Warm-up as You Walk In

Bernouli distribution:

$$Y \sim Bern(z)$$

$$p(y) = \begin{cases} z, & y = 1 \\ 1 - z, & y = 0 \end{cases}$$

What is the log likelihood for three i.i.d. samples, given parameter z:

$$\mathcal{D} = \{y^{(1)} = 1, y^{(2)} = 1, y^{(3)} = 0\}$$

$$L(z) =$$

$$\ell(z) =$$

Introduction to Machine Learning

Logistic Regression

Instructor: Pat Virtue

Announcements

Assignments:

- HW2 (written & programming)
  - Due Tue 2/4, 11:59 pm

Early Feedback

- More mathematical rigor
- Consolidated course notes
- Lots of concepts, how does it all fit together?

Plan

Last time

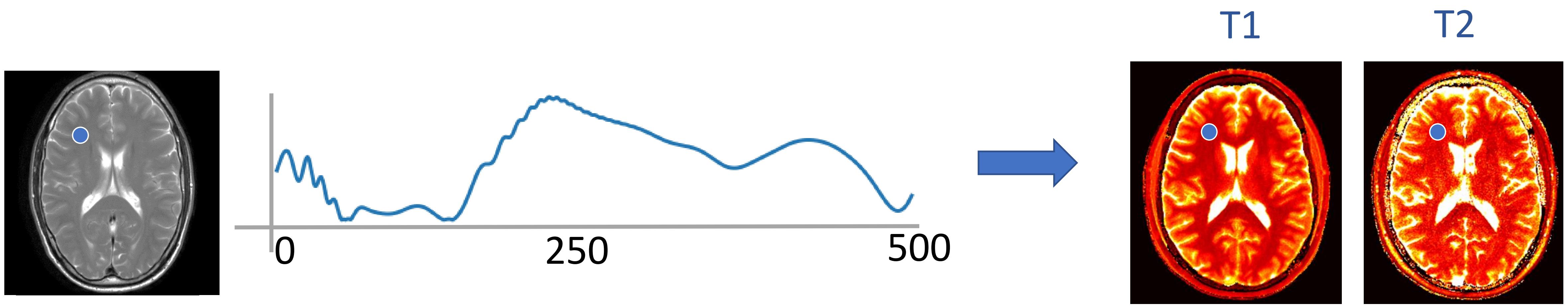
- Likelihood
- Density Estimation
- MLE for Density Estimation

Today

- Wrap up MLE for linear regression
- Classification models
- MLE for logistic regression

