Feb2

Tuesday, February 2, 2021 15:55

"Parametric" (Model Based) vs. "Non-Parametric" (Memory Based) Algorithm

LR: $f(x) = \theta_0 + \theta_1 x_1$ $\quad Y_i = f(\theta) + \varepsilon_i$

① fixed parameters

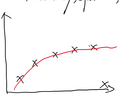
② once we train the model and get the parameters we can erase the training data. ← faster computation when predicting

③ strong assumptions e.g. for LR: Linearity (of parameter) Normality, equal var, Idependent

Non-linear: $f(x) = \frac{\beta_0 x}{\beta_1 + x}$

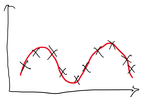
Example:

$f(x) = \theta_0 + \theta_1 x_1$



$f(x) = \theta_0 + \theta_1 \sqrt{x_1}$

$f(x) = \theta_0 + \theta_1 \overline{x_1}$



$f(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_5 x^5$

Choice of the features is very important!

high dimensional data, hard to visualize and hard to choose the features manually.

KNN:

① keep the entire training data.

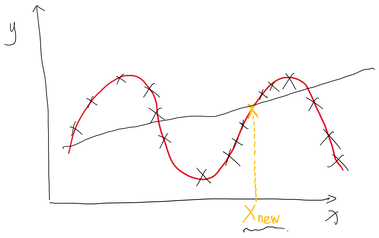
② The amount of stuff we need to keep grows linearly with the size of the training data.

③ Less assumptions

Intuition

Locally Weighted Regression (feature selection is less critical)

E.g.



Local area.

Normally in LR:

$\min_{\theta} J(\theta) = \sum_{i=1}^{m} (f_\theta(x^{(i)}) - y^{(i)})^2$

output $x_{new}^T \theta$

In Locally weighted LR: ← weight / kernel function

we minimize θ $J(\theta) = \sum_{i=1}^{m} w^{(i)} (f_\theta(x^{(i)}) - y^{(i)})^2$

output $x_{new}^T \theta$

What we expect?

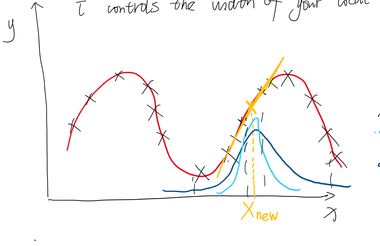
If $|x^{(i)} - x_{new}|$ is small, then $w^{(i)} \approx 1$

If $|x^{(i)} - x_{new}|$ is large, then $w^{(i)} \approx 0$.

$w^{(i)} = \exp\left(\frac{-(x^{(i)} - x_{new})^2}{2\tau^2}\right)$, where τ is a hyperparameter.

Gaussian kernel

τ controls the width of your local area.



Local area.

"~" τ is small

"~" τ is large

$W = \begin{bmatrix} w_1 & & & \\ & w_2 & & \\ & & \ddots & \\ & & & w_m \end{bmatrix}_{m \times m}$

$X_{m \times n}$ design matrix

$J(\theta) = (X\theta - Y)^T W (X\theta - Y)$ weighted Linear regression

$\hat{\theta} = (X^T W X)^{-1} X^T W Y$