Assumptions

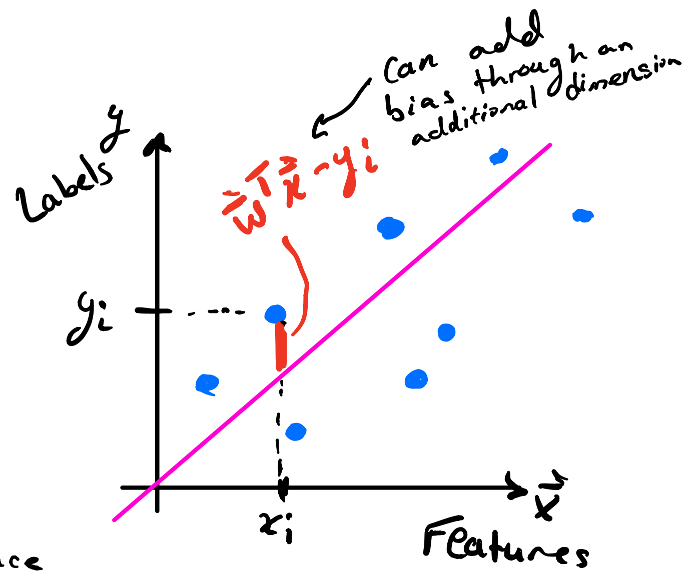
Data: $y_i \in \mathbb{R}$

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

$$\Rightarrow y \mid \bar{x}_i \sim \mathcal{N}(\bar{w}^T \bar{x}_i, \sigma^2)$$

$$\Rightarrow \mathbb{P}(y \mid \bar{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\bar{x}_i^T \bar{w} - y_i)^2}{2\sigma^2}}$$

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

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

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

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

$$\mathbb{P}(A|B,C)\,\mathbb{P}(B|C) = \frac{\mathbb{P}(A\cap B\cap C)}{\mathbb{P}(B\cap C)}\frac{\mathbb{P}(B\cap C)}{\mathbb{P}(C)} = \mathbb{P}(A,B|C)$$

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

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

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

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

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

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

optimize w/ gradient descent

ALWAYS wanna minimize

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

Closed Form: $\vec{w} = (XX^T)^{-1}X\vec{y}^T$; where $X=[x_1,\dots,x_n]$, $y=[y_1,\dots,y_n]$

Estimating with MAP

Additional Model Assumptions:

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

i.e assume $w \sim \mathcal{N}(0, \tau^2)$

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

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

$$\vec{w} = \underset{w}{\operatorname{argmax}} \; \mathbb{P}(\vec{w} \mid D)$$

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

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

constant

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

$$= \underset{w}{\operatorname{argmax}} \; \prod_{i=1}^{n} \left[\mathbb{P}((x_i, y_i) \mid \vec{w}) \right] \mathbb{P}(\vec{w})$$

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

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

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

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

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

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

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

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

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

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

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