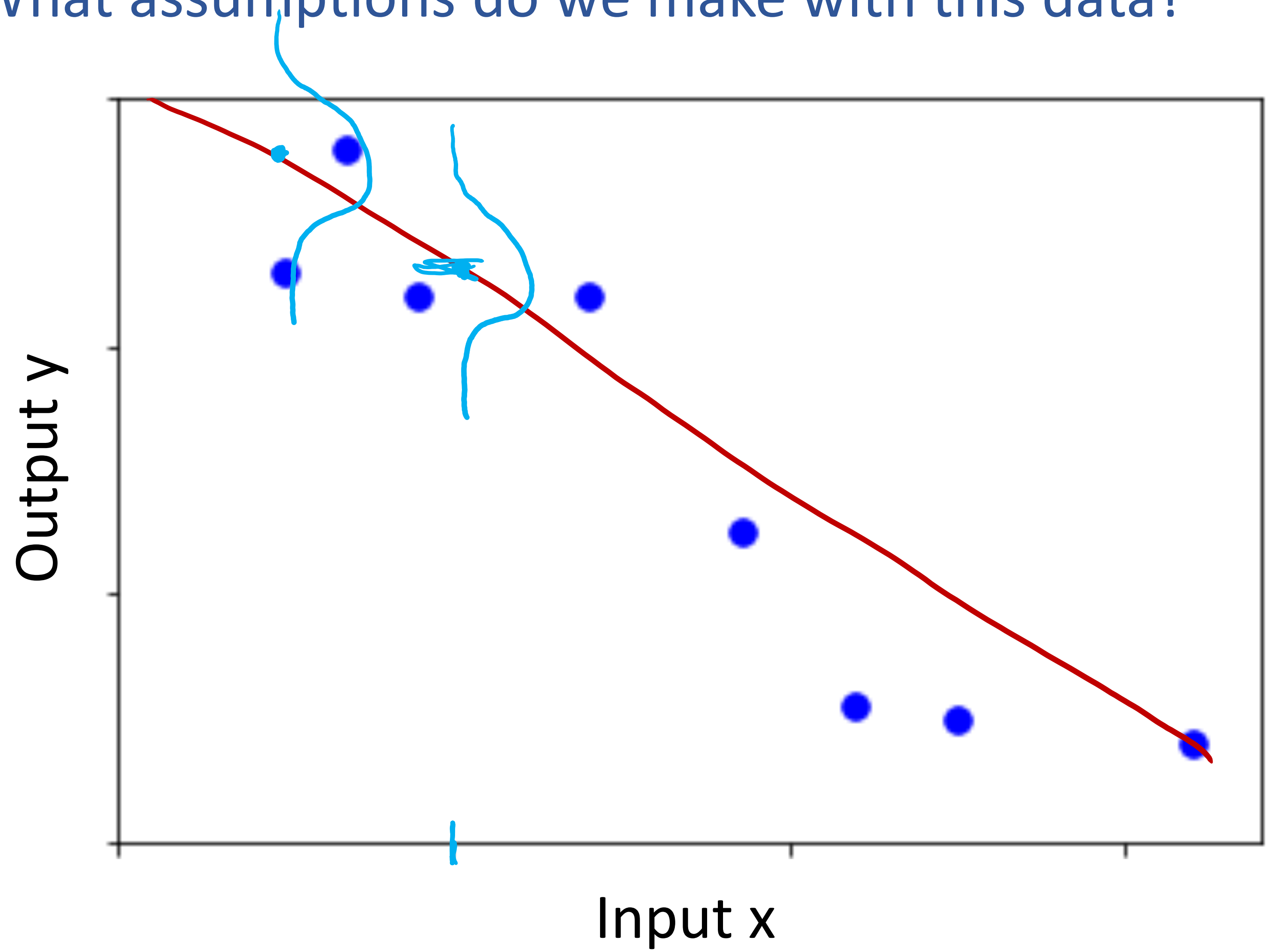
MR Fingerprinting Assumptions

Forgot a really important assumption!!



Assumptions

What assumptions do we make with this data?



$$y = w^T x + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

$$Y \sim N(w^T x, \sigma^2)$$

Modelling $f(Y|X,\theta)$ ←

Density Estimation

$f(D|\underset{\sim}{\theta})$

$X \sim \mathcal{N}(\mu, \sigma^2)$

$f(Y|X,\theta)$

$f(\underline{x}|\underline{\mu, \sigma^2})$

Conditional likelihood

$f(\underline{D}|\underline{\theta})$

$$f(y^{(n)}|x^{(n)}, \vec{w}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(\frac{-(y^{(n)} - w^T x^{(n)})^2}{2\sigma^2}\right)}$$

$$f(\vec{y}|\vec{x}, \vec{w}, \sigma^2) = \prod_{n}^{N} f(y^{(n)}|x^{(n)}, \vec{w}, \sigma^2) \leftarrow$$

MLE for Linear Regression

How does our model of $f(Y|X,\theta)$ with the likelihood function?

$L(\theta)$

Maximum (Conditional) Likelihood Estimate

M(C)LE for Linear Regression

$$e^{z_1} e^{z_2} = e^{z_1+z_2}$$

$$L(\boldsymbol{w}, \boldsymbol{\sigma^2}) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{\left(\frac{-\sum_N (y^{(n)} - \boldsymbol{w}^T \boldsymbol{x}^{(n)})^2}{2\sigma^2}\right)}$$

$$\ell(w, \sigma^2) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\underline{\sum (y^{(n)} - w^T x^{(n)})^2}$$

$$\frac{d\ell}{dw} = 0$$

$$\hat{w}_{ML} = ?$$

$$\ell(\mu) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{\sum_{n=1}^{N} (x^{(n)} - \mu)^2}{2\sigma^2}$$

M(C)LE for Linear Regression

How does M(C)LE optimization relate to least squares optimization?

$$\ell(w, \sigma^2) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\underline{\underline{\sum \left(y^{(n)} - w^T x^{(n)}\right)^2}}$$

$$J(\boldsymbol{w}) = \frac{1}{N2}\left\| \vec{y} - Xw \right\|_2^2$$

Piazza Poll 2:

Does $\min_{\boldsymbol{w}} -\ell(\boldsymbol{w})$ equal $\min_{\boldsymbol{w}} J(\vec{w})$?

