

Discussion 4

This week's resources:

- <https://opusfivedev.com/> for simulating when you want to fix just one thing but your agent is DTM link
- "We Must Pace the Frontier" by Anthropic CEO Dario Amodei link
- "Who Gets to Define the Rules for AI" by Cohere CEO Aidan Gomez link

1 Maximizing Likelihood & Minimizing Cost

Maximum likelihood estimation is often done implicitly when training models. That is, maximizing likelihood and minimizing cost have a very strong correspondence.

Suppose we have a multiclass problem where an individual data point belongs to one of the classes A , T , C , or G . Let $B = \{A, T, C, G\}$. Our dataset looks something like the following.

$$\begin{aligned} r_1 &= A \\ r_2 &= C \\ r_3 &= A \\ &\vdots \\ r_n &= G \end{aligned}$$

We can model this data as a sample from a multinomial distribution. In a multinomial distribution, the order of the data does not matter, and we can equivalently represent our dataset as $(y, c_y)_{y \in B}$, where c_y is the number of items of class y . For instance, $(A, 5)$ means that A appears 5 times in our sample.

The parameters of the multinomial distribution are the number n of trials (data points), the number k of classes (in our example, $k = 4$), and the probabilities p_i of each class i , where $\sum_{i=1}^k p_i = 1$. The probability mass function (PMF) of the multinomial distribution—this is, the probability in n trials of obtaining each class i c_i times—is

$$P(c_1, \dots, c_k) = n! \prod_{i=1}^k \frac{p_i^{c_i}}{c_i!}.$$

In our example, let's use p_A , p_T , p_C , and p_G instead of p_1 , p_2 , p_3 , and p_4 .

- (a) Derive an expression for the likelihood $\mathcal{L}(p_A, p_T, p_C, p_G; c_A, c_T, c_C, c_G)$ of our observed data.

Solution: Since we are working in the discrete case, we simply have $\mathcal{L}(\theta; c) = P(c; \theta)$.

$$\begin{aligned}\mathcal{L}(p_A, p_T, p_C, p_G; c_A, c_T, c_C, c_G) &= P(c_A, c_T, c_C, c_G; p_A, p_T, p_C, p_G) \\ &= n! \prod_{y \in B} \frac{p_y^{c_y}}{c_y!}\end{aligned}$$

- (b) Show that estimating the class probabilities to maximize multinomial likelihood is equivalent to minimizing the cross-entropy cost. Concretely, prove the following.

$$\operatorname{argmax}_{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1=1}} \mathcal{L}(p; c_A, c_T, c_C, c_G) = \operatorname{argmin}_{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1=1}} - \sum_{i=1}^n \sum_{y \in B} \delta_{r_i y} \log p_y$$

where $\mathcal{L}$ is the likelihood you derived in part (a), p is the vector of probabilities per class $\begin{bmatrix} p_A & p_T & p_C & p_G \end{bmatrix}^\top$, and $\delta_{r_i y}$ is the Kronecker delta that outputs 1 if $r_i = y$ and 0 otherwise.

Solution:

$$\begin{aligned}\operatorname{argmax}_{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1=1}} \mathcal{L}(p; c_A, c_T, c_C, c_G) &= \operatorname{argmax}_{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1=1}} n! \prod_{y \in B} \frac{p_y^{c_y}}{c_y!} \\ &= \operatorname{argmax}_{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1=1}} \log \prod_{y \in B} \frac{p_y^{c_y}}{c_y!} \\ &= \operatorname{argmax}_{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1=1}} \sum_{y \in B} \log \frac{p_y^{c_y}}{c_y!} \\ &= \operatorname{argmax}_{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1=1}} \sum_{y \in B} c_y \log p_y - \sum_{y \in B} \log(c_y!)\end{aligned}$$

Note that the $\sum_{y \in B} \log(c_y!)$ term is a constant with respect to p , so it does not affect the optimization problem.

$$\begin{aligned}\operatorname{argmax}_{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1=1}} \sum_{y \in B} c_y \log p_y - \sum_{y \in B} \log(c_y!) &= \operatorname{argmin}_{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1=1}} - \sum_{y \in B} c_y \log p_y \\ &= \operatorname{argmin}_{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1=1}} - \sum_{i=1}^n \sum_{y \in B} \delta_{r_i y} \log p_y\end{aligned}$$

