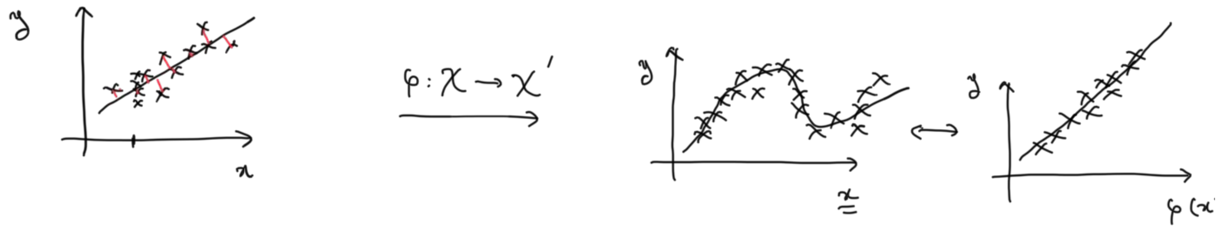
Linear Regression



Setup: Given n data-points $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, $i = 1, \ldots, n$

Goal: Choose/find/"learn" $f: \mathbb{R}^d \to \mathbb{R}$ to predict the value of y for any new given input x.

e.g. $f(x) = w^T x$, $w \in \mathbb{R}^d$

$$= \sum_{i=1}^{d} w_i x_i$$

φ: identity map, i.e $\varphi(x) = x$

e.g. $f(x) = w^T \varphi(x)$

$$= \sum_{j=1}^{m} w_j \varphi_j(x)$$

$\varphi: \mathbb{R}^d \to \mathbb{R}^m$ ↳ feature's space with m-featu

$$\varphi(x) = (\underbrace{\varphi_1(x), \varphi_2(x), \ldots, \varphi_m(x)}_{\text{features / basis functions}})$$

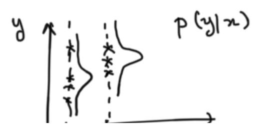
(*) Linear means that the model depends linearly in its parameters.

Choosing a model:

Assume an observation rule, typically taking the form:

$$y = f(x) + \varepsilon$$

Since y may be corrupted by noise, it is natural to model it as a random variable. Then our goal to charactirize its distribution: $p(y|x)$



We'll have to assume a family $p_\theta(y|x)$ parametrized by θ, an then we'll try to estimate

x ϑ given the observed data $\mathcal{D}$.

- Which family to use for $p_\vartheta(y|x)$? ... our first would be a Gaussian

i.e. $p_\vartheta(y|x) = \mathcal{N}(y \mid \mu(x), \sigma^2(x))$, $\vartheta := \{\mu, \sigma^2\}$ ↪ functions in $\mathbb{R}^d$.

a Gaussian random variable — mean — variance

- What $\mu(x)$ and $\sigma^2(x)$ to use?

The simplest choice is to assume that $\sigma^2(x)$ are constants (i.e. independent of x).

$$P_\vartheta(y|x) = \mathcal{N}(y \mid \mu(x), \sigma^2) \,, \quad \mu(x) = W^T x \,, \quad \sigma^2(x) = \sigma^2$$

unknown parameters $\vartheta := \{w_1, \dots, w_d, \sigma^2\}$

<u>Recap:</u>

Our starting point was that: $y = f(x) + \varepsilon$

If we assume $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and $f(x) = w^T x$, then this implies that:

$$\underbrace{p(y|x) = \mathcal{N}(y \mid w^T x, \sigma^2)}_{\text{quantifies the "likelihood" of the observed data according to the observation model we have assumed}} \iff \boxed{y = w^T x + \varepsilon, \ \varepsilon \sim \mathcal{N}(0, \sigma}$$

- Modeling choices we've made so far: $\left\{ \begin{array}{l} \text{mean } \mu(x) \\ \text{latent function } f(x) \end{array} \right. = w^T x \}$, $\{\varepsilon \sim \mathcal{N}(0,$

⊛ How to estimate the unknown model parameters ϑ?