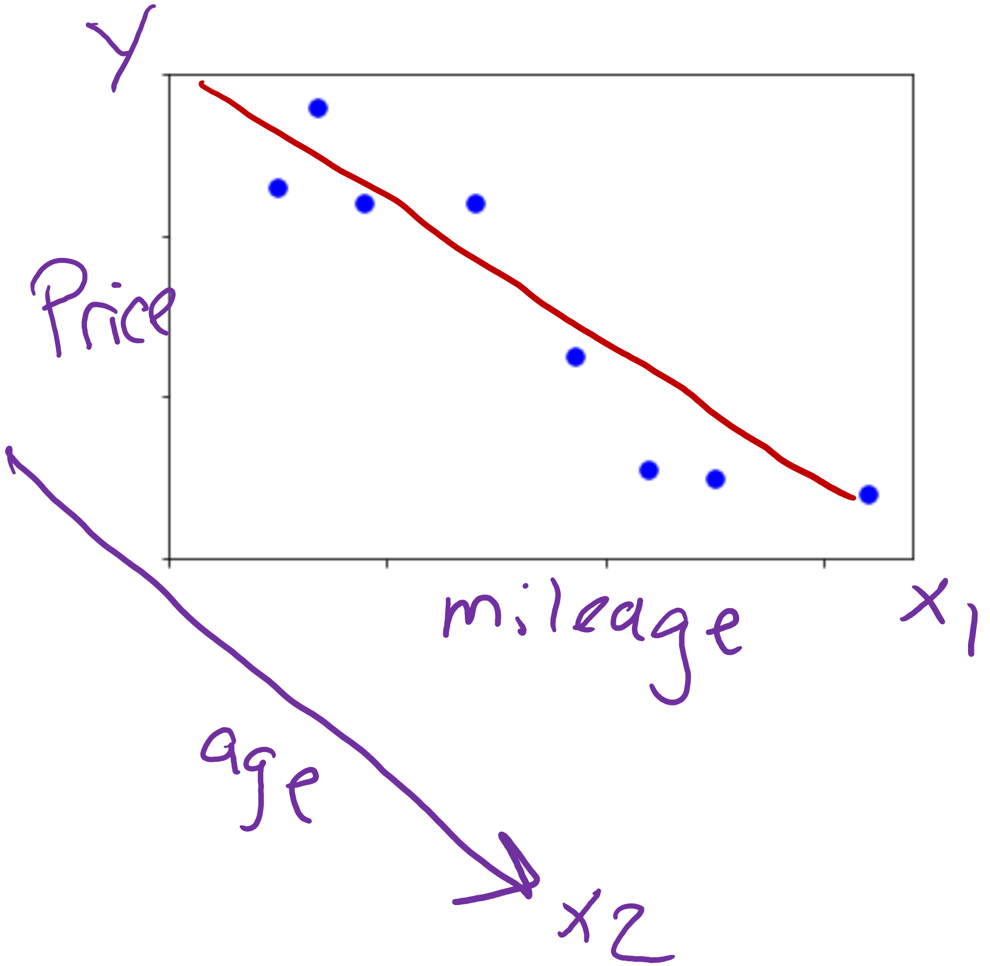
$$\ell(\vec{w}, \sigma^2) = \cancel{-\frac{N}{2}\log(2\pi)} - \cancel{\frac{N}{2}\log(\sigma^2)} - \underbrace{\frac{1}{2\sigma^2} \underline{\underline{\sum\left(y^{(n)} - \vec{w}^T \vec{x}^{(n)}\right)^2}}}$$

