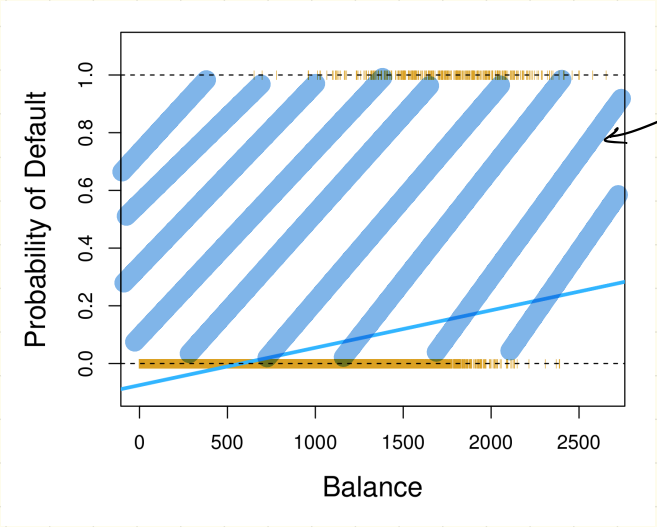
Topic 7:
Logistic regression

1 Motivation

Idea of solving binary classification problem:

1) Build a regression $f_{\hat{w}}(x)$
2) Say that $y(x) = \begin{cases} 0, & f_{\hat{w}}(x) < \frac{1}{2} \\ 1, & f_{\hat{w}}(x) \geq \frac{1}{2} \end{cases}$

Problems to solve:



- Move values of $f_{\hat{w}}$ into interval $[0,1]$
- Think about interpretation of continuous values into $\{0,1\}$.

understan $f_{\hat{w}}$ as probability

1.1 Transformation of $f_w(x)$.

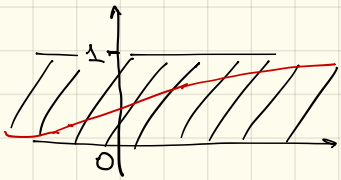
Assumption: we need to solve a binary classification problem "using" linear regression

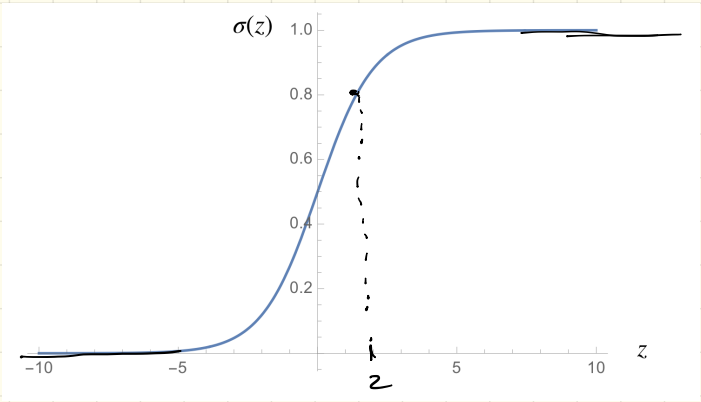
$f_w(x) = \langle x, w \rangle$ ($w \in \mathbb{R}^{d+1}$, weights)

input $\in \mathbb{R}^d$; x augmented $\in \mathbb{R}^{d+1}$

1. Change object/prediction function to $\sigma(f_w(x))$ such that

$\sigma : \mathbb{R} \to [0, 1]$





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

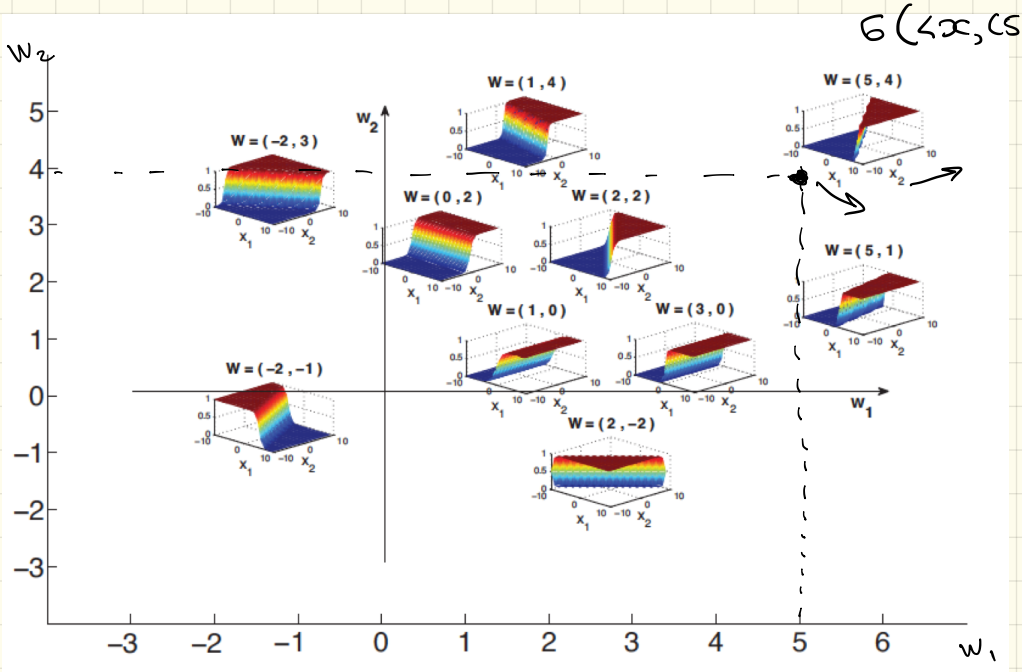
Using the above function we would define such probabilities:

$$p(1 \mid x, w) = \sigma(f_w(x)) = \sigma(\langle x, w \rangle)$$

(1: class label; x: given)

$$p(0 \mid x, w) = 1 - \sigma(f_w(x)) = 1 - \sigma(x, w)$$

$\sigma(\langle x, (5,4) \rangle)$



How to define $\hat{w}$?

1.2 Optimisation of weights.

Similar to classical regression problems we want to maximise the likelihood of getting y_i values for given $x^{(i)}$ and w over varying w.

$\{x^{(i)}, y_i\}_{i=1}^{s}$ – i.i.d.

$x^{(i)} \in \mathbb{R}^d$, $y_i \in \{0,1\}$

$$p(y \mid X, w) \overset{\text{independence}}{=} \prod_{i=1}^{s} p(y_i \mid x^{(i)}, w)$$

split into 2 groups: $y_i = 0$, $y_i = 1$

$$= \prod_{i:\, y_i=0} \underbrace{p(y_i \mid x^{(i)}, w)}_{1-\sigma} \cdot \prod_{i:\, y_i=1} \underbrace{p(y_i \mid x^{(i)}, w)}_{\sigma}$$

Trick: $a^{d} b^{1-d} = \begin{cases} a, & d = 1 \\ b, & d = 0 \end{cases}$

$$p(y \mid X, w) = \prod_{i=1}^{s} p(y_i \mid x^{(i)}, w)$$ ← independence

split into 2 groups:
$\begin{cases} y_i = 0 \\ y_i = 1 \end{cases}$

$$= \prod_{i:\, y_i=0} p(y_i \mid x^{(i)}, w) \cdot \prod_{i:\, y_i=1} p(y_i \mid x^{(i)}, w)$$

$1-\sigma$ $\quad$ σ

Trick: $a^{\alpha} b^{1-\alpha} = \begin{cases} a, & \alpha = 1 \\ b, & \alpha = 0 \end{cases}$

$\alpha \to \sigma$, $1-\alpha \to 1-\sigma$

$$= \prod_{i=1}^{s} \sigma^{y_i}(\langle x^{(i)}, w \rangle) \cdot (1 - \sigma(\langle x^{(i)}, w \rangle))^{1-y_i}$$

$$L(w) = -\log p(y \mid X, w)$$

$\sigma_i = \sigma(\langle x^{(i)}, w \rangle)$

$e_i = e^{\langle x^{(i)}, w \rangle}$

$$= -\log \prod_{i=1}^{s} \sigma_i^{y_i} (1-\sigma_i)^{1-y_i}$$

$$= -\sum_{i=1}^{s} \log \sigma_i^{y_i} (1-\sigma_i)^{1-y_i}$$

$$= -\sum_{i=1}^{s} y_i \log \sigma_i + (1-y_i) \log (1-\sigma_i)$$

$$= -\sum_{i=1}^{s} y_i \log \frac{e_i}{1+e_i} + (1-y_i) \log \left(1 - \frac{e_i}{1+e_i}\right)$$

$$= -\sum_{i=1}^{s} y_i \log e_i - y_i \log(1+e_i) - (1-y_i) \log(1+e_i)$$

$$= -\sum_{i=1}^{s} y_i \langle x^{(i)}, w \rangle - \log\left(1 + e^{\langle x^{(i)}, w \rangle}\right)$$

$$L(w) = \sum_{i=1}^{s} \log\left(1 + e^{\langle x^{(i)}, w \rangle}\right) - y_i \langle x^{(i)}, w \rangle$$

$$\hat{w} = \arg\min_{w} L(w)$$

Questions to answer:

