1 Maximizing Likelihood & Minimizing Cost

Maximum Likelihood Estimation (MLE) is a method for estimating the parameters of a statistical model given observations.

Data Suppose we obtain n discrete *observations* belonging to $B := \{1, 2, 3, 4\}$. Our dataset looks something like the following.

$$\begin{aligned} r_1 &= 1 \\ r_2 &= 1 \\ r_3 &= 3 \\ &\vdots \\ r_n &= 1 \end{aligned}$$

Assumptions Suppose we aim to estimate the occurence probabilities of each class in B based on the observed data. We additionally assume that observations are independent and identically distributed (i.i.d.). In particular, this assumption implies that the order of the data does not matter.

Model Based on these assumptions, a natural model for our data is the 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.

The probability mass function (PMF) of the multinomial distribution—this is, the probability in n trials of obtaining each class i x_i times—is

$$P(x_1, \ldots, x_k) = n! \prod_{i=1}^{k} \frac{p_i^{x_i}}{x_i!}.$$

(a) Derive an expression for the likelihood for this problem. What are the observations? What are the parameters? What parameters are we trying to estimate with MLE?

(b) Typically, the log-likelihood $\ell(\theta) = \log L(\theta)$ is used instead of $L(\theta)$. Write down the expression for $\ell(\theta)$. Why might this be a good idea?

(c) Another idea might be to minimize the cross-entropy based on raw observations, corresponding to the following program

$$\underset{\substack{p\in\mathbb{R}^4_+ \\ \|p\|_1=1}}{\operatorname{argmin}} -\sum_{i=1}^{n}\sum_{y\in B}\delta_{r_i y}\log p_y$$

where p is the vector of probabilities per class $\begin{bmatrix} p_1 & p_2 & p_3 & p_4 \end{bmatrix}^\top$, and $\delta_{r_i y}$ is the Kronecker delta that outputs 1 if $r_i = y$ and 0 otherwise.

Show that this program is equivalent to the MLE program.

2 Independence and Multivariate Gaussians

As described in lecture, a covariance matrix $\Sigma \in \mathbb{R}^{N\times N}$ for a random variable $X \in \mathbb{R}^N$ with the following values, where $\text{cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$ is the covariance between the i-th and j-th elements of the random vector X:

$$\Sigma = \begin{bmatrix} \text{cov}(X_1, X_1) & \dots & \text{cov}(X_1, X_n) \\ \dots & & \dots \\ \text{cov}(X_n, X_1) & \dots & \text{cov}(X_n, X_n) \end{bmatrix}. \tag{1}$$

Recall that the density of an N dimensional Multivariate Gaussian Distribution $\mathcal{N}(\mu, \Sigma)$ is defined as follows when Σ is positive definite:

$$f(x) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}. \tag{2}$$

Here, $|\Sigma|$ denotes the determinant of the matrix Σ.

(a) Consider the random variables X and Y in $\mathbb{R}$ with the following conditions.

(i) X and Y can take values $\{-1, 0, 1\}$.

(ii) When X is 0, Y takes values 1 and -1 with equal probability ($\frac{1}{2}$). When Y is 0, X takes values 1 and -1 with equal probability ($\frac{1}{2}$).

(iii) Either X is 0 with probability ($\frac{1}{2}$), or Y is 0 with probability ($\frac{1}{2}$).

Are X and Y uncorrelated? Are X and Y independent? Prove your assertions. *Hint:* Write down the joint probability of (X, Y) for each possible pair of values they can take.

(b) For $X = [X_1, \cdots, X_n]^\top \sim \mathcal{N}(\mu, \Sigma)$, **verify that if X_i, X_j are independent (for all $i \neq j$), then Σ must be diagonal, i.e., X_i, X_j are uncorrelated.**

(c) Let $N = 2$, $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and $\Sigma = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}$. Suppose $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$. **Show that X_1, X_2 are independent if $\beta = 0$.** Recall that two continuous random variables W, Y with joint density $f_{W,Y}$ and marginal densities f_W, f_Y are independent if $f_{W,Y}(w, y) = f_W(w) f_Y(y)$.

(d) Consider a data point x drawn from an N-dimensional zero mean Multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, as shown above. Assume that Σ^{-1} exists. **Prove that there exists a matrix $A \in \mathbb{R}^{N \times N}$ such that $x^\top \Sigma^{-1} x = \|Ax\|_2^2$ for all vectors x. What is the matrix A?**

3 Least Squares (using vector calculus)

In ordinary least-squares linear regression, we typically have $n > d$ so that there is no $\mathbf{w}$ such that $\mathbf{X}\mathbf{w} = \mathbf{y}$ (these are typically overdetermined systems — too many equations given the number of unknowns). Hence, we need to find an approximate solution to this problem. The residual vector will be $\mathbf{r} = \mathbf{X}\mathbf{w} - \mathbf{y}$ and we want to make it as small as possible. The most common case is to measure the residual error with the standard Euclidean ℓ^2-norm. So the problem becomes:

$$\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2,$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}, \mathbf{w} \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}^n$.

Assume that $\mathbf{X}$ is full rank.

(a) How do we know that $\mathbf{X}^\top \mathbf{X}$ is invertible?

(b) Derive using vector calculus an expression for an optimal estimate for $\mathbf{w}$ for this problem.

(c) What should we do if $\mathbf{X}$ is not full rank?