$$J(\vec{w}) \underbrace{\frac{1}{2N}\|\vec{y} - X\vec{w}\|_2^2}$$

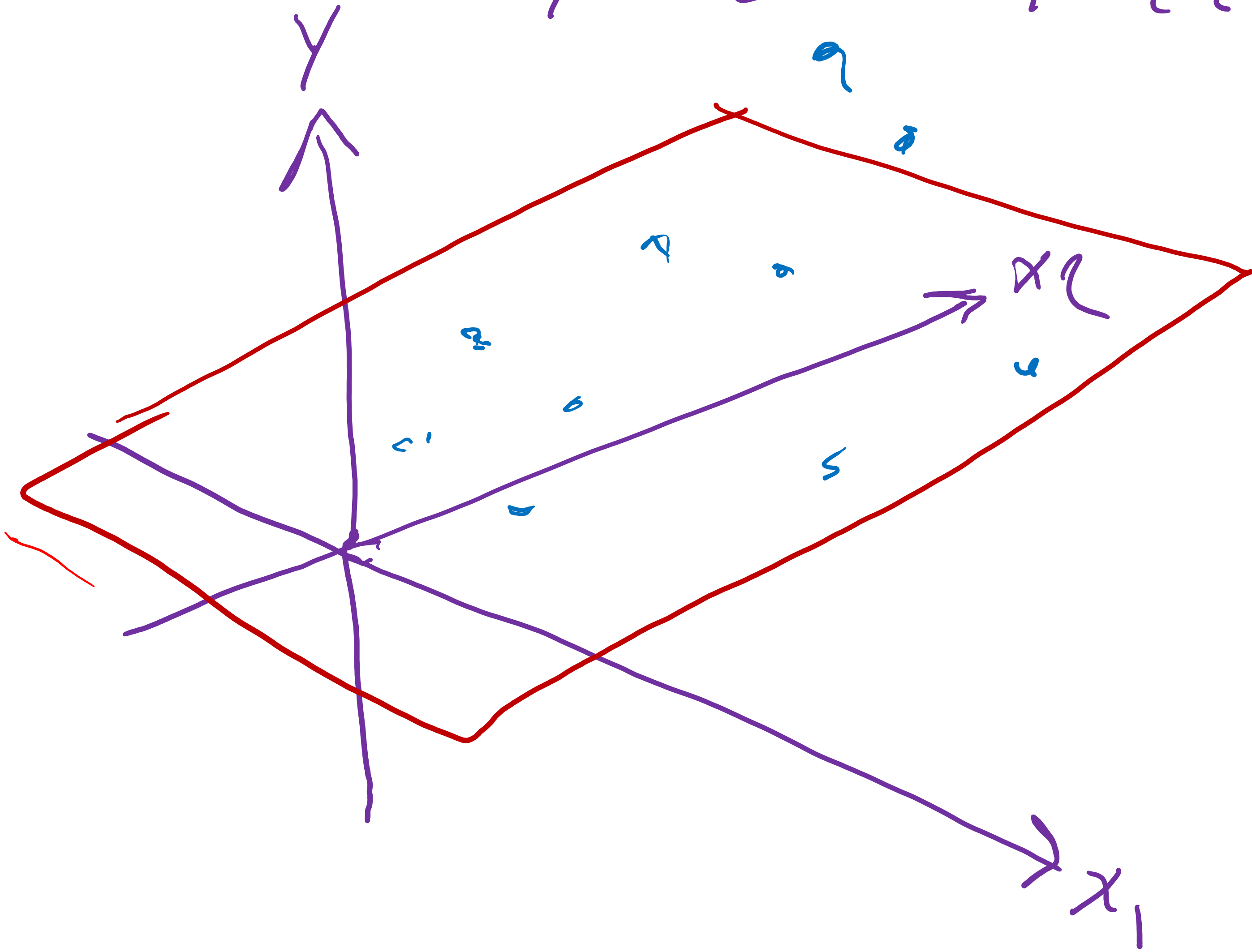
$$2\sum\left(y^{(n)} - w^T x^{(n)}\right) x^{(n)}$$

