Supervised learning $\{x^{(i)} \in \mathbb{R}^D, t^{(i)} \in \mathbb{R}\}_{i=1}^N$ = training set

→ linear regression

through ordinary least squares (OLS) / Residual Sum of Squares

general model $h_\beta(x) = \beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}$

loss $l(h_\beta) = l(\beta) = \frac{1}{2N} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2$ (*)

→ to learn $h_\beta(x)$ from $\{x^{(i)} \in \mathbb{R}^D, t^{(i)} \in \mathbb{R}\}_{i=1}^N$

We can minimize (*) through Gradient descent.

Concretely we start from random initial guess for β_j

We then move iteratively in the direction $d = -\operatorname{grad}_{\beta} \ell(\beta)$

$$\beta^{(k+1)} \leftarrow \beta^{(k)} - \eta \operatorname{grad}_{\beta} \ell(\beta)$$

$$\beta^{(k+1)} \leftarrow \beta^{(k)} - \eta \frac{1}{N} \sum_{i=1}^{N} (t^{(i)} - h_\beta(x^{(i)})) \cdot (-\tilde{x}^{(i)})$$

$$\tilde{x}^{(i)} = [1, x_1^{(i)}, \dots, x_D^{(i)}] \qquad \beta = [\beta_0, \beta_1, \dots, \beta_D]$$

$$h_\beta(x) = \beta_0 + \sum_{j=1}^{D} \beta_j x_j = \beta^T \tilde{x} = [\beta_0 \ \beta_1 \ \cdots \ \beta_D] \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_D \end{bmatrix}$$

2 main possibilities known as Batch gradient descent

(Compute loss/gradient taking into account all the examples from the training set)

Stochastic gradient descent

(loop over the samples and perform one gradient step for each $\{t^{(i)}, x^{(i)}\}$ pair)

in this setting, one pass over the whole training set corresponds to one epoch.

Today → Alternative approach through Normal equations

→ Statistical intuition on the least squares loss
(Why is it a good idea to use the OLS loss?)

→ Non linear data

→ Bias Variance tradeoff

→ Regularization

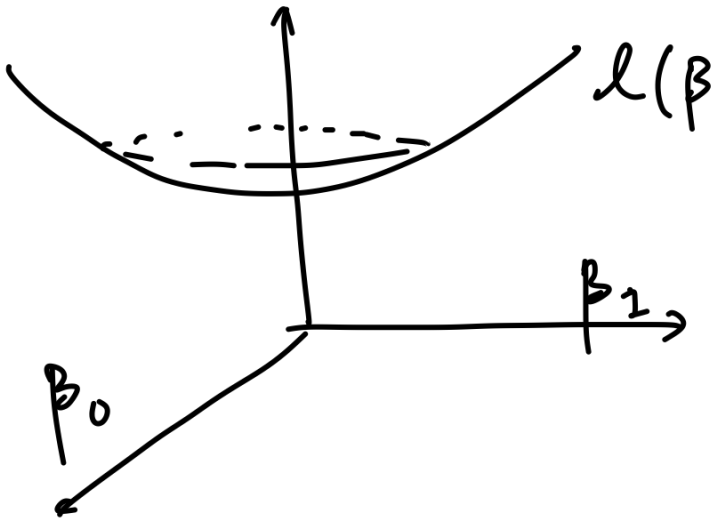
OLS loss: $$l(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}\right)\right)^2$$

$$= \frac{1}{N} \sum_{i=1}^{N} e_i^2 = \frac{1}{N} \vec{e}^T \vec{e}$$

$$e_i = \left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}\right)\right), \quad \tilde{x}^{(i)} = [1, x^{(i)}]$$

$$= \left(t^{(i)} - \beta^T \tilde{x}^{(i)}\right)$$

$$\vec{e} = [e_1, e_2, \dots, e_N]$$



Approach #2

At the β that minimizes $\ell(\beta)$ the derivative must be zero

so in particular, to find the β that minimizes $\ell(\beta)$ we can compute the expression of $\text{grad}\,\ell(\beta) = \left[\frac{\partial \ell}{\partial \beta_0}, \frac{\partial \ell}{\partial \beta_1}, \ldots, \frac{\partial \ell}{\partial \beta_D}\right]$ and then

Solve the equation $\text{grad}\,\ell(\beta) = 0$ for β.

