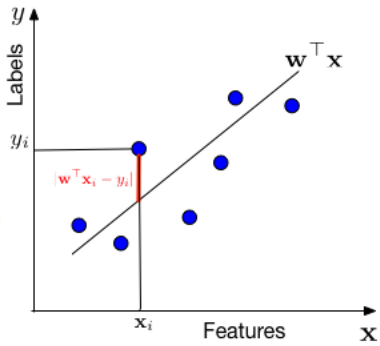
Linear Regression

Cornell CS 4/5780 — Spring 2022

Assumptions

Data Assumption: $y_i \in \mathbb{R}$
Model Assumption: $y_i = \mathbf{w}^T\mathbf{x}_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$
$\Rightarrow y_i|\mathbf{x}_i \sim N(\mathbf{w}^T\mathbf{x}_i, \sigma^2) \Rightarrow P(y_i|\mathbf{x}_i, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\mathbf{x}_i^T\mathbf{w}-y_i)^2}{2\sigma^2}}$



In words, we assume that the data is drawn from a "line" $\mathbf{w}^T\mathbf{x}$ 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 $\mathbf{w}^T\mathbf{x}_i$ and variance σ^2. Our task is to estimate the slope $\mathbf{w}$ from the data.

How can we motivate this model using the central limit theorem?

Estimating with MLE

$$
\begin{aligned}
\hat{\mathbf{w}}_{\text{MLE}} &= \operatorname*{argmax}_{\mathbf{w}} \; P(y_1, \mathbf{x}_1, \ldots, y_n, \mathbf{x}_n|\mathbf{w}) \\
&= \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{n} P(y_i, \mathbf{x}_i|\mathbf{w}) && \text{Because data points are independently} \\
&= \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w})P(\mathbf{x}_i|\mathbf{w}) && \text{Chain rule of probability} \\
&= \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w})P(\mathbf{x}_i) && \mathbf{x}_i \text{ is independent of } \mathbf{w}\text{, we only model } P(y_i|\mathbf{x}) \\
&= \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{n} P(y_i|\mathbf{x}_i, \mathbf{w}) && P(\mathbf{x}_i) \text{ is a constant - can be dropped} \\
&= \operatorname*{argmax}_{\mathbf{w}} \sum_{i=1}^{n} \log\left[P(y_i|\mathbf{x}_i, \mathbf{w})\right] && \log \text{ is a monotonic function} \\
&= \operatorname*{argmax}_{\mathbf{w}} \sum_{i=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(e^{-\frac{(\mathbf{x}_i^T\mathbf{w}-y_i)^2}{2\sigma^2}}\right)\right] && \text{Plugging in probability distribution} \\
&= \operatorname*{argmax}_{\mathbf{w}} \; -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{w} - y_i)^2 && \text{First term is a constant, and } \log(e^z) = z \\
&= \operatorname*{argmin}_{\mathbf{w}} \; \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{w} - y_i)^2 && \text{Scale and switch to minimize}
\end{aligned}
$$

We are minimizing a *loss function*, $l(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{w} - y_i)^2$. This particular loss function is also known as the squared loss or Ordinary Least Squares (OLS). In this form, it has a natural interpretation as the average squared error of the prediction over the training set. OLS can be optimized with gradient descent, Newton's method, or in closed form.

Closed Form Solution: if $\mathbf{X}\mathbf{X}^T$ is invertible, then

$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y}^T \text{ where } \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n} \text{ and } \mathbf{y} = [y_1, \ldots, y_n] \in \mathbb{R}^{1\times n}.$$

Otherwise, there is not a unique solution, and any $\mathbf{w}$ that is a solution of the linear equation

$$\mathbf{X}\mathbf{X}^T\hat{\mathbf{w}} = \mathbf{X}\mathbf{y}^T$$

minimizes the objective.

Estimating with MAP

To use MAP, we will need to make an additional modeling assumption of a prior for the weight $\mathbf{w}$.

$$P(\mathbf{w}) = \frac{1}{\sqrt{2\pi\tau^2}}e^{-\frac{\mathbf{w}^T\mathbf{w}}{2\tau^2}}.$$

With this, our MAP estimator becomes

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

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

Summary

Ordinary Least Squares:

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

Ridge Regression:

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