Linear Regression with Multiple Input Features

$y = w_0 + w_1 x$



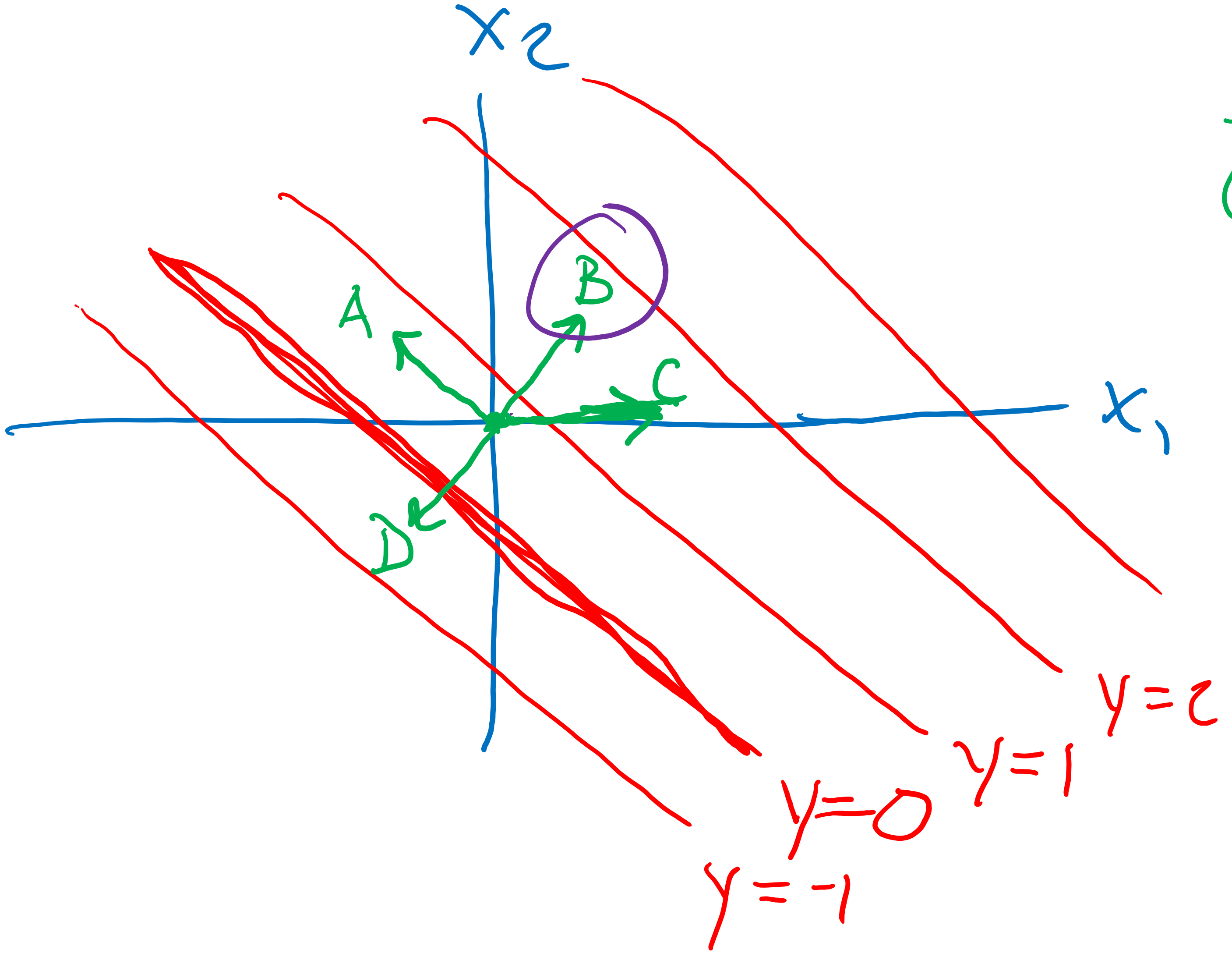
$y = w_0 + w_1 x_1 + w_2 x_2$



Poll 1: Which vector is the correct $\boldsymbol{\theta}$?