Recall that

$$\ell(\beta) = e^T e = \sum_{i=1}^{N} (t^{(i)} - \beta^T \tilde{x}^{(i)})^2$$

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} = \begin{bmatrix} t^{(1)} - \beta^T \tilde{x}^{(1)} \\ t^{(2)} - \beta^T \tilde{x}^{(2)} \\ \vdots \\ t^{(N)} - \beta^T \tilde{x}^{(N)} \end{bmatrix} = \begin{bmatrix} t^{(1)} \\ \vdots \\ t^{(N)} \end{bmatrix} - \begin{bmatrix} 1 & (x^{(1)})^T \\ 1 & (x^{(2)})^T \\ \vdots & \\ 1 & (x^{(N)})^T \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_D \end{bmatrix}$$

$$= \begin{bmatrix} t^{(1)} \\ \vdots \\ t^{(N)} \end{bmatrix} - \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_D^{(1)} \\ \vdots & & & & \\ 1 & x_1^{(N)} & x_2^{(N)} & \dots & x_D^{(N)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_D \end{bmatrix}$$

$$= \vec{t} - \underline{\underline{\tilde{X}}}\, \vec{\beta} \qquad \hookrightarrow \underline{\underline{\tilde{X}}}$$

$\vec{t}$ = target vector

$\underline{\underline{\tilde{X}}}$ = feature matrix (whose rows are the feature vectors augmented with a '1')

From this, $l(\beta)$ can read as

$$l(\beta) = \frac{1}{2N} e^T e = \frac{1}{2N} (\vec{t} - \underline{\underline{\tilde{X}}}\beta)^T (\vec{t} - \underline{\underline{\tilde{X}}}\beta)$$

$$\underline{\underline{\tilde{X}}} = \begin{bmatrix} — \tilde{x}^{(1)} — \\ \vdots \\ — \tilde{x}^{(N)} — \end{bmatrix}$$

$$\ell(\beta) = \frac{1}{2N} t^T t + \frac{1}{2N} \beta^T \tilde{\underline{X}}^T \tilde{\underline{X}} \beta - \frac{1}{2N} \vec{t}^T \tilde{\underline{X}} \beta - \frac{1}{2N} \beta^T \tilde{\underline{X}} \vec{t}$$

$$= \frac{1}{2N} t^T t + \frac{1}{2N} \beta^T \tilde{\underline{X}}^T \tilde{\underline{X}} \beta - \frac{1}{N} \vec{t}^T \tilde{\underline{X}} \beta$$

does not depend on β ②

① term has the form $w^T \beta$ with $w = \vec{t}^T \tilde{\underline{X}}$

$$\text{grad} = \left[\frac{\partial \ell}{\partial \beta_0}, \dots, \frac{\partial \ell}{\partial \beta_D} \right]$$

① gradient of $w^T \beta = \vec{t}^T \tilde{\underline{X}} \beta$

$$= w_0 \beta_0 + w_1 \beta_1 + \dots + w_D \beta_D$$

$$\frac{\partial(w^T\beta)}{\partial\beta_0} = w_0 \qquad \frac{\partial(w^T\beta)}{\partial\beta_1} = w_1 \quad \cdots \quad \frac{\partial(w^T\beta)}{\partial\beta_D} = w_D$$

$$\rightarrow \underset{\beta}{\text{grad}}\left(\vec{t}^T \underline{\underline{\tilde{X}}}\,\beta\right) = \left[\frac{\partial\, \vec{t}^T \underline{\underline{\tilde{X}}}\,\beta}{\partial\beta_0} \quad \cdots \quad \frac{\partial\, \vec{t}^T \underline{\underline{\tilde{X}}}\,\beta}{\partial\beta_D}\right]$$

$$= \underline{\underline{\tilde{X}}}^T \vec{t} \leftarrow (1)$$

term (2) gradient of $\beta^T \underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}}\,\beta$ with respect to β ?

$$\rightarrow [\beta_0, \beta_1] \begin{bmatrix} M_{11} & M_{12} \\ M_{12} & M_{22} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \beta_0^2 M_{11} + \beta_1\beta_0 M_{12} + \beta_1^2 M_{22} + \beta_1\beta_0 M_{12}$$

