Maximizing Data Likelihood Implies Minimizing Cross Entropy

Arthur Tilley

Theorem 1. *Let our model space be*

$$\mathscr{F} = \{Q_f\}_f = \{\{Q_{f,\bar{x}}\}_{\bar{x}}\}_f = \{\{Q(\cdot; f(\bar{x}))\}_{\bar{x}}\}_f$$

some parameterized family of families of probability distributions where

$$Q_{f,\bar{x}}(\bar{y}) = Q(\bar{y}; f(\bar{x}))$$

is the probability of the label $\bar{y}$, given parameters $f(\bar{x})$.

Furthermore, suppose we have some set of data

$$D = \{(\bar{x}^{(i)}, \bar{y}^{(i)})\}_{1 \leq i \leq M}$$

consisting of M observations of features $\bar{x}^{(i)}$ *paired with* labels $\bar{y}^{(i)}$.

Then maximizing data likelihood, that is picking the f which makes D most likely, is equivalent to each of the following:

1. *Minimizing the average cross-entropy*

$$\frac{1}{M} \sum_{i=1}^{M} H(P_i, Q_{f,\bar{x}^{(i)}})$$

   between the model distributions $Q_{f,\bar{x}^{(i)}} = Q(\cdot; f(\bar{x}^{(i)}))$ and the one-hot sample distributions defined as

$$P_i(\bar{y}) = \mathbf{1}_{\bar{y}^{(i)}}(\bar{y}) = \begin{cases} 1 & \text{if } \bar{y} = \bar{y}_i \\ 0 & \text{if } \bar{y} \neq \bar{y}_i \end{cases}$$

2. *Minimizing the average cross-entropy*

$$E_{\bar{x} \sim P} H(P_{\bar{x}}, Q_{f,\bar{x}}) = \sum_{\bar{x} \in D_{\bar{x}}} P(\bar{x}) H(P_{\bar{x}}, Q_{f,\bar{x}})$$

   where

(a) $D_{\bar{x}}$ is just the set of all $\bar{x}$ that occur anywhere in the data:

$$D_{\bar{x}} = \{\bar{x} : \bar{x} = \bar{x}^{(i)} \text{ for some } 1 \leq i \leq M\}$$

(b) $P(\bar{x})$ is the sample distribution of $\bar{x}$ in D:

$$P(\bar{x}) = \frac{\#_D(\bar{x})}{M} = \frac{|\{i : 1 \leq i \leq M \text{ and } \bar{x}^{(i)} = \bar{x}\}|}{M}$$

(c) $P_{\bar{x}}(\bar{y})$ is the sample distribution of $\bar{y}$ given *$\bar{x}$ in D:*

$$P_{\bar{x}}(\bar{y}) := \frac{\#_D(\bar{x}, \bar{y})}{\#_D(\bar{x})} = \frac{|\{i : 1 \leq i \leq M \text{ and } (\bar{x}^{(i)}, \bar{y}^{(i)} = (\bar{x}, \bar{y})\}|}{|\{i : 1 \leq i \leq M \text{ and } \bar{x}^{(i)} = \bar{x}\}|}.$$

Concisely, this is to say that

1. $$argmax_f Pr(D|f) = argmin_f \frac{1}{M} \sum_{i=1}^{M} H(P_i, Q_{f,\bar{x}^{(i)}}), \text{ and}$$

2. $$argmax_f Pr(D|f) = argmin_f E_{\bar{x} \sim P} H(P_{\bar{x}}, Q_{f,\bar{x}}).$$

Proof.

Toward showing both equivalences we first notice that:

$$argmax_f Pr(D|f) = argmax_f \prod_i Pr((\bar{y}^{(i)}, \bar{x}^{(i)})|f) \quad \text{(since data is i.i.d)}$$

$$= argmax_f \prod_i Pr(\bar{y}^{(i)}|f, \bar{x}^{(i)}) \quad \text{(features } \bar{x}^i \text{ are independent of } f)$$

$$= argmax_f \prod_i Q(\bar{y}^{(i)}; f(\bar{x}^{(i)})) = argmax_f \prod_i Q_{f,\bar{x}^{(i)}}(\bar{y}^{(i)}) \quad \text{(by definition of } Q_{f,\bar{x}})$$

$$= argmax_f \log \prod_i Q_{f,\bar{x}^{(i)}}(\bar{y}^{(i)}) \quad \text{(since log is monotonic increasing)}$$

$$= argmax_f \sum_{i=1}^{M} \log Q_{f,\bar{x}^{(i)}}(\bar{y}^{(i)}) \quad \text{(since } \log xy = \log x + \log y)$$

$$= argmax_f \frac{1}{M} \sum_{i=1}^{M} \log Q_{f,\bar{x}^{(i)}}(\bar{y}^{(i)}) \quad \text{(positive constant multiplication is monotonic increasing)}$$

$$= argmin_f -\frac{1}{M}\sum_{i=1}^{M} \log Q_{f,\bar{x}^{(i)}}(\bar{y}^{(i)}) \quad \text{(insert minus sign and change } max \text{ to } min) \qquad (*)$$

We have labelled this last expression with a star (*) because we will use it to show both equivalences in the claim.

