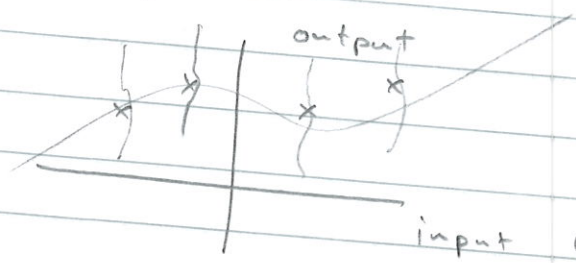
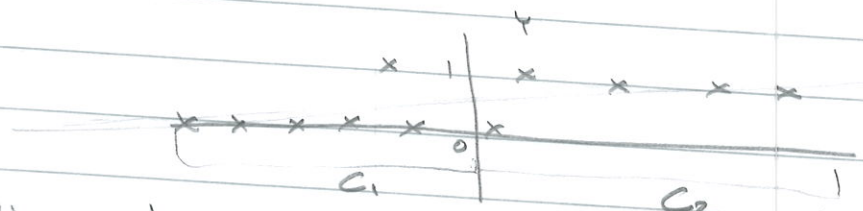
Classification ; In particular logistic reg

Classification and regression are close cousins. In probabilistic regression we model the conditional distribution of output given input (features).



In 2-class classification we do the same except the output is discrete rather than continuous, eg. $y_i = 1 \Rightarrow x_i \in C_1$. The goal is to learn a conditional distribution of the class label given the input.

Imagine, in 1-D input for now, that we have data like



it's clear that a linear regressor fit to such data will do something with a decision rule like $x_i \in C_1$ if $y_i < .5$ but such a rule a) doesn't allow a sharp decision boundary and b) doesn't have a coherent probabilistic interpretation: i.e, what does output = .7 mean?

What we might like is 1) a "noise" distribution that penalizes misclassification appropriately and 2) a controllable function that stretches input space.

The Bernoulli distribution is just the ticket for 1). For any given y_i we can write

$$p(y_i|\dots) = p_i^{y_i}(1-p_i)^{1-y_i}$$

which if we know $p(y_i=1) = p_i$ then "getting it right" has likelihood p_i and wrong $1-p_i$. If $p_i = 0.9$ and we observe $y_i = 1$ then we're better than if we observe $y_i = 0$. The only question is, what's p_i?

We only know that p_i must be between 0 and 1; the choice is arbitrary past that.

For now, assume $x_i \in \mathbb{R}$. Let's consider

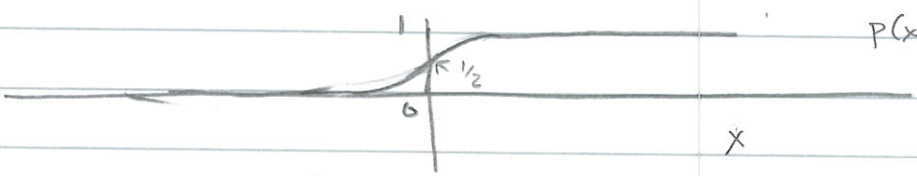
$$p_i = \frac{1}{1+\exp(-x_i\beta)}$$

the logistic sigmoid function. More generally

$$\sigma(x) = \frac{1}{1+\exp(-x)}$$

For large values of x, $\lim_{x\to\infty}\sigma(x) = 1$, and for small $\lim_{x\to-\infty}\sigma(x) = 0$.

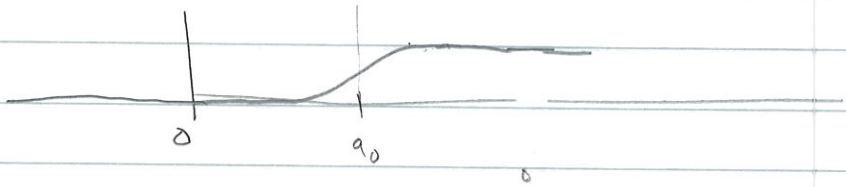
This function looks like



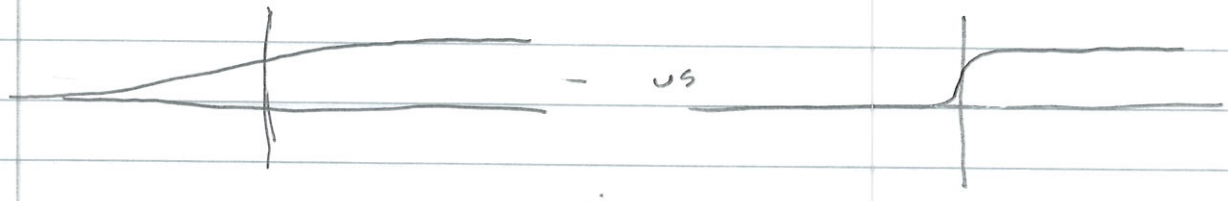
This is a non-linear function that stretches

One way shift this function by adding an offset

$$\sigma(x+a_0) = \frac{1}{1+\exp(-x-a_0)}$$



or make the slope more or less steep by adding a weight in front of x



via

$$\sigma(a_1 x + a_0) = \frac{1}{1+\exp(-a_1 x - a_0)}$$

The logistic sigmoid function has some nice properties which can be verified algebraically, for instance

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

and

$$\frac{d\sigma}{dx} = \sigma(1-\sigma) .$$

With these choices, (likelihood and nonlinear transform, we get logistic regression classification which imposes the following classification likelihood

$$p(\vec{y}=\{y_1 \cdots y_N\} \mid X=\{\vec{x}_1, \cdots \vec{x}_N\}; \vec{\beta}) = \prod_{n=1}^{N} \sigma(\vec{x}_n^T \vec{\beta})^{y_i} \left(1-\sigma(\vec{x}_n^T \vec{\beta})\right)^{1-y_i}$$

where $\vec{x}_i \in \mathbb{R}^{D+1}$, $\vec{\beta} \in \mathbb{R}^{D+1}$, $y_i = \{0,1\}$ and $\vec{x}_i = \begin{bmatrix} 1 \\ \vec{x}_i \end{bmatrix}$

like usual