$\rightarrow$ take $\frac{\partial \ell}{\partial \beta_0} = 2\beta_0 M_{11} + 2\beta_1 M_{12}$

$\frac{\partial \ell}{\partial \beta_1} = 2\beta_1 M_{22} + 2\beta_0 M_{12}$

$$\text{grad}\, \ell = \begin{bmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \frac{\partial \ell}{\partial \beta_1} \end{bmatrix} = 2 \underbrace{M}\, \underbrace{\beta}$$

For ② gradient is simply given by $2 \underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}}^T \beta \frac{1}{2N}$

Combining ① & ② the gradient of $\ell(\beta)$ w.r.t β is given

by $\frac{1}{N} \underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}} \beta - \frac{1}{N} \underline{\underline{\tilde{X}}}^T \vec{t} = \underset{\beta}{\text{grad}}\, \ell(\beta)$

In order to find the β that minimizes $l(\beta)$ we set grad $l(\beta)$ to 0 and solve the resulting equations

$$\frac{1}{N} \tilde{\underline{X}}^T \tilde{\underline{X}} \beta - \frac{1}{N} \tilde{\underline{X}}^T \vec{t} = 0$$

$$\tilde{\underline{X}}^T \tilde{\underline{X}} \beta = \tilde{\underline{X}}^T \vec{t}$$

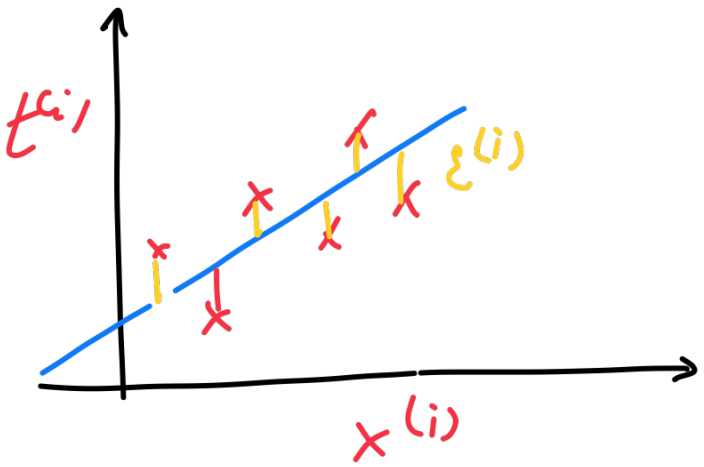
Normal equations

$$\beta = \left(\tilde{\underline{X}}^T \tilde{\underline{X}} \right)^{-1} \tilde{\underline{X}}^T \vec{t} \quad (**)$$

→ provided $\tilde{\underline{X}}^T \tilde{\underline{X}}$ is invertible, we can compute β using $(**)$

Why is it a good idea to choose β as the model that minimizes the least squares loss?

Assume ① $t^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_D x_D^{(i)} + \varepsilon^{(i)}$



Further assume ② $\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma)$

Finally assume ③ $\varepsilon^{(i)}$ independent.

(i.i.d
independent
identically distributed)

Only randomness in $t^{(i)}$'s is due to $\varepsilon^{(i)}$. In particular the probability of observing a particular $t^{(i)}$ is given by

$$p(t^{(i)}) = p(\varepsilon^{(i)} = t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_D x_D^{(i)}))$$

Assumptions ① + ②

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - \beta^T \tilde{x}^{(i)})^2}{2\sigma^2}\right)$$