Toward showing the first equivalence (1.), let $D_{\bar{y}}$ be the set of all $\bar{y}$ values that occur anywhere in the data. Then we can see that the function inside the $argmin$ in (*) is just

$$-\frac{1}{M}\sum_{i=1}^{M} \log Q_{f,\bar{x}^{(i)}}(\bar{y}^{(i)}) = -\frac{1}{M}\sum_{i=1}^{M}\left(\sum_{\bar{y}\in D} \mathbf{1}_{\bar{y}_i}(\bar{y}) \log Q_{f,\bar{x}^{(i)}}(\bar{y})\right)$$

(This inner sum is just a big disjunction (*or*-statement) over all possible values that $\bar{y}^{(i)}$ could take; only one term is ever non-zero at a time.)

$$= -\frac{1}{M}\sum_{i=1}^{M}\left(\sum_{\bar{y}\in D} P_i(\bar{y}) \log Q_{f,\bar{x}^{(i)}}(\bar{y})\right) \quad \text{(by definition of } P_i \text{)}$$

$$= -\frac{1}{M}\sum_{i=1}^{M}\left(E_{P_i} \log Q_{f,\bar{x}^{(i)}}\right) \quad \text{(by definition of expectation)}$$

$$= \frac{1}{M}\sum_{i=1}^{M} H(P_i, Q_{f,\bar{x}^{(i)}}) \quad \text{(by definition of cross-entropy)}$$

Thus have shown equivalence (1)

$$argmax_f Pr(D|f) = argmin_f \frac{1}{M}\sum_{i=1}^{M} H(P_i, Q_{f,\bar{x}^{(i)}})$$

To show the second equivalence (2.), we note that, when we get to (*) above, we may instead partition the sum (currently over all points in the data) into sums over groups determined by all possible values $\bar{x}$ and $\bar{y}$ that appear for any point in the data. When rewriting the sums as over the groups, we just need to remember to multiply each term by the number of times that that particular value appeared in the data. Thus, again, letting $D_{\bar{x}}$ and $D_{\bar{y}}$ represent the sets of all $\bar{x}$ and $\bar{y}$ values appearing anywhere in the data, we see that the function inside the $argmin$ in (*) is just

$$-\frac{1}{M}\sum_{i=1}^{M} \log Q_{f,\bar{x}^{(i)}}(\bar{y}^{(i)}) = -\frac{1}{M}\sum_{\bar{x}\in D_{\bar{x}}}\sum_{\bar{y}\in D_{\bar{y}}} \#_D(\bar{x},\bar{y}) \log Q_{f,\bar{x}}(\bar{y}) \quad \text{(grouping the sum)}$$

$$= -\frac{1}{M}\sum_{\bar{x}\in D_{\bar{x}}} \#_D(\bar{x}) \sum_{\bar{y}\in D_{\bar{y}}} \frac{\#_D(\bar{x},\bar{y})}{\#_D(\bar{x})} \log Q_{f,\bar{x}}(\bar{y}) \quad \text{(multiply and divide by } \#_D(\bar{x}))$$

$$= -\frac{1}{M} \sum_{\bar{x} \in D_{\bar{x}}} \#_D(\bar{x}) \sum_{\bar{y} \in D_{\bar{y}}} P_{\bar{x}}(\bar{y}) \log Q_{f,\bar{x}}(\bar{y}) \quad \text{(Definition of } P_{\bar{x}} \text{)}$$

$$= -\frac{1}{M} \sum_{\bar{x} \in D_{\bar{x}}} \#_D(\bar{x}) E_{P_{\bar{x}}} \log Q_{f,\bar{x}} \quad \text{(Definition of expectation)}$$

$$= \frac{1}{M} \sum_{\bar{x} \in D_{\bar{x}}} \#_D(\bar{x}) H(P_{\bar{x}}, Q_{f,\bar{x}}) \quad \text{(Definition of cross-entropy)}$$

$$= \sum_{\bar{x} \in D_{\bar{x}}} P(\bar{x}) H(P_{\bar{x}}, Q_{f,\bar{x}}) \quad \text{(Definition } P\text{)}$$

$$= E_{\bar{x} \sim P} H(P_{\bar{x}}, Q_{f,\bar{x}}) \quad \text{(Definition expectation.)}$$

And this proves equivalence (2).

$$argmax_f Pr(D|f) = argmin_f E_{\bar{x} \sim P} H(P_{\bar{x}}, Q_{f,\bar{x}}).$$

□