MLPR so far

Train, Val, Test → splits, generalization, Gaussian stats, CLT

Pre-processing: look at data, one hot encoding, taking logs

"Generative model"

Bayes classifiers + Gaussians

Today: revisit "Discriminative models" fit $P(y \mid \mathbf{x})$ directly

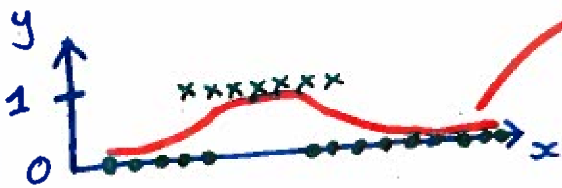
Linear regression

- with basis fn's
- regularization
- Bayesian prediction
- ∞ basis fn's: GPs

Start of non-linear models, fitted with gradient methods

Later will do Bayesian prediction with non-linear models

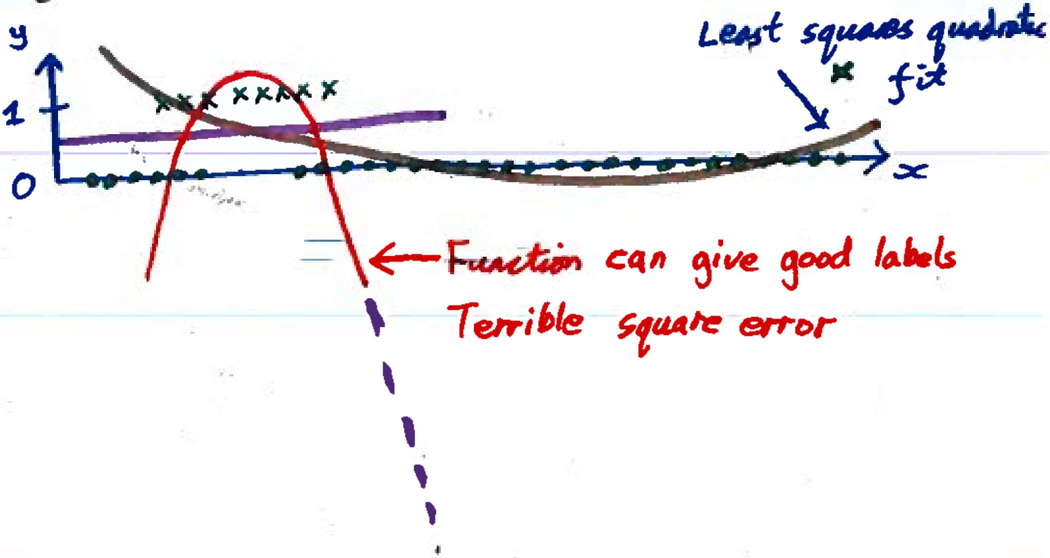
Regressing on Labels



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

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

Often bad idea:



Gradients for least squares cost

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

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

$$= \underline{y}^T\underline{y} - \underbrace{(X\underline{w})^T\underline{y} - \underline{y}^T(X\underline{w})}_{\text{Tut 2 Q1}}$$

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

"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

Initialise $\underline{w}$ (to $\underline{0}$?)

for $t = 1 \ldots T$:

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

"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

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

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

$$\dots w_i h_i + \dots$$

$$\left. \dots w_D h_D \right)$$

$$= h_i$$

Matrix Cookbook

Normal Equations approach

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

all weights are happy where they are

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

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

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

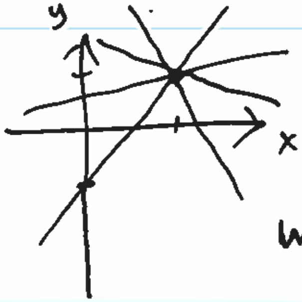
? $X^{-1} \cancel{(X^{-T} X^T)} \underline{y}$ II

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

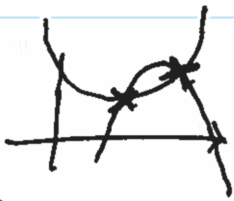
$N \times D$

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

When is there a unique solution?



$D = 2$
$N = 1$



Different functions same train cost

We need $N \geq D$ for unique solution.

One-hot encoding in R

		x_1	x_2	x_3	bias x_0
"red"	$\rightarrow$	1	0	$\cancel{0}$	1
"blue"	$\rightarrow$	0	1	$\cancel{0}$	1
"green"	$\rightarrow$	0	0	$\cancel{1}$	1

Imagine I have weights $\underline{w}$

New weights $\tilde{\underline{w}}$:

$$\tilde{w}_1 = w_1 + \delta$$
$$\tilde{w}_2 = w_2 + \delta$$
$$\tilde{w}_3 = w_3 + \delta$$
$$\tilde{w}_0 = w_0 - \delta$$

Same function $\Rightarrow \underline{w}^T\underline{x} = \tilde{\underline{w}}^T\underline{x}$ for all $\underline{x}$ (so same train cost)

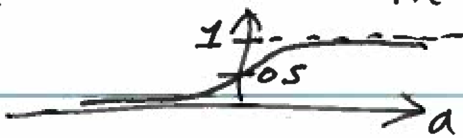
$$x_1 + x_2 + x_3 + x_0 = 2$$
$$x_3 = 2 - x_1 - x_2 - x_0$$

Logistic Regression

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

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

$$\sigma(a) = \frac{1}{1+e^{-a}}$$



Loss Function

Could use square loss again:

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

Interpretation:

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

Maximum Likelihood, maximize prob. of the data given $\underline{w}$:

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

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:

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

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

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

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

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

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

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