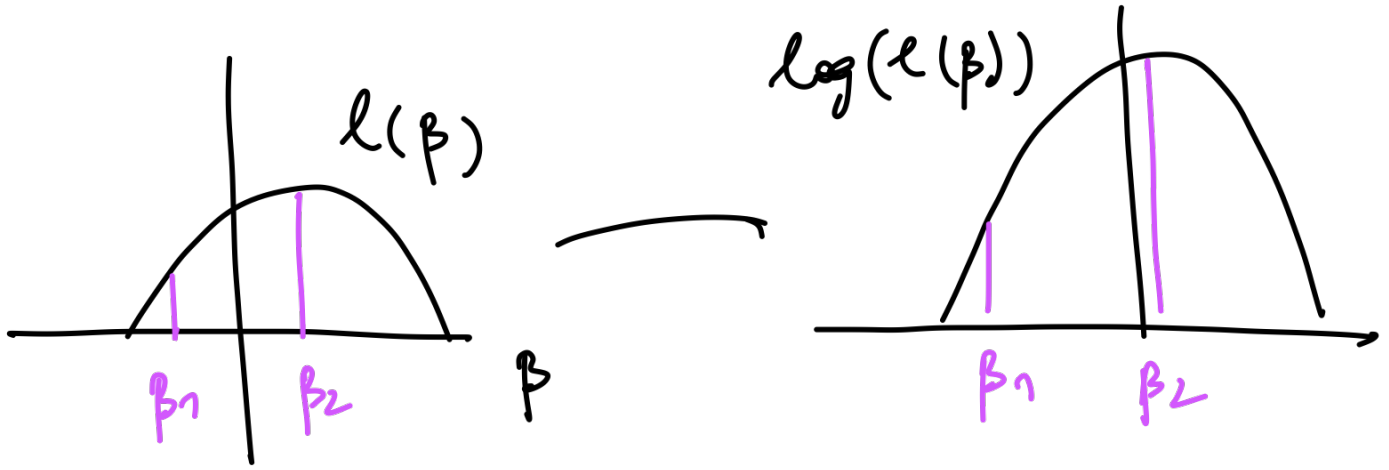
From Assumption ③ the probability of observing $\{t^{(i)}\}_{i=1}^N$ given $\{x^{(i)}\}_{i=1}^N$ is given by

When viewed as a function of β = likelihood

$$p(\{t^{(i)}\}_{i=1}^N \mid x^{(i)}) = \prod_{i=1}^N p(t^{(i)}) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - \beta^T \tilde{x}^{(i)})^2}{2\sigma^2}\right) \quad (*)$$

Given a set of observations $\{t^{(i)}\}_{i=1}^{N}$ and a distribution on those observations $p(\{t^{(i)}\}_{i=1}^{N})$ that is parameterized by β a good approach is to look for the value of β that leads to $p(\{t^{(i)}\}_{i=1}^{N})$ being maximized.

in order to maximize (*) with respect to β we first get rid of the product by taking the log

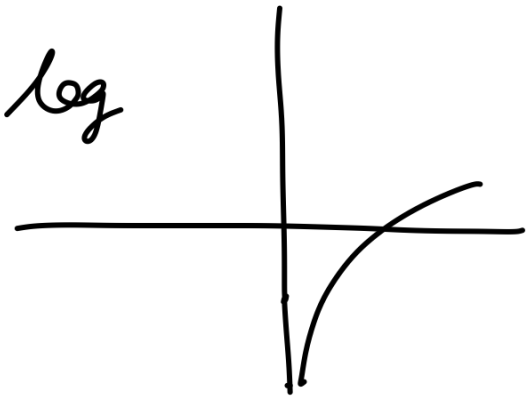


for β_1, β_2

$\ell(\beta_1) < \ell(\beta_2)$

$\log(\ell(\beta_1)) <$

$\log(\ell(\beta_2))$



$\Rightarrow$ As log is an increasing function, looking for $\underset{\beta}{\operatorname{argmax}}\, \ell(\beta)$ is equivalent to $\underset{\beta}{\operatorname{argmax}}\, \log(\ell(\beta))$

Applying this idea to (*)

$$\log\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - \beta^T \tilde{x}^{(i)})^2}{2\sigma^2}\right)\right) \rightarrow \text{log likelihood}$$

$$= \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(t^{(i)} - \beta^T \tilde{x}^{(i)})^2}{2\sigma^2}$$

does not depend on β

We can then find β^* by maximizing

$$LL = -\sum_{i=1}^{N} \frac{(t^{(i)} - \beta^T \tilde{x}^{(i)})^2}{2\sigma^2} \leftarrow \beta_{MLE}$$

Maximum likelihood estimator for β

→ the β that maximize the likelihood is the β that minimizes the OLS loss

When the $\epsilon^{(i)}$'s are generated following assumptions ① to ③ minimizing the OLS loss will return the maximum likelihood estimates for β.