$\theta = [w_1 \; w_2]^T$

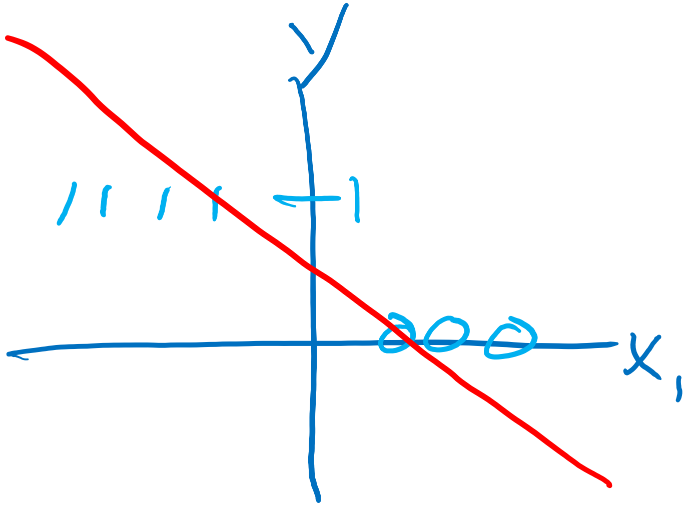
$\vec{\theta}^T \vec{x}$

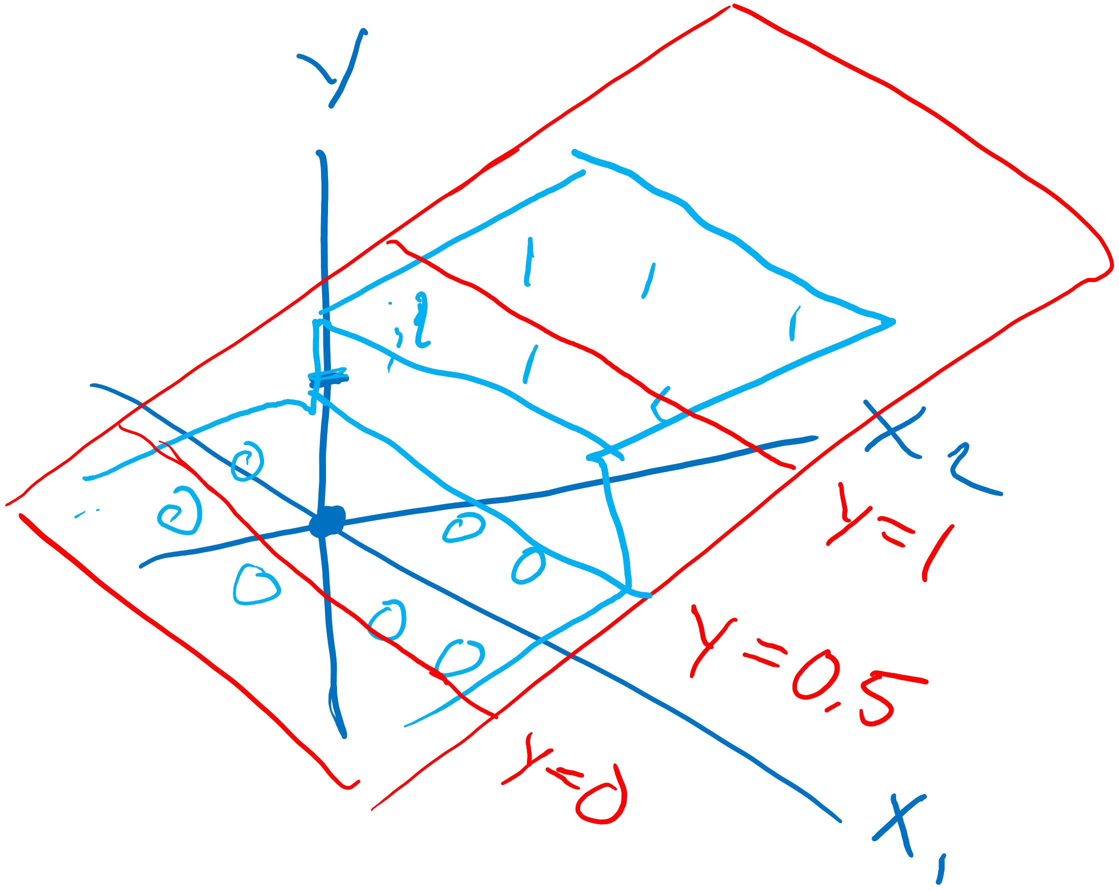


E ?

