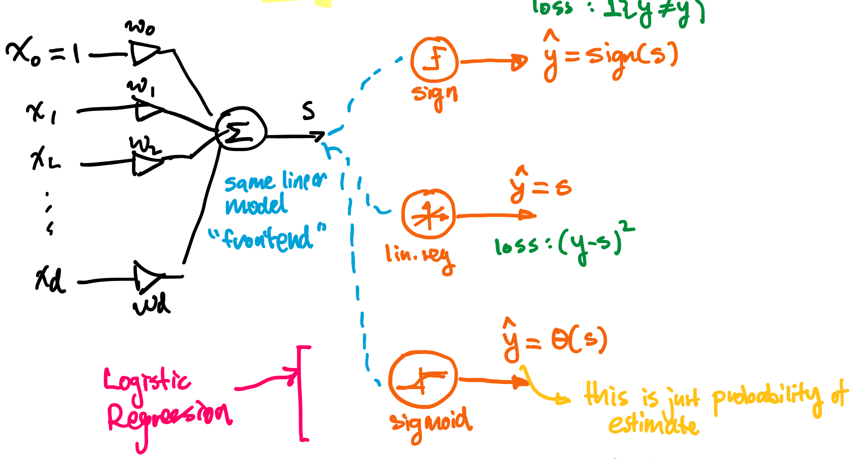
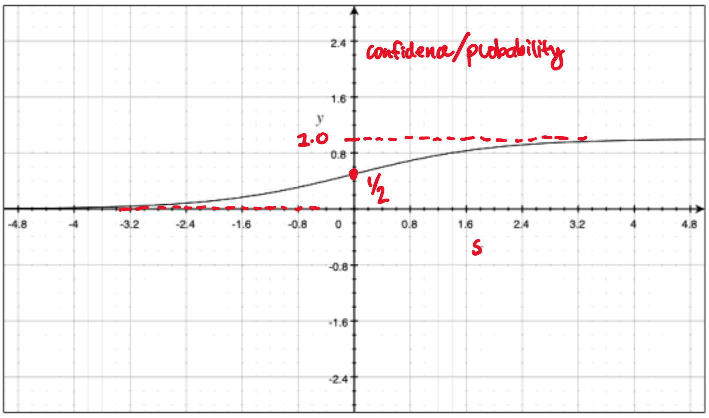
Logistic Regression

Tuesday, January 21, 2020 13:35

Still dealing w/ linear model $w^T x$



$$\theta = \frac{e^s}{1+e^s}$$



Interpretation:

$$\theta(s) = \mathbb{P}(y=1 \mid \underline{x})$$

or

probability estimate $\quad \hat{P}_{\underline{w}}(y \mid \underline{x}) = \text{Bernoulli}\left(y \mid \theta(\underline{w}^T \underline{x})\right)$

$$\hat{P}_{\underline{w}}(y=1 \mid \underline{x}) = \theta(\underline{w}^T \underline{x}) = \frac{e^{\underline{w}^T \underline{x}}}{1+e^{\underline{w}^T \underline{x}}}$$

$$\hat{P}_{\underline{w}}(y=-1 \mid \underline{x}) = 1-(\ldots) = \frac{e^{-\underline{w}^T \underline{x}}}{1+e^{-\underline{w}^T \underline{x}}}$$

compact $\Rightarrow \hat{P}_{\underline{w}}(y \mid \underline{x}) = \dfrac{e^{y\underline{w}^T \underline{x}}}{1+e^{y\underline{w}^T \underline{x}}}$

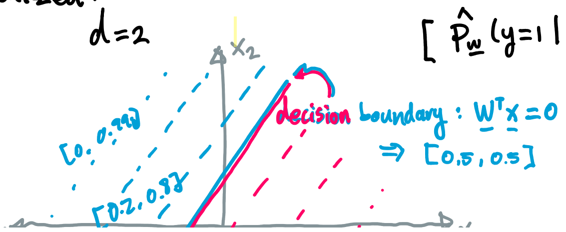
Log odds (Logit) function: $\log\left(\dfrac{\theta(s)}{1-\theta(s)}\right) = \underline{W}^T \underline{S}$

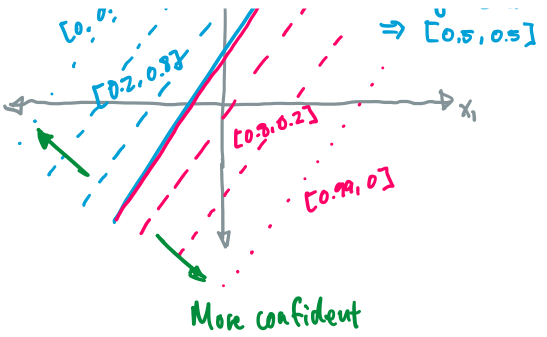
Visualized:

$d=2$

probability based on $\underline{w}$

$$\left[\hat{P}_{\underline{w}}(y=1 \mid \underline{x}),\ \hat{P}_{\underline{w}}(y=-1 \mid \underline{x})\right]$$





Loss Function: Measures uncertainty

$$\text{Loss} = \underbrace{-\log \hat{p}_{\underline{w}}(y|\underline{x})}_{\text{this is entropy : shows "uncertainty" or "surprise"}}$$

ex. $\hat{p} = 0.8$

$\hat{p}$	$y=1$ (true value)	$y=-1$
0.8	0.22	1.61
10^{-3}	10	10^{-4}

← we check the probability of the true 'y' value

?

