1
1	π_1	1	0	0	0
2	π_2	0	1	0	0
4	π_4	0	0	0	1
4	π_4	0	0	0	1
1	π_1	1	0	0	0
3	π_3	0	0	1	0
4	π_4	0	0	0	1
2	π_2	0	1	0	0
4	π_4	0	0	0	1
3	π_3	0	0	1	0
3	π_3	0	0	1	0
4	π_4	0	0	0	1
3	π_3	0	0	1	0
3	π_3	0	0	1	0
1	π_1	1	0	0	0

	n_1	n_2	n_3	n_4
Counts	4	2	5	6

$$\begin{aligned}\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) &= \prod_{i=1}^{n} P_X(x_i) \\ &= \pi_1^{n_1}\, \pi_2^{n_2}\, \pi_3^{n_3}\, \pi_4^{n_4}\end{aligned}$$

The probability of the observed sequence depends only on the counts, but...

$$\mathbb{P}(N_1 = n_1, \ldots, N_4 = n_4) = \binom{n}{n_1, n_2, n_3, n_4} \pi_1^{n_1}\, \pi_2^{n_2}\, \pi_3^{n_3}\, \pi_4^{n_4}$$

with the *multinomial coefficient*

$$\binom{n}{n_1, n_2, n_3, n_4} := \frac{n!}{n_1!\, n_2!\, n_3!\, n_4!},$$

which counts the number of ways that a sequence of n numbers in $\{1, 2, 3, 4\}$ in which 1,2,3 and 4 appears respectively exactly n_1, n_2, n_3, n_4 times.

Multinomial random variable

Multinomial pmf

A vector of discrete r.v. $(N_1, \ldots, N_K)$ is said to follow jointly a multinomial distribution with parameters n and $(\pi_1, \ldots, \pi_K)$ and we write $(N_1, \ldots, N_K) \sim \mathcal{M}\big(n, (\pi_1, \ldots, \pi_k)\big)$ if $N_1 + \ldots + N_K = n$ and

$$\mathbb{P}(N_1 = n_1, \ldots, N_K = n_K) = \binom{n}{n_1, \ldots, n_K} \pi_1^{n_1} \ldots \pi_K^{n_K},$$

with $\binom{n}{n_1,\ldots,n_K} := \frac{n!}{n_1! \ldots n_K!}$, the *multinomial coefficient*.

Remark:

$$(N_1, N_2) \sim \mathcal{M}\big(n, (\pi_1, 1 - \pi_1)\big) \quad \Leftrightarrow \quad N_1 \sim \text{Bin}(n, \pi_1).$$

The "Multinoulli"?

What happens for $(N_1, \ldots, N_K) \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K))$?

- $(N_1, \ldots, N_K)$ is a vector of counts such that $N_1 + \ldots + N_k = 1$ so it has to be an

  indicator vector !

- Moreover $1! = 0! = 1$ so $\binom{n}{n_1, \ldots, n_K} = 1$ for all possible values of $n_1, \ldots, n_K$.

- So if $(Z_1, \ldots, Z_K) \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K))$, then it is an indicator vector and

$$\boxed{\mathbb{P}(Z_1 = z_1, \ldots, Z_K = z_k) = \pi_1^{z_1} \ldots \pi_K^{z_K}}$$

- The distribution of the indicator vector is therefore the counterpart of the Bernoulli and becomes the Bernoulli for $K = 2$.

The Bernoulli, the Binomial, the "Multinoulli" and the Multinomial

$Z \sim \text{Ber}(\pi)$	$(Z_1, \ldots, Z_K) \sim \mathcal{M}(1, \pi_1, \ldots, \pi_K)$
$P_Z(z) = \pi^z (1-\pi)^{1-z}$	$P_{\boldsymbol{Z}}(\boldsymbol{z}) = \pi_1^{z_1} \ldots \pi_K^{z_K}$
$N_1 \sim \text{Bin}(n, \pi)$	$(N_1, \ldots, N_K) \sim \mathcal{M}(n, \pi_1, \ldots, \pi_K)$
$P_{N_1}(n_1) = \binom{n}{n_1} \pi^{n_1} (1-\pi)^{n-n_1}$	$P_{\mathbf{N}}(\mathbf{n}) = \binom{n}{n_1 \;, \ldots ,\; n_K} \pi_1^{n_1} \ldots \pi_K^{n_K}$

with

$$\binom{n}{i} = \frac{n!}{(n-i)!i!} \qquad \text{and} \qquad \binom{n}{n_1 \;, \ldots ,\; n_K} = \frac{n!}{n_1! \ldots n_K!}$$