

Assumptions

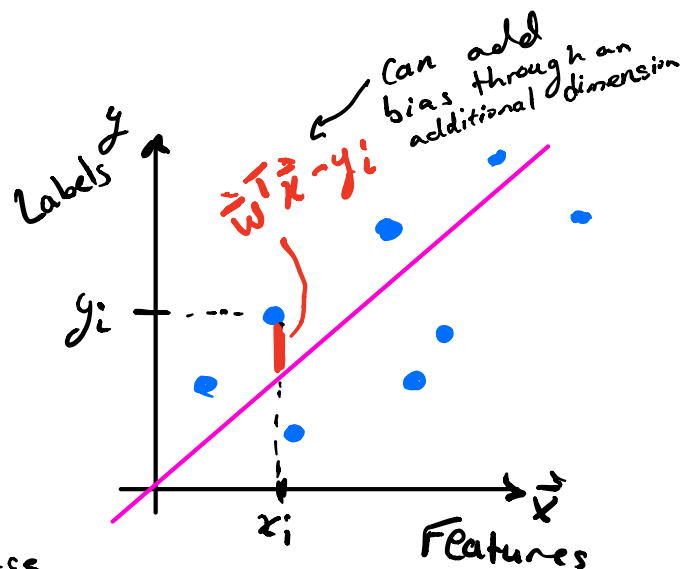
Data: $y_i \in \mathbb{R}$

Model: $y_i = \vec{w}^T \vec{x}_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

$$\Rightarrow y_i | \vec{x}_i \sim \mathcal{N}(\vec{w}^T \vec{x}_i, \sigma^2)$$

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

In words, this means we can assume the data is drawn from a "line" $\vec{w}^T \vec{x}$ through the origin. For each data point w/ features \vec{x}_i , the label y is drawn from a Gaussian w/ mean $\vec{w}^T \vec{x}_i$ and variance σ^2 .



Our task is to estimate the slope \vec{w} from the data.

Estimating with MLE

$$\vec{w} = \operatorname{argmax}_{\vec{w}} P((y_1, \vec{x}_1), (y_2, \vec{x}_2), \dots, (y_n, \vec{x}_n) | \vec{w})$$

$$= \operatorname{argmax}_{\vec{w}} \prod_{i=1}^n P(y_i, \vec{x}_i | \vec{w})$$

By independence

$$= \operatorname{argmax}_{\vec{w}} \prod_{i=1}^n P(y_i | \vec{x}_i, \vec{w}) P(\vec{x}_i | \vec{w})$$

$$\begin{aligned} P(A|B,C) P(B|C) \\ &= \frac{P(A \cap B \cap C)}{P(B \cap C)} \frac{P(B \cap C)}{P(C)} \\ &= P(A, B | C) \end{aligned}$$

$$= \operatorname{argmax}_{\vec{w}} \prod_{i=1}^n P(y_i | \vec{x}_i, \vec{w}) P(\vec{x}_i)$$

\vec{x}_i independent of \vec{w}
 $\neq i$

$$= \operatorname{argmax}_{\vec{w}} \prod_{i=1}^n P(y_i | \vec{x}_i, \vec{w})$$

$P(\vec{x}_i)$ is a constant

$$= \operatorname{argmax}_{\vec{w}} \sum_{i=1}^n \log(P(y_i | \vec{x}_i, \vec{w}))$$

log is a monotonic function

$$= \operatorname{argmax}_{\vec{w}} \sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\vec{x}_i^T \vec{w} - y_i)^2}{2\sigma^2}}\right)$$

$$= \operatorname{argmax}_{\vec{w}} \sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(e^{-\frac{(\vec{x}_i^T \vec{w} - y_i)^2}{2\sigma^2}}\right)$$

$$= \operatorname{argmax}_{\vec{w}} \sum_{i=1}^n -\frac{(\vec{x}_i^T \vec{w} - y_i)^2}{2\sigma^2}$$

ALWAYS wanna minimize

$$= \operatorname{argmin}_{\vec{w}} \frac{1}{n} \sum_{i=1}^n (\vec{x}_i^T \vec{w} - y_i)^2$$

optimize w/
gradient
descent

Ordinary
least squares

$\frac{1}{2\sigma^2}$ is a constant
which doesn't change
 \vec{w} . dividing by n instead
makes the loss more
tractable.

$$\text{Closed Form: } \vec{w} = (X^T X)^{-1} X^T \vec{y}; \text{ where } X = [\vec{x}_1, \dots, \vec{x}_n], \vec{y} = [y_1, \dots, y_n]$$

Estimating with MAP

Additional Model Assumptions:

$$P(\vec{w}) = \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{\vec{w}^T \vec{w}}{2\tau^2}} \quad \leftarrow \text{i.e. assume } w \sim \mathcal{N}(0, \tau^2)$$

MAP maximizes w from $P(w | \text{dataset})$

$$\vec{w} = \underset{w}{\operatorname{argmax}} P(\vec{w} | D)$$

Note: $D = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_n, y_n)\}$

$$= \underset{w}{\operatorname{argmax}} P(\vec{w} | (\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_n, y_n))$$

$$= \underset{w}{\operatorname{argmax}} \frac{P((\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_n, y_n) | \vec{w}) P(\vec{w})}{P((\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_n, y_n))} \quad \leftarrow \text{constant}$$

$$= \underset{w}{\operatorname{argmax}} P((\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_n, y_n) | \vec{w}) P(\vec{w})$$

$$= \underset{w}{\operatorname{argmax}} \prod_{i=1}^n [P((x_i, y_i) | \vec{w})] P(\vec{w})$$

$$= \underset{w}{\operatorname{argmax}} \prod_{i=1}^n [P(y_i | \vec{x}_i, \vec{w}) P(\vec{x}_i | \vec{w})] P(\vec{w})$$

$$= \underset{w}{\operatorname{argmax}} \prod_{i=1}^n [P(y_i | \vec{x}_i, \vec{w}) P(\vec{x}_i)] P(\vec{w})$$

$$= \underset{w}{\operatorname{argmax}} \prod_{i=1}^n [P(y_i | \vec{x}_i, \vec{w})] P(\vec{w})$$

$$= \underset{w}{\operatorname{argmax}} \sum_{i=1}^n \log(P(y_i | \vec{x}_i, \vec{w})) + \log(P(\vec{w}))$$

$$= \underset{w}{\operatorname{argmax}} \sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\vec{x}_i^T \vec{w} - y_i)^2}{2\sigma^2}}\right) + \log\left(\frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{\vec{w}^T \vec{w}}{2\tau^2}}\right)$$

$$= \underset{\vec{w}}{\operatorname{argmax}} \sum_{i=1}^n \left[-\frac{1}{2\sigma^2} (\vec{x}_i^T \vec{w} - y_i)^2 - \frac{\vec{w}^T \vec{w}}{2\tau^2} \right]$$

$$= \underset{\vec{w}}{\operatorname{argmin}} \sum_{i=1}^n (\vec{x}_i^T \vec{w} - y_i)^2 - \frac{n\sigma^2}{\tau^2} \vec{w}^T \vec{w}$$

$$= \underset{\vec{w}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (\vec{x}_i^T \vec{w} - y_i)^2 - \lambda \vec{w}^T \vec{w} \quad ; \quad \lambda = \frac{\sigma^2}{n\tau^2}$$

↑

Known as ridge regression. Differs from OLS in that there's L_2 regularization

Closed form: $\vec{w} = (X X^T - \lambda I)^{-1} X \vec{y}^T$

where $X = [x_1, \dots, x_n], y = [y_1, \dots, y_n]$