Section 02: Maximum Likelihood Estimation, Linear Algebra

Maximum Likelihood Estimation

From MLE to optimization

In this section, we formulate maximum likelihood estimation for different noise densities as different minimization problems. Specifically, we'll see how each noise distribution corresponds to a specific objective function.

We consider the linear measurement model, $y_i = x_i^\top w + \epsilon_i$ for $i = 1, 2, \ldots, m$. The noise ϵ_i for different measurements are all independent and identically distributed. Per the principle of maximum likelihood estimation, we seek to maximize

$$\log p_w(y) = \log \prod_{i=1}^{m} p(y_i - x_i^\top w).$$

(a) Show that when the noise measurements follow a Gaussian distribution ($\epsilon_i \sim \mathcal{N}(0, \sigma^2)$), the maximum likelihood estimate of w is the solution to $\min_w \|Xw - y\|_2^2$.

(b) Show that when the noise measurements follow a Laplacian distribution ($p(z) = (1/2a)\exp(-|z|/a)$), the maximum likelihood estimate of w is the solution to $\arg\min_w \|Xw - y\|_1$.

(c) Show that when the noise measurements follow a uniform distribution ($p(z) = (1/2a)$ on $[-a, a]$), the maximum likelihood estimate is any w satisfying $\|Xw - y\|_\infty \leq a$.

From optimization to MLE

In the previous problem (part a), we showed that for different noise distributions, the maximum likelihood estimation problem can be formulated as minimization problems with different objectives. Here, we flip our perspective and show the opposite: suppose you have any penalty minimization problem $\min_w \sum_{i=1}^m \phi(y_i - x_i^\top w)$, then we claim that we can express it equivalently as a maximum likelihood estimation problem with observations y_i and noise density p. What is the expression for the density function p? Your answer should be in terms of the function ϕ.

Linear Algebra Review

Let $X \in \mathbb{R}^{m \times n}$. X may not have full rank. We explore properties about the four fundamental subspaces of X.

Subspaces of X

What is the rowspace, columnspace, nullspace, and rank of $X = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6. \end{pmatrix}$.

Connections between subspaces of X

Check the following facts.

(a) The rowspace of X is the columnspace of $X^\top$, and vice versa.

(b) The nullspace of X and the rowspace of X are orthogonal complements. This can be written in shorthand as $\text{Null}(X) = \text{Range}(X^\top)^\perp$. This is further equivalent to saying $\text{Range}(X^\top) = \text{Null}(X)^\perp$.

(c) The nullspace of $X^\top$ is orthogonal to the columnspace of X. This can be written in shorthand as $\text{Null}(X^\top) = \text{Range}(X)^\perp$.

Linear algebra facts for linear regression

We saw in Linear Regression (Week 2) that the closed form expression for linear regression without an offset involves the term $(X^\top X)^{-1}$.

(a) Is it true that the matrix $X^\top X$ is always symmetric and positive semidefinite?

(b) State and prove the connection between the nullspace of X and the nullspace of $X^\top X$. That is, your statement should look like one of the following: $\text{Null}(X) \subseteq \text{Null}(X^\top X)$, or $\text{Null}(X) \supseteq \text{Null}(X^\top X)$ or $\text{Null}(X) = \text{Null}(X^\top X)$.

(c) Is it true that $X^\top X$ is always invertible?

(d) Based on the above fact about the connection between the nullspaces of $X \in \mathbb{R}^{m \text{ x } n}$ and $X^\top X$ and the expression for linear regression without an offset (that we referred to two problems above), justify the use of "tall skinny (m»n)" data matrix X as opposed to a "short wide (m«n)" matrix X.

(e) The columnspace and rowspace of $X^\top X$ are the same, and are equal to the rowspace of X. (Hint: Use the relationship between nullspace and rowspace.)