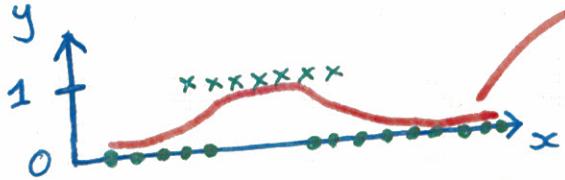
Regressing on Labels

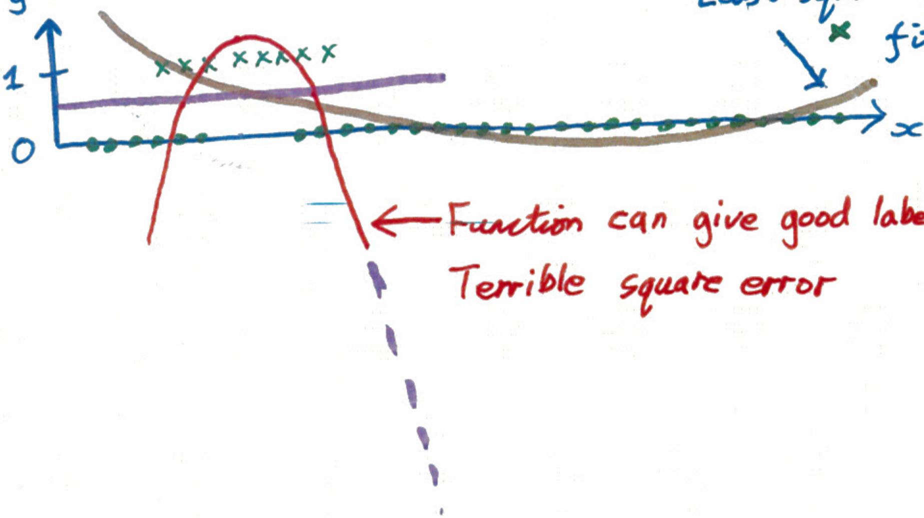


$f(x) \approx P(y=1|x)$

(If enough data and basis functions, linear least squares works!)

Often bad idea:

Least squares quadratic fit



Function can give good labels
Terrible square error

Gradients for Least Squares

Residuals $\underline{r} = \underline{y} - X\underline{w}$

NxD design matrix of inputs

Dx1 parameters

Nx1 training labels

Cost $\underline{r}^T\underline{r} = (\underline{y} - X\underline{w})^T(\underline{y} - X\underline{w})$

$$= \underline{y}^T\underline{y} - 2\underline{w}^T X^T \underline{y} + \underline{w}^T X^T X \underline{w}$$

Gradients $\nabla_{\underline{w}}[\underline{r}^T\underline{r}] = -2X^T\underline{y} + 2X^TX\underline{w}$

$$\left[= -2X^T \underbrace{(\underline{y} - X\underline{w})}_{\underline{r}} \right]$$

Gradient descent:

$$\underline{w} \leftarrow \underline{w} - \eta \nabla_{\underline{w}}[\underline{r}^T\underline{r}]$$

step-size, "small" 0.01?

Normal Equations approach

$\nabla_{\underline{w}} [\underline{r}^T \underline{r}] = \underline{0}$ at least min. squares solution

$$(X^T X)\underline{w} = X^T y$$

So if the best $\underline{w}$ is unique:

$$\underline{w} = \underbrace{(X^T X)^{-1} X^T}_{\text{Pseudo-Inverse}} y$$

$$X^{-1} \underbrace{\cancel{X^{-T} X^T}}_{I} y$$

$$\cancel{\underline{w} = X^{-1} y}$$

"$\underline{w} = X \backslash y$"

Matlab:

~~inv(X' * X) * (X' * y)~~

(X' * X) \ (X' * y)

More accurate than

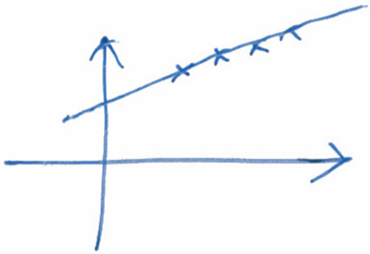
R puzzle

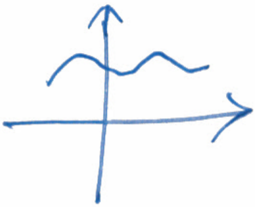
						bias
"red"	$\rightarrow$	1	0	~~0~~	why?	1
"blue"	$\rightarrow$	0	1	~~0~~		1
"green"	$\rightarrow$	0	0	~~1~~		1

For unique $\underline{w}$ to min. sum of squares

$(X^T X)$ is invertible

$\rightarrow$ Full rank.



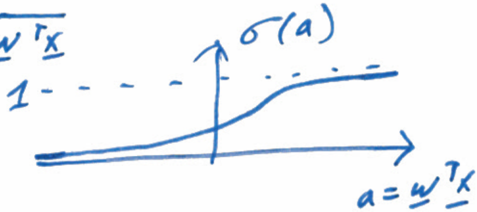


Logistic Regression

could $\phi(\underline{x})$

$$f(\underline{x};\underline{w}) = \sigma(\underline{w}^T\underline{x}) \qquad f \in [0,1]$$

$$= \frac{1}{1+e^{-\underline{w}^T\underline{x}}}$$



Loss Function

Could use square loss again:

$$\sum_{n=1}^{N} \left(y^{(n)} - f(\underline{x}^{(n)};\underline{w})\right)^2$$

Normal interpretation:

$$p(y=1\,|\,\underline{x}) \simeq f(\underline{x};\underline{w})$$

Could maximize likelihood of the parameters.

Likelihood prob. of data given the parameters.

Prob $p(\{y^{(n)}\}\,|\,X,\underline{w}) = \prod_n p(y^{(n)}\,|\,\underline{x}^{(n)},\underline{w})$

Or minimize negative log probability

$$NLL = -\sum_{n:\,y^{(n)}=1} \log \sigma(\underline{w}^T\underline{x}) - \sum_{n:\,y^{(n)}=0} \log(1-\sigma(\underline{w}^T\underline{x}))$$

I like to make labels $\{-1, +1\}$

$$z^{(n)} = 2y^{(n)} - 1$$

Useful fact:

$$1 - \sigma(a) = \sigma(-a)$$

$$NLL = -\sum_{n=1}^{N} \log \underbrace{\sigma(z^{(n)}\underline{w}^T\underline{x}^{(n)})}_{\text{Prob. of being correct, } \sigma_n}$$

$$\nabla_{\underline{w}} NLL = -\sum_{n=1}^{N} \nabla_{\underline{w}} \log \sigma_n$$

$$= -\sum_{n=1}^{N} \frac{1}{\sigma_n} \nabla_{\underline{w}} \sigma_n \qquad \text{(Chain rule)}$$

$$= -\sum_{n=1}^{N} \frac{1}{\cancel{\sigma_n}} \cancel{\sigma_n}(1-\sigma_n) \nabla_{\underline{w}} z^{(n)}\underline{w}^T\underline{x}^{(n)}$$

$$\frac{\partial \sigma(a)}{\partial a} = \sigma(a)(1-\sigma(a))$$

$$= -\sum_{n=1}^{N} (1-\sigma_n)\, z^{(n)}\, \underline{x}^{(n)}$$