Sufficient Statistics

In many settings, the observed data comes from a very large alphabet or has a very high dimension

- We frequently want to preprocess the data first before doing inference

This is known as creating **representations** or **features** from the data

Leads to two questions:

- What representations are lossless with respect to the data?
- How can we quantify how compact the representation is?

For binary hypothesis testing, we have seen that the likelihood ratio is sufficient

For exponential families, we saw that the natural statistics and log base distribution fully determine the distribution

- For computing MLE, we only need the natural statistics

NonBayesian Setting

We first consider a nonBayesian setting where the parameters of interest are nonrandom

- They are fixed, but unknown

A statistic is a vector-valued function $t(y)$ of the data y

- We can then express the distribution of y as:

$$p_y(y;x) = p_{y|t}(y|t(y)) \cdot p_t(t(y);x)$$

- We can express the distribution of t as:

$$p_t(t'|x) = \sum_{y \in \mathcal{Y} | t(y) = t'} p_y(y;x)$$

Sufficient Statistic

A statistic t is sufficient if $p_{y|t}(\cdot \mid \cdot; x)$ does not depend on x

This means that once we know the value of t, any unremaining uncertainty about y does not have anything to do with x

- We have a similar notion with likelihood functions
- If we have our likelihood $L_{y=y}(x) = p_y(y;x)$ and the likelihood $L_{t=t(y)}(x) = p_t(t(y);x)$, then we expect that these should convey the same information to us
- As a result, we expect that they should be proportional to each other, which is conveyed by the following theorem:

Likelihood Characterization

A statistic t is sufficient iff the following is not a function of x for all y:

$$\frac{L_{y=y}(x)}{L_{t=t(y)}(x)} = \frac{p_y(y;x)}{p_t(t(y);x)}$$

Proof

- We note this ratio is equal to $p_{y|t}(y|t(y);x)$ which is not a function of x iff $t(y)$ is sufficient based on the definition

We next have a necessary and sufficient condition for sufficient statistics:

Neyman Factorization Theorem

A statistic t is sufficient iff there exist functions such that:
$p_y(y; x) = a(t(y), x) \cdot b(y)$$

Proof Sketch

- For the forward direction, assume t is a sufficient statistic
  - Then we can set $b(y) = p_{y|t}(y|t(y))$ and $a(t, x) = p_t(t; x)$
- For the reverse direction, we can write out $p_{y|t}$ as p_y/p_t
  - We can expand the bottom using the above definition and then factor $a(t, x)$ out of both numerator and denominator of fraction
  - We can then cancel out and get a fraction that does not depend on x

Minimal Sufficient Statistics

Given a sufficient statistic, we can augment it with any function of the data and still have it be sufficient

- In fact, y itself is a sufficient statistic
- We search for the statistic that is the most "compact"

One natural idea is to use dimension as a measure of compactness

- However, we can encode the entire data into a single real number via binary representation

Instead, we have the following dimension-free notion:

Minimal Sufficiency

A sufficient statistic s is minimal if for any other sufficient statistic t there exists a function g such that $s = g(t)$

With this definition, the minimal sufficient statistic is not unique

Bayesian Setting

When we choose to model the unknown parameter as a random variable x, the conditional distribution $p_{y|x}(y|x)$ replaces $p_y(y; x)$

Sufficient Statistic (Bayesian Form)

A statistic t is sufficient if:

$$p_{y|t,x}(y|t(y), x) = p_{y|t}(y|t(y))$$

Belief Characterization

A statistic t is sufficient iff:

$$p_{x|y}(\cdot|y) = p_{x|t}(\cdot|t(y))$$

Bayesian Version of Neyman Factorization

A statistic t is sufficient iff:

$$p_{y|x}(y|x) = p_{t|x}(t(y)|x)p_{y|t}(y|t(y))$$

Markov Chains

Markov chains provide a particularly useful way to view the relationship between the parameters, data, and statistic

A statistic can be described via a Markov constraint:

$$x \leftrightarrow y \leftrightarrow t$$

- This describes that t can't depend directly on x, but rather only y

We can then interpret a sufficeint statistic with a Markov relation:

$$x \leftrightarrow t \leftrightarrow y$$

Partitions

The most conceptually valuable interpretation of a sufficient statistic is as a **partition** of the space $\mathcal{Y}$ of observations

- For the purposes of inference, two observations y_1 and y_2 that lead to likelihood functions that are proportional to each other provide the same amount of information
- We can use this to partition the space of observations into equivalence classes
- We can do something similar with sufficient statistics

Sufficient Statistic, Partition Characterization

A statistic t is sufficient iff for all y_1 and y_2 such that $t(y_1) = t(y_2)$, we have $L_{y=y_1}(x) \propto L_{y=y_2}(x)$ where the constant of proportionality only depends on y_1 and y_2

Minimal Sufficient Statistic, Partition Characterization

A statistic t is minimal iff for all y_1 and y_2 such that $L_{y=y_1}(x) \propto L_{y=y_2}(x)$ we have $t(y_1) = t(y_2)$

We then have that a minimal sufficient statistic partitions our space $\mathcal{Y}$ into classes that have the same t value

- A sufficient statistic merely partitions our space into a space of proportional likelihood functions
- A minimal sufficient statistic can be thought of as a maximally-coarse partition
- Any sufficient statistic is merely a finer partition of a minimal sufficient statistic

Existence and Computation of Minimal Sufficient Statistics

We restrict attention to the case of finite $\mathcal{Y}$

- It follows immediately that a minimal sufficient always exist via the following procedure:

1. Select $x_0 \in \mathcal{X}$ arbitrarily
2. Order the elements of $\mathcal{Y}$ arbitrarily, labeling them $y_1, \ldots$
3. Initialize $m = 1$
4. Construct the normalized likelihood function $\tilde{L}_i(x) = \frac{L_{y=y_i}(x)}{L_{y=y_i}(x_0)}$
5. Make the assignment $t(y_1) = 1$ and let $f_1 = \tilde{L}_1$
6. For each $i \geq 2$, in order do:

- If $\tilde{L}_i = f_{m'}$ for any $m' \in \{1, \ldots, m\}$, assign $t(y_i) = m'$
- Otherwise, increment m, assign $t(y_i) = m$, and let $f_m = \tilde{L}_i$

When $\mathcal{X}$ is finite, the computational complexity is $O(|X||Y|^2)$

- This can't be carried out exactly when $\mathcal{X}$ is infinite, only approximately
- This cost is also prohibitively large for large alphabets
- This is generally to establish the existence, and there is generally no known computationally feasible construction

In general, there isn't even a practical methodology that is guaranteed to resolve whether some statistic is minimal

- While necessary and sufficient conditions are not known, we do know some sufficient ones

Complete Sufficient Statistic

A sufficient statistic is complete if any function $\phi(\cdot) : t(\mathcal{Y}) \rightarrow \mathbb{R}$ satisfying:

$$\mathbb{E}\left[\phi(t(y)\right] = 0$$

must satisfy:

$$P(\phi(t(y)) = 0) = 1$$

Theorem

A sufficient statistic t is minimal if it is complete

Completeness is not necessary for minimality, it is merely a sufficient condition that can be easier to test sometimes