... via Maximum Likelihood Estima (MLE)

<u>Setup:</u> Given $\mathcal{D} := \{(x_1, y_1), \dots, (x_n, y_n)\}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, $\vartheta = \{w, \sigma$

<u>Model:</u> $\boxed{y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(y_i \mid \underline{w^T x_i}, \underline{\sigma^2})}$, i.e. our observed outputs are distributed according to a Gaussian

i.i.d. = independent and identically distributed identically distributed likelihood.

<u>Then</u> : $\theta_{MLE} = \arg\max_{\theta} p(\mathcal{D}|\theta)$,

The likelihood is : $p(\mathcal{D}|\theta) = p(y|x,\theta)$

$$= \underbrace{p\left(y_1, y_2, \dots, y_n \,\middle|\, x_1, x_2, \dots, x_n, w_1, w_2, \dots, w_d, \sigma^2\right)}_{\text{joint distribution of all observed outputs, given the corresponding inputs and model parameters!}}$$

$$\overset{\text{i.i.d.}}{=} \prod_{i=1}^{n} p(y_i|x_i,\theta) = \prod_{i=1}^{n} \underbrace{\mathcal{N}(y_i \mid w^T x_i, \sigma^2)}_{\text{Gaussian pdf}}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - w^T x_i)^2\right]$$

can be re-written in vectorized form :

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{n}(y_i - w^T x_i)^2}_{\text{sum of squares}}\right]$$

$$\sum_{i=1}^{n}(y_i - w^T x_i)^2 = (y - Xw)^T(y - Xw) \quad (1)$$

where : $\underset{n\times 1}{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $\underset{n\times d}{X} = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & & \\ x_{n1} & \cdots & x_{nd} \end{bmatrix}$, $\underset{d\times 1}{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$

Recall, we want to compute $\theta_{MLE} = \arg\max_{\theta} p(\mathcal{D}|\theta)$

$$\Updownarrow$$

$$\theta_{MLE} = \arg\min -\log p(\mathcal{D}|\theta)$$

$(1) \implies -\log p(\mathcal{D}|\theta) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(y - Xw)^T(y - Xw) := \mathcal{L}(\theta$

$\theta := \{w, \sigma$

We need to compute the critical points of $\mathcal{L}(w)$

i.e. the points for which $\nabla_w \mathcal{L}(w) = 0$.

Let's focus on estimating w_{MLE} (i.e. for now assume that σ^2 is know

<u>First notice</u> : $\frac{1}{2}(y-Xw)^T(y-Xw) = \frac{1}{2}(y^Ty - y^TXw - (Xw)^Ty + (Xw)^T(Xw))$

$$= \frac{1}{2}y^Ty - \underbrace{y^TXw}_{w^TX^Ty} + w^TX^TXw$$

Now we can solve for :

$$\nabla_w \mathcal{L}(w) = 0 \implies \boxed{w_{MLE} = (X^TX)^{-1}X^Ty}$$

* $(X^TX)^{-1}$ need to be invertible.
$d \times d$.

X^TX is invertible if the calums of X are linearly independent.

To check whether w_{MLE} is actually a minimizer we can examine the second-derivative of our loss $\mathcal{L}(w)$:

$$\nabla_w^2 \mathcal{L}(w) = X^TX$$ $\rightarrow$ this is a symmetric positive-definite matrix, which implies that indeed w_{MLE} is a (global) minimum

<u>Remark</u> :

In the case of linear regression with basis functions :

$$w_{MLE} = (\Phi^T\Phi)^{-1}\Phi^Ty \quad , \quad \varphi : \underset{\text{input space}}{\mathbb{R}^d} \longrightarrow \underset{\text{feature space}}{\mathbb{R}^m}$$

where

$$\underset{n \times m}{\Phi} = \begin{bmatrix} \varphi_1(x_1) & \cdots & \varphi_m(x_1) \\ \vdots & & \\ \varphi_1(x_n) & \cdots & \varphi_m(x_n) \end{bmatrix}$$