KL divergence and maximum likelihood

Last modified: 2026-11-25 01:48:21 (PST)

The Kullback–Leibler divergence of one model from another is the number of bits per value that a code built for the model wastes, by the cross-entropy decomposition. This page applies that to data, and shows that the model a maximum likelihood estimate picks is the model with the least waste.

Maximum likelihood

Theorem 0.1 (Maximum likelihood minimizes the divergence from the data)**.** *Let $x_1, \dots, x_n$ be data, and let $\hat{\mathbb{p}}$ be their empirical distribution. Let q_θ be probability mass functions with $q_\theta(v) > 0$ wherever $\hat{\mathbb{p}}(v) > 0$, for parameter values θ. Then θ maximizes the likelihood $\prod_{i=1}^n q_\theta(x_i)$ exactly when it minimizes the Kullback–Leibler divergence $D_{\mathrm{KL}}(\hat{\mathbb{p}} \parallel q_\theta)$.*

Proof

Proof. Let n_v be the number of data equal to v. Then $n_v = n\ \hat{\mathbb{p}}(v)$.
Group the terms of the log-likelihood by value:

$$\sum_{i=1}^{n} \log_2 q_\theta(x_i) = \sum_{v} n\ \hat{\mathbb{p}}(v) \log_2 q_\theta(v). \tag{1}$$

Divide by n. The right side is minus the cross-entropy $\mathbb{H}(\hat{\mathbb{p}}, q_\theta)$:

$$\frac{1}{n} \sum_{i=1}^{n} \log_2 q_\theta(x_i) = -\mathbb{H}(\hat{\mathbb{p}}, q_\theta). \tag{2}$$

By the cross-entropy decomposition, the cross-entropy is the entropy of $\hat{\mathbb{p}}$ plus the divergence:

$$-\mathbb{H}(\hat{\mathbb{p}}, q_\theta) = -\mathbb{H}(\hat{\mathbb{p}}) - D_{\mathrm{KL}}(\hat{\mathbb{p}} \parallel q_\theta). \tag{3}$$

The entropy $\mathbb{H}(\hat{\mathbb{p}})$ does not depend on θ. So θ maximizes the average log-likelihood exactly when it minimizes $D_{\mathrm{KL}}(\hat{\mathbb{p}} \,\|\, q_\theta)$. The logarithm increases, and dividing by n keeps the order, so θ maximizes the average log-likelihood exactly when it maximizes the likelihood. □

Example 0.1 (The divergence from the data for coin flips)**.**
A coin is flipped 100 times, and it lands heads 70 times. The chunk below follows Theorem 0.1 for the Bernoulli models with the probability of heads $0.1, 0.2, \ldots, 0.9$. It prints the divergence of each model from the empirical distribution, the average log-likelihood, and minus the entropy of the empirical distribution minus the divergence.

```
n_flips <- 100
n_heads <- 70
empirical <- c(tails = n_flips - n_heads, heads = n_heads) / n_flips
success_prob <- seq(0.1, 0.9, by = 0.1)
empirical_entropy <- -sum(empirical * log2(empirical))
divergence <- numeric(length(success_prob))
avg_loglik <- numeric(length(success_prob))
for (j in seq_along(success_prob)) {
  model <- c(1 - success_prob[j], success_prob[j])
  divergence[j] <- sum(empirical * log2(empirical / model))
  avg_loglik[j] <- sum(empirical * log2(model))
}
decomposition <- -empirical_entropy - divergence
data.frame(
  success_prob,
  divergence = round(divergence, 4),
  avg_loglik = round(avg_loglik, 4),
  decomposition = round(decomposition, 4)
)
#>   success_prob divergence avg_loglik decomposition
#> 1          0.1     1.4897    -2.3710       -2.3710
#> 2          0.2     0.8406    -1.7219       -1.7219
#> 3          0.3     0.4890    -1.3702       -1.3702
#> 4          0.4     0.2651    -1.1464       -1.1464
#> 5          0.5     0.1187    -1.0000       -1.0000
#> 6          0.6     0.0312    -0.9125       -0.9125
#> 7          0.7     0.0000    -0.8813       -0.8813
#> 8          0.8     0.0406    -0.9219       -0.9219
#> 9          0.9     0.2217    -1.1030       -1.1030
```

The last two columns agree, as the proof says. The divergence is smallest, at 0 bits, and

the average log-likelihood is largest for the model with probability of heads 0.7, which is the fraction of heads in the data, so that model is the empirical distribution.

Remark 0.1 (One labeled example). For a single labeled example, $n = 1$, and the empirical distribution puts all its mass on the observed label. Then Theorem 0.1 says that the model with the largest probability for that label is the model with the smallest divergence from it. That is the case that the cross-entropy loss remark describes.

References