Linear Regression

In this lecture we will learn about Linear Regression.

Assumptions

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

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

Estimating with MLE

$$
\begin{aligned}
\mathbf{w} &= \operatorname*{argmax}_{\mathbf{w}} \sum_{i=1}^{n} log(P(y_i|\mathbf{x}_i, \mathbf{w})) \\
&= \operatorname*{argmax}_{\mathbf{w}} \sum_{i=1}^{n} \left[log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + log\left(e^{-\frac{(\mathbf{x}_i^\top \mathbf{w} - y_i)^2}{2\sigma^2}}\right) \right] \\
&= \operatorname*{argmax}_{\mathbf{w}} -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 \\
&= \operatorname*{argmin}_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2
\end{aligned}
$$

The loss is thus $l(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2$ AKA square loss or Ordinary Least Squares (OLS). OLS can be optimized with gradient descent, Newton's method, or in closed form.

Closed Form: $\mathbf{w} = (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{X}\mathbf{y}^\top$

Estimating with MAP

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

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

This formulation is known as Ridge Regression. It has a closed form solution of:
$\mathbf{w} = (\mathbf{X}\mathbf{X}^\top + \lambda^2 \mathbf{I})^{-1} \mathbf{X}\mathbf{y}^\top$

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.
- $l2$-regularization.
- Closed form: $\mathbf{w} = (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I})^{-1} \mathbf{X}\mathbf{y}^\top$.