■

- (c) Suppose we have data in the form (x_i, y_i) , where y_i is the observed label for a corresponding point x_i . Our goal is to fit some function $g_\theta(x)$ to the available data (θ contains our function's parameters). We can model the unobserved error between our observations and the ground truth as i.i.d. Gaussian with mean 0 and variance σ^2 . Hence, we can model the vector of observed labels itself as $y \sim \mathcal{N}(g_\theta(x), \sigma^2 I)$. A core aspect of machine learning is using available data to optimize the function parameters θ (training). Now that we have a probabilistic model of y , we can try MLE! Show that the resulting optimal θ^* from performing MLE on Gaussian likelihood is equivalent to the θ^* we obtain by minimizing squared error. Concretely, prove the following.

$$\operatorname{argmax}_{\theta} \mathcal{L}(\theta, \sigma^2; y) = \operatorname{argmin}_{\theta} \|g_\theta(x) - y\|_2^2$$

where $\mathcal{L}$ is the Gaussian likelihood for you to define. Note that, in the continuous case, $\mathcal{L}(\theta; x) = f(x; \theta)$, where f is the PDF of our distribution.

Solution:

$$\begin{aligned} \operatorname{argmax}_{\theta} \mathcal{L}(\theta, \sigma; y) &= \operatorname{argmax}_{\theta} \prod_i f(y_i; \theta, \sigma^2) \\ &= \operatorname{argmax}_{\theta} \log \prod_i f(y_i; \theta, \sigma^2) \\ &= \operatorname{argmax}_{\theta} \sum_i \left(-\frac{(y_i - g_\theta(x_i))^2}{2\sigma^2} - \log \sqrt{2\pi} - \log \sigma \right) \\ &= \operatorname{argmin}_{\theta} \sum_i (y_i - g_\theta(x_i))^2 \\ &= \operatorname{argmin}_{\theta} \|g_\theta(x) - y\|_2^2 \end{aligned}$$

■

2 Closed-Form Solution for ℓ_2 -Regularized Least Squares (Ridge)

Consider a data matrix $X \in \mathbb{R}^{N \times D+1}$ with rows $x_n^\top$, a target vector $t \in \mathbb{R}^N$, and a parameter vector $w \in \mathbb{R}^{D+1}$. For a regularization hyperparameter $\lambda \geq 0$, define the ridge objective:

$$E(w) = \|Xw - t\|_2^2 + \lambda \|w\|_2^2.$$

- (a) Write $E(w)$ in expanded quadratic form $w^\top Aw - 2b^\top w + c$ by identifying A , b , and c in terms of X , t , and λ .

Solution: Expand:

$$\|Xw - t\|_2^2 = (Xw - t)^\top (Xw - t) = w^\top X^\top Xw - 2t^\top Xw + t^\top t.$$

Thus

$$E(w) = w^\top X^\top Xw - 2t^\top Xw + t^\top t + \lambda w^\top w = w^\top (X^\top X + \lambda I)w - 2(X^\top t)^\top w + t^\top t.$$

So $A = X^\top X + \lambda I$, $b = X^\top t$, and $c = t^\top t$.

- (b) Compute the gradient $\nabla_w E(w)$ using the identities

$$\nabla_w \|Xw - t\|_2^2 = 2X^\top (Xw - t), \quad \nabla_w \|w\|_2^2 = 2w.$$

Solution:

$$\nabla_w E(w) = 2X^\top (Xw - t) + 2\lambda w = 2(X^\top Xw - X^\top t + \lambda w) = 2((X^\top X + \lambda I)w - X^\top t).$$

- (c) Set the gradient to zero and derive the *normal equations* for ridge regression. Solve for w .

Solution: Stationarity $\nabla_w E(w) = 0$ gives

$$(X^\top X + \lambda I)w = X^\top t.$$

Assuming $X^\top X + \lambda I$ is invertible (see next part), the unique minimizer is

$$\boxed{\hat{w} = (X^\top X + \lambda I)^{-1} X^\top t}.$$

- (d) Justify why $X^\top X + \lambda I$ is invertible for $\lambda > 0$ (and discuss when it might fail for $\lambda = 0$).

Solution: For any $v \neq 0$,

$$v^\top (X^\top X + \lambda I)v = \|Xv\|_2^2 + \lambda \|v\|_2^2.$$

If $\lambda > 0$, this sum is strictly positive, so $X^\top X + \lambda I$ is **positive definite** and therefore invertible. If $\lambda = 0$, invertibility requires $X^\top X$ to be positive definite, i.e., X must have full column rank ($\text{rank}(X) = d$). Otherwise the solution is not unique (the Moore–Penrose pseudoinverse yields the minimum-norm solution).