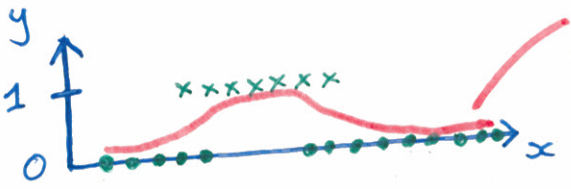
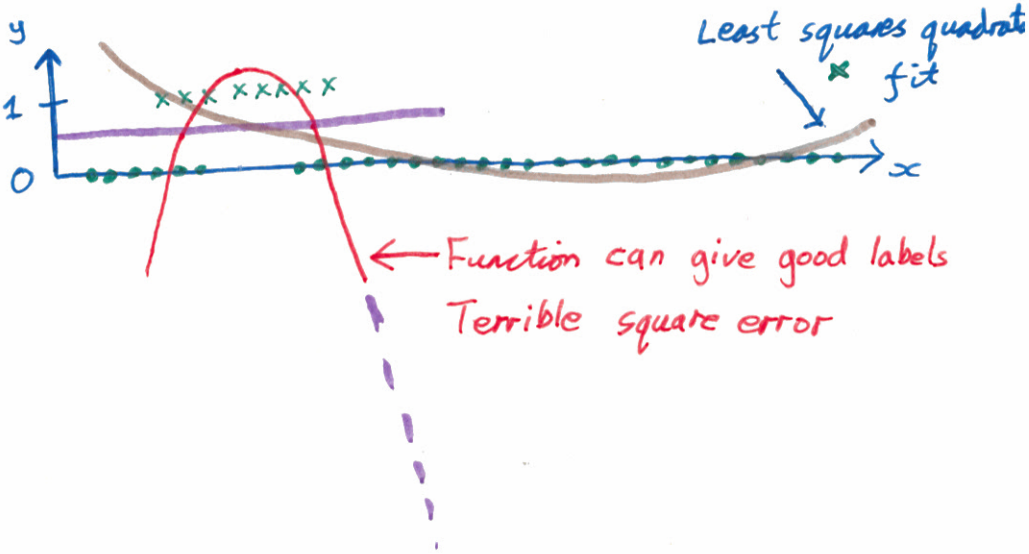
Regressing on Labels



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

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

Often bad idea:



Put up MLPR 2016 L8 (7) again

Gradients for least squares

X: $N \times D$ design matrix of all inputs

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

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

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

Gradient descent:

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

η: step-size, small.

$$\nabla_{\underline{w}} \; \underline{w}^T \underline{b} = \begin{bmatrix} \frac{\partial \underline{w}^T \underline{b}}{\partial w_1} \\ \frac{\partial \underline{w}^T \underline{b}}{\partial w_2} \\ \vdots \\ \vdots \end{bmatrix} = \underline{b}$$

$$\frac{\partial}{\partial w_i} \sum_j w_j b_j = \frac{\partial}{\partial w_i} \left[b_1 w_1 + b_2 w_2 + \ldots \right.$$
$$\left. b_i w_i + \ldots b_D w_D \right]$$
$$= b_i$$

$$\nabla_w [\underline{r}^T \underline{r}] = \underline{0}$$

$$(X^T X)\underline{w} = X^T y \quad \text{at optimal } \underline{w}$$

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

I

$X^{-1} X^{-T} X^T y$

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

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

Matlab:

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

$X^T X \quad D \times D$

X^TX DXD Matrix

If rank < D not invertible.

R puzzle

One hot encoding

Turns integer k → [00…0100…0]

k is in $\{1, 2, \ldots K\}$

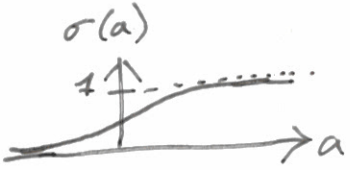
$k-1$ long

↑ kth position

Logistic Regression

$$f(\underline{x}; \underline{w}) = \sigma(\underline{w}^T \underline{x})$$

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



Loss Function

Could use square loss again:

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

Normally say

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

and use maximum likelihood. Maximize prob of data

Or minimize negative log probability.

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

What I do is make labels $\{-1, +1\}$

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

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

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

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

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

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

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

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