Linear Regression MLE

As a starting point for linear regression, we assume that the labels y_i are a linear function of the corresponding features $\mathbf{x}_i$ with additive Gaussian noise ϵ:

$$y_i = \mathbf{w}^\top \mathbf{x}_i + \epsilon$$

The Gaussian noise ϵ is mathematically drawn from a Gaussian (Normal) random variable $\epsilon \sim \mathcal{N}(0, \sigma)$. This means that the noise is sampled from a Gaussian probability distribution with mean 0 and standard deviation σ.

> ☆ **Key Points**
>
> The maximum likelihood estimator maximizes the likelihood of the observed data given the model assumptions.
>
> The MLE expression is mathematically equivalent to the minimization of the mean squared error.

If ϵ is sampled with mean 0 and standard deviation σ, then y_i (given $\mathbf{x}_i, \mathbf{w}$) must be sampled from a Gaussian random variable with mean $\mathbf{w}^\top \mathbf{x}_i$ and standard deviation σ. This follows directly from the definition of $y_i = \mathbf{w}^\top \mathbf{x}_i + \epsilon$ (mean of ϵ gets shifted by $+\mathbf{w}^\top \mathbf{x}_i$).
We can then get the exact conditional probability distribution of $y_i | \mathbf{x}_i ; \mathbf{w}$:

$$P(y_i | \mathbf{x}_i | \mathbf{w}) = \frac{1}{2\pi\sigma^2} e^{\frac{(y_i - (w)^2 \mathbf{x}_i)^2}{2\sigma^2}}$$

Given such a probability distribution, we need to find the parameter vector $\mathbf{w}$ that best explains our data. To this end, we can deploy MLE, which will seek the $\mathbf{w}$ that maximizes the likelihood of our observed labels (given their feature vectors):

$$\begin{aligned}
\mathbf{w} &= \arg\max_{\mathbf{w}} \prod_{i=1}^{n} \mathrm{P}(y_i | \mathbf{x}_i ; \mathbf{w}) \\
&= \arg\max_{\mathbf{w}} \sum_{i=1}^{n} \log[\mathrm{P}(y_i | \mathbf{x}_i ; \mathbf{w})] && \text{(log converts a product of probabil} \\
&= \arg\max_{\mathbf{w}} \sum_{i=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \log\left(e^{\frac{(\mathbf{x}_i^\top \mathbf{w} - y_i)^2}{2\sigma^2}} \right) \right] && \text{(P} \\
&= \arg\max_{\mathbf{w}} -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 && \text{(First t} \\
&= \arg\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 && \left(\frac{1}{n} \text{ makes the loss interp}
\end{aligned}$$

This derivation shows that the MLE estimate of $\mathbf{w}$ is the minimizer of the following **loss function**:

$$l(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2$$

This particular loss function is also known as the squared loss or **ordinary least squares (OLS)**. OLS can be optimized with gradient descent, Newton's method, or in closed form (i.e. solving for the exact analytical solution). It has a nice explanation: the linear prediction $\mathbf{w}^\top \mathbf{x}_i$ should minimize the squared error from the true label y_i.