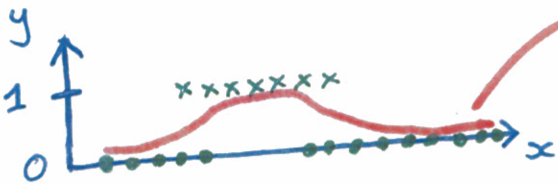
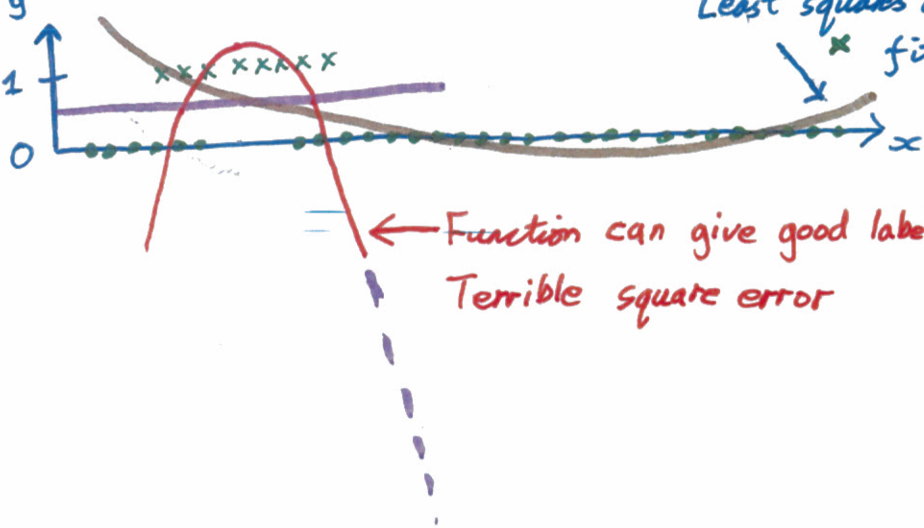
Regressing on Labels



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

(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 cost

Residuals $\underset{N\times 1}{\underline{r}} = \underline{y} - \underbrace{X\underline{w}}_{\underline{f}}$

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

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

"Gradient" vector of partial derivatives

$$\nabla_{\underline{w}}[\underline{r}^T\underline{r}] = \underline{0} - 2X^T\underline{y} + 2X^TX\underline{w}$$

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

Gradient Descent

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

"eta" small number 0.01?

(Scratch working)

$$\nabla_{\underline{w}} \left[\underline{w}^T \underline{h} \right] = \begin{bmatrix} \frac{\partial \underline{w}^T \underline{h}}{\partial w_1} \\ \frac{\partial \underline{w}^T \underline{h}}{\partial w_2} \\ \vdots \\ \frac{\partial \underline{w}^T \underline{h}}{\partial w_D} \end{bmatrix} = \underline{h}$$

some vector

$$\frac{\partial \underline{w}^T \underline{h}}{\partial w_i} = \frac{\partial}{\partial w_i} \sum_j w_j h_j$$

$$= \frac{\partial}{\partial w_i} \left(w_1 h_1 + w_2 h_2 + \dots \right.$$
$$\left. w_i h_i + \dots w_D h_D \right)$$

$$= h_i$$

Matrix Cookbook.

Normal Equations approach

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

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

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

If $(X^T X)^{-1}$ exists

Matlab:

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

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

w = X\y

BLAS

R Puzzle

$$
\begin{array}{ll}
 & \overbrace{\begin{array}{cccccc} & & & & & \text{bias} \\ 1 & 0 & \cancel{0} & & & 1 \\ 0 & 1 & \cancel{0} & & & 1 \\ 0 & 0 & \cancel{1} & & & 1 \\ x_1 & x_2 & x_3 & & & x_D \end{array}}^{X} \\
\text{"red"} \rightarrow & \\
\text{"blue"} \rightarrow & \\
\text{"green"} \rightarrow &
\end{array}
$$

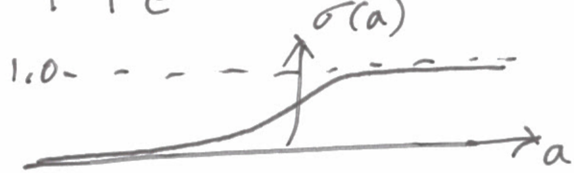
$$x_1 + x_2 + x_3 + x_D = 2$$

$$x_3 = 2 - x_1 - x_2 - x_D$$

Logistic Regression

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

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



Loss Function

Could use square loss again:

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

Normal interpretation:

$$p(y=1 \mid \underline{x}) \approx f(\underline{x};\underline{w})$$

Likelihood probability of data given $\underline{w}$

$$P(\{y^{(n)}\} \mid X, \underline{w}) = \prod_n p(y^{(n)} \mid \underline{x}^{(n)}; \underline{w})$$

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:

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

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

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

$$= -\sum_n \frac{1}{\sigma_n} \nabla_w \sigma_n \quad \text{(Chain rule)}$$

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

$$= \underline{\underline{-\sum_n (1-\sigma_n) z^{(n)} \underline{x}^{(n)}}}$$