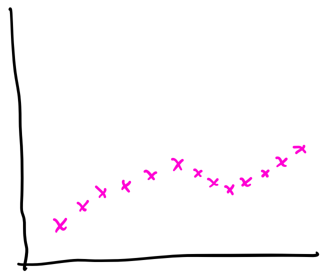
Locally Weighted Linear Regression



To evaluate h at a certain x:

LR does

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

LWLR does

1. Fit θ to minimize $\frac{1}{2} \sum_i w^{(i)} (y^{(i)} - \theta^T x^{(i)})^2$
2. Return $\theta^T x$

$w^{(i)}$ weight:

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

τ large

τ small

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

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

Note: algorithm is prone to overfitting, but useful for situations with lots of low-density data

Probabilistic Understanding of Linear Regression | Least Squares

Assume $y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$ ← error (unmodeled effects, random noise)

where $\underbrace{\varepsilon^{(i)} \sim N(0, \sigma^2)}_{\text{i.i.d}}$

Given $x^{(i)}, \theta$ $y^{(i)}$ distributed $N(\theta^T x^{(i)}, \sigma^2)$

$\Rightarrow y^{(i)} \mid x^{(i)}; \theta \sim N(\theta^T x^{(i)}, \sigma^2)$

↑ is parametrized by

$$\mathcal{L}(\theta) = p(\vec{y} \mid x; \theta)$$
$$= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)$$
$$= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

How to estimate θ?

Maximum Likelihood: Choose θ that makes train data as probable as possible

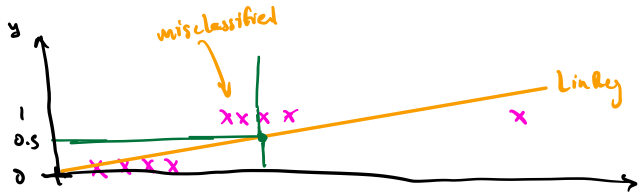
Likelihood of parameters: probability of data (y)

Choose θ to maximize $\mathcal{L}(\theta)$

$$\ell(\theta) = \log \mathcal{L}(\theta) \qquad \text{"log likelihood"}$$
$$= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(\frac{-(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$
$$= \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right) + \log \exp(\ldots)$$
$$= \underbrace{m \log\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)}_{\text{doesn't depend on } \theta} + \underbrace{\sum_{i=1}^{m} -\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}}_{\text{max only this}}$$

$\Rightarrow$ Maximizing $\ell(\theta)$ is the same as minimizing $\frac{1}{2\cancel{\sigma^2}} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2 = J(\theta)$.

Logistic Regression.

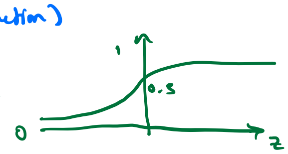


LinReg can output numbers $<< 0$ or $>> 1$

With LogReg, want $h_\theta(x)$ to be in $[0, 1]$

Logistic function (sigmoid function)

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



Probabilistic interpretation

$$h_\theta(x) = p(y=1 \mid x; \theta)$$

$$1 - h_\theta(x) = p(y=0 \mid x; \theta)$$

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

If $y=1$: $h_\theta(x)^1 (1-h_\theta(x))^0 = p(y=1 \mid x; \theta)$

If $y=0$: $h_\theta(x)^0 (1-h_\theta(x))^1 = p(y=0 \mid x; \theta)$

Maximum Likelihood

$$\mathcal{L}(\theta) = p(\vec{y} \mid x; \theta)$$

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

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

Log likelihood:

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

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

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

Want: find θ to maximize $\ell(\theta)$; use gradient ascent

Gradient ascent.

Like gradient descent, but uphill

$$\theta_j := \theta_j + \alpha \frac{\partial}{\partial \theta_j} \ell(\theta)$$

Repeat until converge: {

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

$h_\theta(x) = g(\theta^T x)$

for $j = 0 \ldots n$

}

Optimization looks the same as linreg, but the parameters are used differently.

Newton's Method.

To maximize $\ell(\theta)$ $\theta \in \mathbb{R}$

find θ s.t. $f(\theta) = \frac{d}{d\theta} \ell(\theta) = 0$

Repeatedly:

$$\theta^{(t+1)} := \theta^{(t)} - \frac{f(\theta^{(t)})}{f'(\theta^{(t)})} = \theta^{(t)} - \frac{\ell'(\theta^{(t)})}{\ell''(\theta^{(t)})}$$

This converges very quickly!

$$\theta^{(t+1)} = \theta^{(t)} - H^{-1} \nabla_\theta \ell$$

where $H \in \mathbb{R}^{n+1 \times n+1}$

$$H_{ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\theta)$$