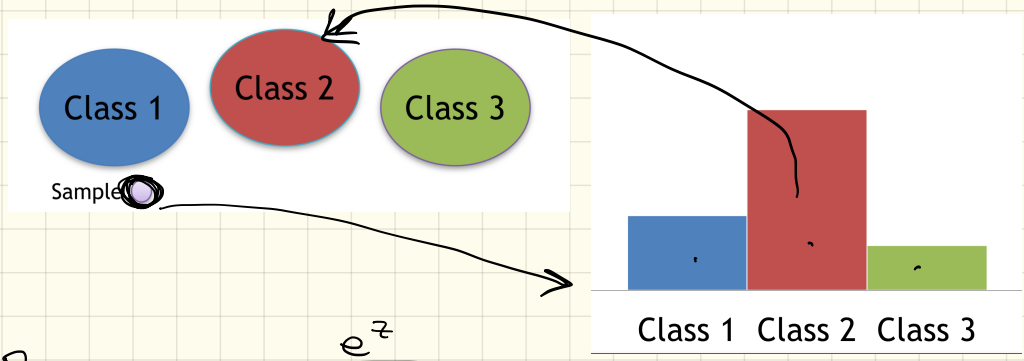
1. How to extend the method to multinomial regression?
2. How to solve corresponding optimisation problems?

2. Multinomial logistic regression

If one wants to classify inputs into n classes then one would be interested in defining $p(j \mid x, w)$ for $j = \overline{1, n}$.

Restrictions:

- $p(j \mid x, w) \in [0, 1]$
- $\sum_{j=1}^{n} p(j \mid x, w) = 1$



Binary: $z \to \frac{e^z}{1+e^z}$

Multinomial case: $(z_1, z_2, \ldots, z_n) \to (p_1, \ldots, p_n)$

softmax

Def: The softmax function

$\sigma : \mathbb{R}^n \to \mathbb{R}^n$ is

$$\text{softmax} = \sigma(z = (z_1, \dots, z_n))$$

$$= \left(\frac{e^{z_1}}{\sum_{j=1}^{n} e^{z_j}} \quad \frac{e^{z_2}}{\sum_{j=1}^{n} e^{z_j}} \quad \cdots \quad \frac{e^{z_n}}{\sum_{j=1}^{n} e^{z_j}} \right)$$

Remark: binary multinomial

$$\frac{e^z}{1+e^z} = \sigma(z) \longleftrightarrow \text{“}\sigma((0, z))\text{”}$$

$$\bar{x} \rightsquigarrow \langle x, w \rangle = z \xrightarrow{\text{plug}} (0, z)$$

$e^0 \to \left(\frac{1}{1+e^z} \quad \frac{e^z}{1+e^z} \right)$

Origin of softmax name?

$$\text{argmax}(x_1, x_2, \dots, x_n) = p$$

$$= (0 \; 0 \dots 0 \; \underbrace{1}_{p} \; 0 \dots 0)$$

$$\text{softmax}(x_1, x_2, \dots, x_n) =$$

$$= \left(\frac{e^{x_1}}{\sum_{j=1}^{n} e^{x_j}}, \dots, \frac{e^{x_n}}{\sum_{j=1}^{n} e^{x_j}} \right)$$

largest $x_n \to 1$, $x_p \to \infty$

p

2.2 Setting up a multinomial logistic regression problem

How to get n dimensional vector from d-dimensional input?

$$f_W(\bar{x}) = \left(\langle x, w^{(1)} \rangle \quad \langle x, w^{(2)} \rangle \quad \dots \langle x, w^{(n)} \rangle \right)$$

$$\sigma(\bar{x}, W) = \left(\frac{e^{\langle x, w^{(1)} \rangle}}{\sum_{j=1}^{k} e^{\langle x, w^{(j)} \rangle}} \quad \dots \quad \frac{e^{\langle x, w^{(n)} \rangle}}{\sum_{j=1}^{k} e^{\langle x, w^{(j)} \rangle}} \right)$$

$$p(y_i = k \mid x^{(i)}, W) = (\sigma(x^{(i)}, W))_k$$

$$\hat{W} = \arg\min_W \sum_{i=1}^{s} \log \left(\sum_{j=1}^{k} e^{\langle x^{(i)}, w^{(j)} \rangle} \right) - \sum_{i=1}^{s} \sum_{j=1}^{k} \mathbb{1}_{y_i = j} \langle x^{(i)}, w^{(k)} \rangle$$

2.3 Solving logistic regression.

Binary classification:

$$\hat{w} = \arg\min_w \left\{ \sum_{i=1}^{n} \log\left(1 + e^{\langle x^{(i)}, w \rangle}\right) - \sum_{i=1}^{n} y_i \langle x^{(i)}, w \rangle \right\}$$

