Section 02: Solutions

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 $\min_w \|Xw - y\|_2^2$.

Solution:

When $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, the density is given by the expression $p(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-z^2/2\sigma^2}$. This implies that the log likelihood function is

$$\ell(w) = \log p_w(y) = m \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} \|Xw - y\|_2^2$$

From the definition of MLE we get:

$$\hat{w}_{mle} = \arg\max m \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2} \|Xw - y\|_2^2$$

We cancel out terms that do not contribute to maximizing w to get:

$$= \arg\max -\|Xw - y\|_2^2$$

We then change to argmin by negating:

$$= \arg\min \|Xw - y\|_2^2$$

Therefore, the maximum likelihood estimate of w is $\hat{w}_{mle} = \arg\min \|Xw - y\|_2^2$, as claimed.

(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 $\arg\min_w \|Xw - y\|_1$.

Solution:

When ϵ_i distributed on a Lapalcian distribution which has the form: For $a > 0, p(z) = (1/2a)\exp(-|z|/a)$, we have that the log likelihood function is:

$$\ell(w) = \log p_w(y) = m \log \frac{1}{2a} \sum_{i=1}^{m} \frac{-1}{a} |y_i - x_i^T w|$$

From the definition of MLE we get:

$$\hat{w}_{mle} = \arg\max m \log \frac{1}{2a} + \sum_{i=1}^{m} \frac{-1}{a} |y_i - x_i^T w|$$

We cancel terms that do not contribute to maximizing w to get:

$$= \arg\max \sum_{i=1}^{m} (-1) |y_i - x_i^T w|$$

We then change to argmin by negating to get:

$$= \arg\min \sum_{i=1}^{m} |y_i - x_i^T w|$$

Which is equivalent to

$$= \arg\min \|y - Xw\|_1$$

By definition of L1 norm, $\|y - Xw\|_1 = \|Xw - y\|_1$, so we have that the $\hat{w}_{mle} = \arg\min \|Xw - y\|_1$, as claimed.

(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 $\|Xw - y\|_\infty \leq a$.

Solution:

When ϵ_i distributed on a uniform distribution which has the form $p(z) = (1/2a)$ on $[-a, a]$, we have that the log likelihood function (using an indicator variable) is :

$$mlog\frac{1}{2a} + \sum_{i-1}^{m} \log 1\{-a \leq y_i - x_i^T w \leq a\}$$

From the definition of MLE we get:

$$\hat{w}_{mle} = \arg\max mlog\frac{1}{2a} + \sum_{i-1}^{m} \log 1\{-a \leq y_i - x_i^T w \leq a\}$$

We cancel terms that do not contribute to maximizing w to get:

$$= \arg\max \sum_{i-1}^{m} \log 1\{-a \leq y_i - x_i^T w \leq a\}$$

We simplify this to be:

$$= \arg\max \begin{cases} 0 & \text{if } \forall i - a \leq y_i - x_i^T w \leq a \\ -\infty & \text{otherwise} \end{cases} \tag{1}$$

We simplify more to get:

$$= \arg\max \begin{cases} 0 & \text{if } \|y - Xw\|_\infty \leq a \\ -\infty & \text{otherwise} \end{cases} \tag{2}$$

Thus the maximum likelihood estimate is any w satisfying $\|y - Xw\|_\infty \leq a$ which is the same as $\|Xw - y\|_\infty \leq a$, as claimed.

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.

Solution:

A penalty minimization problem of the type provided can be interpreted as a maximum likelihood estimation problem with noise density $p(z) = \frac{e^{-\phi(z)}}{\int e^{-\phi(u)} du}$.

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 $X = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6. \end{pmatrix}$.

Solution:

- Rowspace is the **span** (i.e., *the set of all linear combinations*) of the rows of X. Therefore, in this example, it is the subspace of vectors of the form $(1 \cdot x + 4 \cdot y, 2 \cdot x + 5 \cdot y, 3 \cdot x + 6 \cdot y)$ for all x and y.
- Columnspace (a.k.a. Range(X)) is the span of the columns of X. In this example, it is the subspace of vectors of the form $(1 \cdot x + 2 \cdot y + 3 \cdot z, 4 \cdot x + 5 \cdot y + 6 \cdot z)$ for all x, y, and z.
- Nullspace (a.k.a. Null(X)) is the set of vectors v such that $Xv = 0$. In this example, the nullspace is the subspace spanned by $(1, -2, 1)$.
- The matrix X can be reduced to the form $\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \end{pmatrix}$. This matrix has submatrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, which has rank 2. Observe that the third column, $\begin{pmatrix} -1 \\ 2 \end{pmatrix}$, is in the columnspace of this first submatrix.

Connections between subspaces of X

Check the following facts.

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

Solution:

The matrix $X^\top$ is $\begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$. The rows of X are the columns 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 Null$(X) =$ Range$(X^\top)^\perp$. This is further equivalent to saying Range$(X^\top) =$ Null$(X)^\perp$.

Solution:

A vector $v \in$ Null(X) if and only if $Xv = 0$, which is true if and only if for every row X_i of X, $\langle X_i, v \rangle = 0$. This is precisely the condition that v is perpendicular to each row of X, which is the stated claim.

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

Solution:

This is seen by applying the previous result to $X^\top$.

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?

Solution:

Yes. Symmetry can be checked by computing the transpose. For any vector u, we have $u^\top X^\top X u = \|Xu\|_2^2 \geq 0$.

(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: Null$(X) \subseteq$ Null$(X^\top X)$, or Null$(X) \supseteq$ Null$(X^\top X)$ or Null$(X) =$ Null$(X^\top X)$.

Solution:

We have, Null$(X) =$ Null$(X^\top X)$. Let $v \in$ Null(X). Then, one can check that $X^\top X v = 0$, leading to $v \in$ Null$(X^\top X)$, which proves Null$(X) \subseteq$ Null$(X^\top X)$. For the other direction, let $0 \neq v \in$ Null$(X^\top X)$. Then, $0 = v^\top X^\top X v = \|Xv\|_2^2$, which implies $v \in$ Null(X). Therefore, Null$(X^\top X) \subseteq$ Null(X), which finishes the proof.

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

Solution:

No, this isn't always the case. Since Null$(X) =$ Null$(X^\top X)$ (see the answer to the previous question), the matrix $X^\top X$ is not invertible if X has a non-empty nullspace.

(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.

Solution:

If X is "short and wide", it has a non-empty nullspace. Therefore, $X^\top X$ is not invertible.

(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.)

Solution:

$X^\top X$ is symmetric, and previous parts, we have rowspace$(X^\top X) =$ columnspace$((X^\top X)^\top) =$ columnspace$(X^\top X)$. By previous parts again, we have: rowspace$(X^\top X) =$ Null$(X^\top X)^\perp =$ Null$(X)^\perp =$ rowspace(X).