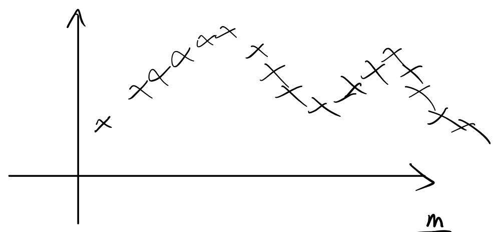
Locally weighted regression.

Non-parametric Learing algorithm: keep all the traing set in memory.



To evalute h for certain x

Fit $\cancel{\theta}$ ~~to~~ minimize $\frac{1}{2}\sum_i (y^{(i)} - \theta^T x^{(i)})^2$

Fit θ to minimize $\sum_{i=1}^{m} w^{(i)} (y^{(i)} - \theta^T x^{(i)})^2$

$\downarrow$

weighting function $= \exp\left(-\frac{(x^{(i)} - x)^2}{2}\right)$

if $|x^{(i)} - x|$ is small, $w^{(i)} \approx 1$

if $|x^{(i)} - x|$ is large, $w^{(i)} \approx 0$

Why squares?

Assume $y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$, $\boxed{\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)}$

$$P(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\varepsilon^{(i)2}}{2\sigma^2}\right)$$

$$\Rightarrow p(y^{(i)} | x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y^{i} - \theta^T x^{i})^2}{2\sigma^2}\right)$$

$\downarrow$

parameterized by θ, not a r.v.

$$\mathcal{L}(\theta) = \prod_{i=1}^{m} p(y^{(i)} | x^{(i)}; \theta)$$

"Likelihood of parameters" = "Probability of data"

$$= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$l(\theta) = \log(\mathcal{L}(\theta))$$

$$= \sum_{i=1}^{m} \left[\log \frac{1}{\sqrt{2\pi}\sigma} + \log \exp(\cdots)\right]$$

$$= m \cdot \log \frac{1}{\sqrt{2\pi}\sigma} + \sum_{i=1}^{m} - \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}$$

MLE: maximum likelihood estimation

$\downarrow$

log - likeli hood

$\downarrow$

$\frac{1}{\ }\sum^{m} (y^{(i)} - \theta^T x^{(i)})^2$

choose θ to minimize: $\frac{1}{2}\sum_{i=1}(y^{(i)}-\theta^T x^{(i)})^2$.

$\Downarrow$

Cost function Linear Regression.

Assumption: error terms are Gaussian & IID

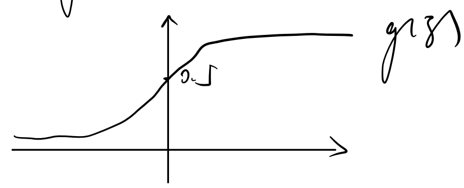
Classification problem. (Binary)

Logistic Regression.

Want $h_\theta(x) \in [0,1]$

$$g(z) = \frac{1}{1+e^{-z}}$$

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$



$$p(y=1 \mid x;\theta) = h_\theta(x)$$
$$p(y=0 \mid x;\theta) = 1-h_\theta(x)$$
$$> \; p(y \mid x;\theta) = h_\theta(x)^y(1-h_\theta(x))^{1-y}$$

$$L(\theta) = \prod_{i=1}^{m} h_\theta(x^{(i)})^{y^{(i)}}(1-h_\theta(x^{(i)}))^{1-y^{(i)}}$$

$$l(\theta) = \log L(\theta)$$
$$= \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))$$

Choose θ to maximize $l(\theta)$

Batch gradient descent: $\theta_j = \theta_j + \alpha \frac{\partial}{\partial \theta_j} l(\theta)$ $\Leftarrow$ Logistic Reg

See lecture notes

$\Updownarrow$

$$\theta_j = \theta_j + \alpha \sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)}$$

$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$ $\Leftarrow$ Liner Reg

$\downarrow$

Same form with Linear Regression, but with different definition of $h_\theta(x)$