$w^{(0)} = (0, \dots, 0)$
$w^{(j)} = w$

Multinomial classification

$$\hat{W} = \arg\min_W \left\{ \sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} e^{\langle x^{(i)}, w^{(j)} \rangle}\right) - \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{1}_{y_i = j} \langle x^{(i)}, w^{(j)} \rangle \right\}$$

2 classes $\{0, 1\}$

$$- \sum_{i=1}^{n} \mathbb{1}_{y_i = 0} \cdot \langle x^{(i)}, \overline{0} \rangle + \mathbb{1}_{y_i = 1} \cdot \langle x^{(i)}, w \rangle$$

$j = 0$ $\quad$ $j = 1$

$y_i \in \{0,1\} \quad \mathbb{1}_{y_i = 1} \in \{0,1\}$

Gradient descent ?

2.3.1 Gradient descent for logistic regression.

$$L_2(w) = \sum_{i=1}^{s} \left\{ \log\left(1 + e^{\langle x^{(i)}, w \rangle}\right) - y_i \langle x^{(i)}, w \rangle \right\}$$

$$L_n(W) = \sum_{i=1}^{s} \left\{ \log\left(\sum_{j=1}^{k} e^{\langle x^{(i)}, w^{(j)} \rangle} \right) - \sum_{j=1}^{k} \mathbb{1}_{y_i = j} \langle x^{(i)}, w^{(j)} \rangle \right\}$$

$$W^{(k+1)} = W^{(k)} - \tau \nabla L_m(W^{(k)}).$$

↑ vector if $m = 2$
matrix if $m > 2$.

Binary-

$\nabla L_2(w)$ - ? $\Rightarrow$ need to find $\frac{\partial L_2(w)}{\partial w_p}$

- What is the derivative $\frac{\partial}{\partial w_p} f(\langle x^{(i)}, w \rangle)$

$$\frac{\partial}{\partial w_p} f\left(x_0^{(i)} w_0 + x_1^{(i)} w_1 + \dots + x_d^{(i)} w_d \right)$$

only one has w_p

$$= f'\left(\langle x^{(i)}, w \rangle\right) \cdot x_p^{(i)}$$

$$\frac{\partial}{\partial w_p} L_2(w) = \sum_{i=1}^{s} \left\{ \frac{\partial}{\partial w_p} \log\left(1 + e^{\langle x^{(i)}, w \rangle}\right) - \frac{\partial}{\partial w_p} y_i \cdot \langle x^{(i)}, w \rangle \right\}$$

$f(z) = \log(1+e^z)$, $f'(z) = \sigma(z)$

$f(z) = y_i \cdot z$, $f'(z) = y_i$

$$= \sum_{i=1}^{s} \left\{ \sigma\left(\underbrace{\langle x^{(i)}, w \rangle}_{(Xw)_i} \right) \cdot x_p^{(i)} - y_i \cdot x_p^{(i)} \right\}$$

$= X_{ip}$

$$X = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_d^{(1)} \\ \vdots & & & & \\ 1 & x_1^{(s)} & x_2^{(s)} & \dots & x_d^{(s)} \end{pmatrix}$$

$$\frac{\partial}{\partial w_p} L_2(w) = \sum_{i=1}^{s} \left\{ \frac{\partial}{\partial w_p} \log\left(1 + e^{\langle x^{(i)}, w \rangle}\right) - \frac{\partial}{\partial w_p} y_i \cdot \langle x^{(i)}, w \rangle \right\}$$

$f(z) = \log(1+e^z)$, $f'(z) = \sigma(z)$

$f(z) = y_i \cdot z$, $f'(z) = y_i$

$$= \sum_{i=1}^{s} \left\{ \sigma\left(\underbrace{\langle x^{(i)}, w \rangle}_{(Xw)_i}\right) \cdot x_p^{(i)} - y_i \cdot x_p^{(i)} \right\}$$

$= X_{ip}$

$$X = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_d^{(1)} \\ \vdots & & & & \\ 1 & x_1^{(s)} & x_2^{(s)} & \dots & x_d^{(s)} \end{pmatrix}$$

$$Xw = \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix} \begin{pmatrix} | \\ w \\ | \end{pmatrix} = \begin{pmatrix} \langle x^{(1)}, w \rangle \\ \langle x^{(2)}, w \rangle \end{pmatrix} \Rightarrow \langle x^{(i)}, w \rangle = (Xw)_i$$

X $\quad$ w

$$= \sum_{i=1}^{s} X_{ip} \left(\sigma\left((Xw)_i\right) - y_i \right) = \left[X^T \left(\sigma(Xw) - y \right) \right]_p$$

$X_{ip} = X_{pi}^T$

$\sigma(Xw)$ - ? coordinate-wise application of σ.

∇L_2