Training Set $\mathcal{D} = \{(\underline{x}_1, y_1), (\underline{x}_2, y_2), \dots\}$

$$e_n(\underline{w}) = -\log\left(\hat{p}_{\underline{w}}(y_n|\underline{x}_n)\right)$$

Loss on $(\underline{x}_n, y_n) = \log\left(1 + e^{-y_n \underline{w}^T \underline{x}_n}\right)$
for each set of data

$$E_{in}(\underline{w}) = \frac{1}{N}\sum_{n=1}^{N} e_n(\underline{w})$$

To find $\underline{w}^*$: $\underline{w}^* = \underset{\underline{w} \in \mathbb{R}^{d+1}}{\arg\min}\ E_{in}(\underline{w})$

⇒ $E_{in}(\underline{w})$ is convex function of $\underline{w}$

⇒ With Regularization:

$E_{in}(\underline{w}) + \lambda \|\underline{w}\|^2$ is also convex for $\lambda \geq 0$

So this means there's an unique global minimum

⚠ but unlike linear regression, we don't have a closed form solution.

Maximum Likelihood Viewpoint

$$\hat{p}_{\underline{w}}(y_n|\underline{x}_n) = \Theta(y_n \underline{w}^T \underline{x}_n) = \frac{1}{1 + e^{y_n \underline{w}^T \underline{x}_n}}$$

Likelihood:

$$\hat{P}_{\underline{w}}(y_1,\dots y_N \mid \underline{x}_1 \dots x_N) = \underbrace{\prod_{n=1}^{N} \hat{P}_{\underline{w}}(y_n \mid x_n)}_{0.5 \times 0.6 \times 0.1 \times \dots}$$

we want to find $\underline{w}$ such that we maximizes likelyhood

AKA:

$$\Rightarrow \underline{w}^* = \underset{\underline{w} \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \prod_{n=1}^{N} \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n)$$

log of products $\equiv$ sum of logs
(easier to compute)

$$= \underset{\underline{w} \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \log\left(\prod_{n=1}^{N} \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n)\right)$$

$$= \underset{\underline{w} \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \sum_{n=1}^{N} \log\left(\hat{P}_{\underline{w}}(y_n \mid \underline{x}_n)\right)$$

$$= \underset{\underline{w} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \underbrace{\left(\frac{1}{N}\right) \sum_{n=1}^{N} \underbrace{-\log\left(\hat{P}_{\underline{w}}(y_n \mid x_n)\right)}_{e_n(\underline{w})}}_{E_{in}(\underline{w})}$$

Cross Entropy Viewpoint

discrete alphabets

$$\mathcal{S} = \{s_1, s_2 \dots s_M\}$$

probability dist.

$$\mathcal{P} = \{p_1, p_2 \dots p_M\}$$

this looks like entropy

$$H(P) = -\sum_{i=1}^{M} p_i (\log(p_i))$$

$$\text{Cross Entropy } CE(P,Q) = -\sum_{i=1}^{M} p_i \cdot \log(q_i)$$

$$= H(P) + \underbrace{D_{KL}(P \parallel Q)}_{\text{distance b/w } P \,\&\, Q}$$

▷ The log-loss function can be viewed as cross entropy

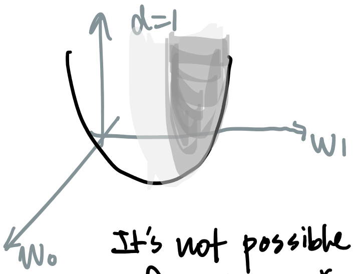
$$e_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n)$$

$$P(\{y_n=+1\}, \{y_n=-1\}) = -1\{y_n=+1\} \cdot \log \hat{P}_{\underline{w}}(+1 \mid \underline{x}_n) - 1\{y_n=-1\} \log \hat{P}_{\underline{w}}(-1 \mid \underline{x}_n)$$

$$P(\{y_n=+1\},\{y_n=-1\}) = -\underbrace{\mathbb{1}\{y_n=+1\}}_{p_i}\cdot\underbrace{\log \hat{p}_{\underline{w}}(+1|\underline{x}_n)}_{\log(q_i)} - \underbrace{\mathbb{1}\{y_n=-1\}}_{p_i}\underbrace{\log \hat{p}_{\underline{w}}(-1|\underline{x}_n)}_{\log(q_i)}$$

How do we minimize $E_{in}(\underline{w})$?

Similar to linear reg. $E_{in}(\underline{w})$ is concave:



It's not possible to solve for $\underline{w}^*$ for $E_{in}(\underline{w}^*)=\emptyset$

$\Rightarrow$ iterative solution using GRADIENT DESCENT