Classification Models

Linear Regression

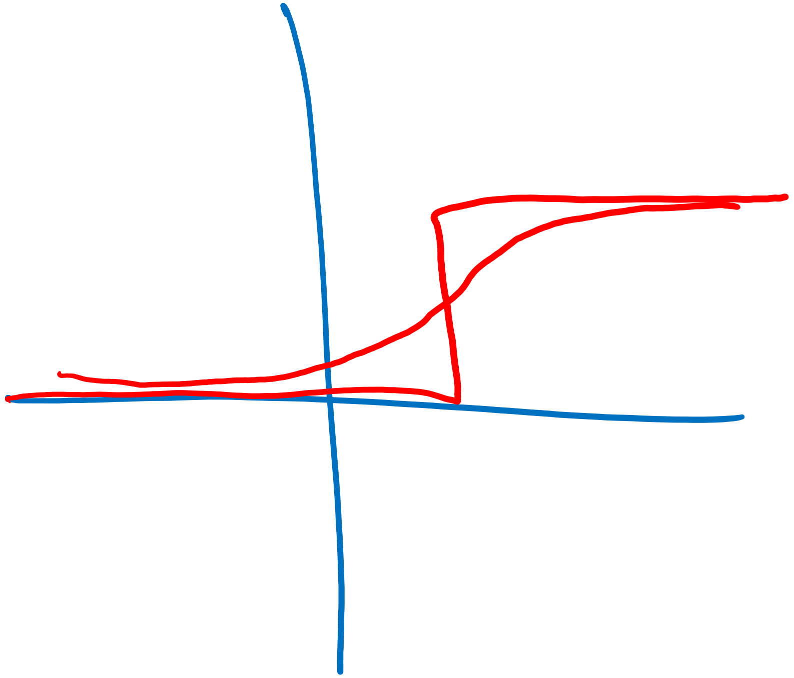




$$y = w_0 + x^T w$$

Classification Models

Linear Regression with Decision Boundary



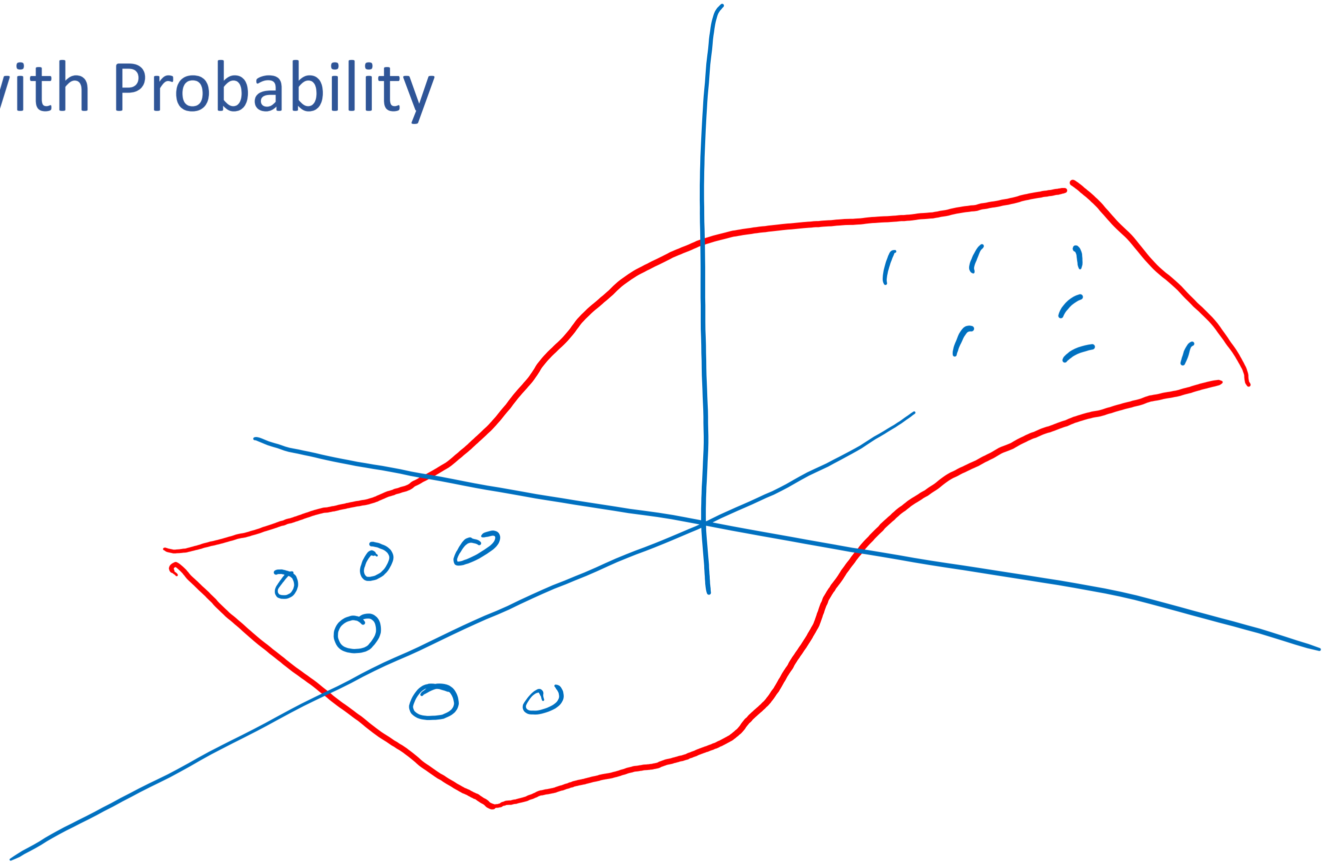
$$z = w^T x$$

$$y = \begin{cases} 1 & z \geq 0.5 \\ 0 & z < 0.5 \end{cases}$$

$$y = step(z, 0.5)$$

Classification Models

Linear Regression with Probability

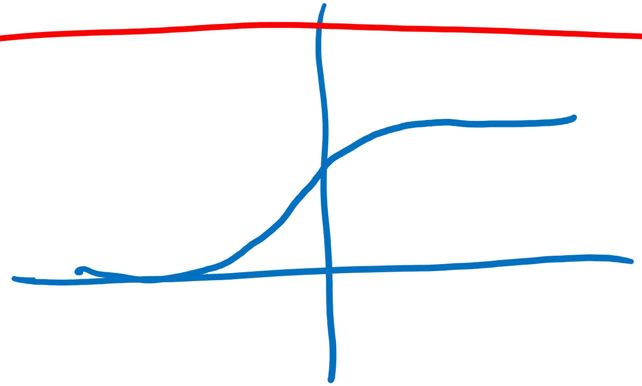


Modelling $p(Y|X,\theta)$

Bernoulli distribution of logistic function of linear model

$$z = x^T w$$

$$g = g(x^T w) \qquad g(z) = \frac{1}{1+e^{-z}}$$



$$y \sim \mathrm{Bern}(g) = \begin{cases} g & y=1 \\ 1-g & y=0 \end{cases}$$

MLE for Bernoulli

Bernoulli distribution:

$$Y \sim Bern(z)$$

$$p(y) = \begin{cases} z, & y = 1 \\ 1 - z, & y = 0 \end{cases}$$

What is the log likelihood for three i.i.d. samples, given parameter z?

