Maximum Likelihood

Advanced Topics in High-Performance Computing

Faisal Qureshi
We now turn our attention to probabilistic view of linear regression.

Using vector calculus

$\Theta: (1.811322, 0.524238)$
Univariate Gaussian distribution

\[ N(x|\mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2\sigma^2} (x - \mu)^2 \right) \]

\( \mu \) is the center of mass or mean

\( \sigma^2 \) is the variance

\( \mu \) and \( \sigma^2 \) are sufficient statistics

Sampling from a Gaussian

\[ x \sim N(\mu, \sigma^2) \]
Multivariate Gaussian distribution

Gaussian distribution in $d$-dimensions

$$
\mathcal{N}(\mathbf{x} | \mu, \Sigma) = \frac{1}{\sqrt{|\Sigma|} (2\pi)^{\frac{d}{2}}} \exp \left( -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right)
$$

$x, \mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$

Example: Gaussian in 2D
Covariance

Covariance between two random variables $X$ and $Y$ measures the degree to which these variables are linearly related.

$$\text{cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

$\mathbb{E}[X]$ is the expected value of the random variable $X$.

$$\mathbb{E}[X] = \int xp(x)dx = \mu$$
Covariance matrix $\Sigma$

If $x \in \mathbb{R}^d$ random vector, its covariance matrix $\Sigma$ is defined as follows:

$$
\Sigma = \text{cov}[x] = \\
\begin{bmatrix}
\text{var}[X_1] & \text{cov}[X_1, X_2] & \cdots & \text{cov}[X_1, X_d] \\
\text{cov}[X_2, X_1] & \text{var}[X_2] & \cdots & \text{cov}[X_2, X_d] \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}[X_d, X_1] & \text{cov}[X_d, X_2] & \cdots & \text{var}[X_d]
\end{bmatrix}
$$
Likelihood example

Consider the points: \( y_1 = 1 \), \( y_2 = 0.5 \) and \( y_3 = 1.5 \). The points are drawn from a Gaussian with unknown mean \( \theta \) and \( \sigma^2 = 1 \).

\[
y_i \sim \mathcal{N}(\theta, 1)
\]

Points are independent so

\[
P(y_1, y_2, y_3|\theta) = P(y_1|\theta)P(y_2|\theta)P(y_3|\theta)
\]

Our goal is to find the Gaussian (i.e., find its mean, since variance is already given) that maximizes the likelihood of this data.

From Nando de Freitas
Linear regression

Consider data points \((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})\). Our goal is to learn a function \(f(x)\) that returns (predict) the value \(y\) given an \(x\).
The likelihood for linear regression

Let’s assume that targets $y(i)$ are corrupted by Gaussian noise with 0 mean and $\sigma^2$ variance

$$y^{(i)} = \theta^T x^{(i)} + \mathcal{N}(0, \sigma^2)$$

$$= \mathcal{N}(\theta^T x^{(i)}, \sigma^2)$$

In higher dimensions, we write:

$$y^{(i)} = \mathcal{N}(\theta^T x^{(i)}, \sigma^2)$$

Why assume Gaussian noise?

- Mathematically convenient
- A reasonably accurate assumption in practice
- Central Limit Theorem
The likelihood for linear regression

Under the assumption that each \( y^{(i)} \) is i.i.d., we can write the likelihood of \( y \) given data \( X \) as follows:

\[
p(y | X; \theta, \sigma) = \prod_{i=1}^{N} p(y^{(i)} | x^{(i)}; \theta, \sigma)
= \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-1/2} e^{-\frac{1}{2\sigma^2} \left(y^{(i)} - \theta^T x^{(i)}\right)^2}
= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - \theta^T x^{(i)}\right)^2}
= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2} (y - X\theta)^T (y - X\theta)}
\]

Aside: the “;” above indicate that we are following the frequentist approach, and we do not treat \( \theta \) as a random variable. Rather we view \( \theta \) as having some true value that we are trying to estimate.
Probability of data given parameters

Loss for linear regression

\[ C(\theta) = (y - X\theta)^T (y - X\theta) \]

Probability of data given parameters is related to the loss for linear regression that we obtained before.
Maximum likelihood estimation (1)

The maximum likelihood estimate (MLE) of $\theta$ is obtained by maximizing $p(y|X, \theta, \sigma)$

$$
\theta_{ML} = \arg \max_{\theta} \prod_{i=1}^{N} p(y^{(i)}|x^{(i)}; \theta, \sigma)
$$

Log likelihood

$$
p(y|X; \theta, \sigma) = (2\pi \sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2} (y-X\theta)^T(y-X\theta)}
$$
Maximum likelihood estimation (2)
Making predictions using MLE

For a previously unseen data $x^*$, the target $y^*$ can be obtained as follows:

$$y^* \sim \mathcal{N}(\theta^T_{ML} x^*, \sigma^2)$$
Entropy

Entropy $H$ is a measure of uncertainty associated with a random variable.

$$H(X) = - \sum_x p(x|\theta) \log p(x|\theta)$$

Example

Entropy of a Gaussian in $D$ dimensions

$$H(\mathcal{N}(\mu, \Sigma)) = \frac{1}{2} \ln \left[ (2\pi e)^D |\Sigma| \right]$$
Kullback-Leibler divergence

*Kullback-Leibler* (KL) divergence is a measure of how much two probability distributions diverge from each other.

For discrete probability distributions

\[ D_{KL} (P \parallel Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \]

For continuous probability distributions

\[ D_{KL} (P \parallel Q) = \int p(x) \log \frac{p(x)}{q(x)} dx \]
Kullback-Leibler divergence

MLE: For $\mathcal{M}$ data from some distribution $p(x | \theta_0)$ the MLE minimizes the KL divergence (KULLBACK-LEIBLER divergence)

$$\hat{\theta} = \arg \max_{\theta} \prod_{i=1}^{m} p(x^{(i)} | \theta)$$

$$= \arg \max_{\theta} \sum_{i=1}^{m} \log p(x^{(i)} | \theta)$$

$$= \arg \max_{\theta} \frac{1}{m} \sum_{i=1}^{m} \log p(x^{(i)} | \theta)$$

$$- \frac{1}{m} \sum_{i=1}^{m} \log p(x^{(i)} | \theta_0)$$

$$= \arg \max_{\theta} \left[ \frac{1}{m} \sum_{i=1}^{m} \log \frac{p(x^{(i)} | \theta)}{p(x^{(i)} | \theta_0)} \right]$$

$$= \arg \min_{\theta} \left[ \frac{1}{m} \sum_{i=1}^{m} \log \frac{p(x^{(i)} | \theta_0)}{p(x^{(i)} | \theta)} \right]$$

$$\Rightarrow \lim_{m \to \infty} \arg \min_{\theta} \int \log \frac{p(x | \theta_0)}{p(x | \theta)} p(x | \theta_0) \, dx$$

Figure 1:
MLE and KL divergence

It turns out that for i.i.d. (independant, identically distributed) data from a some (unknown true) distribution \( p(x|\theta_{\text{True}}) \) MLE minimizes the Kullback-Leibler (KL) divergence.
Ridge regression and Bayes rule

Previously we saw the loss function for ridge regression

\[ C(\theta) = (y - X\theta)^T (y - X\theta) + \delta^2 \theta^T \theta \]

We can cast the above in probabilistic terms

\[ p(y|x, \theta) = \frac{1}{Z_1} e^{-((y - X\theta)^T (y - X\theta))} \]

Then

\[ p(\theta) = \frac{1}{Z_2} e^{-\delta^2 \theta^T \theta} \]

becomes prior.
Summary

- We developed a probabilistic view of linear regression.