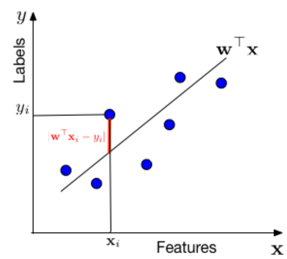
Linear Regression

Assumptions

Data Assumption: $y_i \in \mathbb{R}$
Model Assumption: $y_i = \mathbf{w}^\top \mathbf{x}_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$

$\Rightarrow y_i|\mathbf{x}_i \sim N(\mathbf{w}^\top \mathbf{x}_i, \sigma^2) \Rightarrow P(y_i|\mathbf{x}_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\mathbf{x}_i^\top \mathbf{w} - y_i)^2}{2\sigma^2}}$



In words, we assume that the data is drawn from a "line" $\mathbf{w}^\top \mathbf{x}$ through the origin (one can always add a bias / offset through an additional dimension, similar to the Perceptron). For each data point with features $\mathbf{x}_i$, the label y is drawn from a Gaussian with mean $\mathbf{w}^\top \mathbf{x}_i$ and variance σ^2. Our task is to estimate the slope $\mathbf{w}$ from the data.

Estimating with MLE

$$
\begin{aligned}
\mathbf{w} &= \operatorname*{argmax}_{\mathbf{w}} P(y_1, \mathbf{x}_1, \ldots, y_n, \mathbf{x}_n|\mathbf{w}) \\
&= \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{n} P(y_i, \mathbf{x}_i|\mathbf{w}) && \text{Because data points are independently samplec} \\
&= \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w}) P(\mathbf{x}_i|\mathbf{w}) && \text{Chain rule of probability} \\
&= \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w}) P(\mathbf{x}_i) && \mathbf{x}_i \text{ is independent of } \mathbf{w}\text{, we only model } P(y_i|\mathbf{x} \\
&= \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w}) && P(\mathbf{x}_i) \text{ is a constant - can be droppe} \\
&= \operatorname*{argmax}_{\mathbf{w}} \sum_{i=1}^{n} \log[P(y_i|\mathbf{x}_i, \mathbf{w})] && \log \text{ is a monotonic functio} \\
&= \operatorname*{argmax}_{\mathbf{w}} \sum_{i=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(e^{-\frac{(\mathbf{x}_i^\top \mathbf{w} - y_i)^2}{2\sigma^2}}\right)\right] && \text{Plugging in probability distributio} \\
&= \operatorname*{argmax}_{\mathbf{w}} -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 && \text{First term is a constant, and } \log(e^z) = \\
&= \operatorname*{argmin}_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 && \text{Always minimize; } \frac{1}{n} \text{ makes the loss interpretable (average squared error}
\end{aligned}
$$

We are minimizing a *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.

Closed Form: $\mathbf{w} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}^\top$ where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathbf{y} = [y_1, \ldots, y_n]$.

Estimating with MAP

Additional Model Assumption: $P(\mathbf{w}) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{\mathbf{w}^\top \mathbf{w}}{2\tau^2}}$

$$\begin{aligned}
\mathbf{w} &= \operatorname*{argmax}_{\mathbf{w}} P(\mathbf{w}|y_1, \mathbf{x}_1, \ldots, y_n, \mathbf{x}_n) \\
&= \operatorname*{argmax}_{\mathbf{w}} \frac{P(y_1, \mathbf{x}_1, \ldots, y_n, \mathbf{x}_n|\mathbf{w}) P(\mathbf{w})}{P(y_1, \mathbf{x}_1, \ldots, y_n, \mathbf{x}_n)} \\
&= \operatorname*{argmax}_{\mathbf{w}} P(y_1, \mathbf{x}_1, \ldots, y_n, \mathbf{x}_n|\mathbf{w}) P(\mathbf{w}) \\
&= \operatorname*{argmax}_{\mathbf{w}} \left[\prod_{i=1}^{n} P(y_i, \mathbf{x}_i|\mathbf{w})\right] P(\mathbf{w}) \\
&= \operatorname*{argmax}_{\mathbf{w}} \left[\prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w}) P(\mathbf{x}_i|\mathbf{w})\right] P(\mathbf{w}) \\
&= \operatorname*{argmax}_{\mathbf{w}} \left[\prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w}) P(\mathbf{x}_i)\right] P(\mathbf{w}) \\
&= \operatorname*{argmax}_{\mathbf{w}} \left[\prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w})\right] P(\mathbf{w}) \\
&= \operatorname*{argmax}_{\mathbf{w}} \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i, \mathbf{w}) + \log P(\mathbf{w}) \\
&= \operatorname*{argmin}_{\mathbf{w}} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 + \frac{1}{2\tau^2} \mathbf{w}^\top \mathbf{w} \\
&= \operatorname*{argmin}_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 + \lambda ||\mathbf{w}||_2^2 \qquad \lambda = \frac{\sigma^2}{n\tau^2}
\end{aligned}$$

This objective is known as Ridge Regression. It has a closed form solution of: $\mathbf{w} = (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I})^{-1} \mathbf{X}\mathbf{y}^\top$, where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $\mathbf{y} = [y_1, \ldots, y_n]$.

Summary

Ordinary Least Squares:
- $\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2$.
- Squared loss.
- No regularization.
- Closed form: $\mathbf{w} = (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{X}\mathbf{y}^\top$.

Ridge Regression:
- $\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 + \lambda ||\mathbf{w}||_2^2$.
- Squared loss.
- $l2\text{-regularization.}$
- Closed form: $\mathbf{w} = (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I})^{-1} \mathbf{X}\mathbf{y}^\top$.