$$\mathcal{D} = \{y^{(1)} = 1, y^{(2)} = 1, y^{(3)} = 0\}$$

$$L(z) = z \cdot z \cdot (1-z) \qquad = \prod_n z^{y^{(n)}} (1-z^{(n)})^{1-y^{(n)}}$$

$$\ell(z) = \log z + \log z + \log(1-z) = \sum_n y^{(n)} z^{(n)} + (1-y)(1-z)$$

MLE for Bernoulli

$y^{(3)^2}$

Bernoulli distribution:

$$Y \sim Bern(z)$$

$$p(y) = \begin{cases} z, & y = 1 \\ 1 - z, & y = 0 \end{cases}$$

What is the log likelihood for three i.i.d. samples, given parameter z?

$$\mathcal{D} = \{y^{(1)} = 1, y^{(2)} = 1, y^{(3)} = 0\}$$

$$L(z) = z \cdot z \cdot (1-z) = \prod_n z^{y^{(n)}} (1-z)^{1-y^{(n)}}$$

$$\ell(z) = \log z + \log z + \log(1-z) = \sum_n y^{(n)} \log z + (1-y^{(n)}) \log(1-z)$$

MLE for Bernoulli

Bernoulli distribution:

$$Y \sim Bern(z)$$

$$p(y) = \begin{cases} z, & y = 1 \\ 1 - z, & y = 0 \end{cases}$$

What is the log likelihood for three i.i.d. samples, given parameter z?

$$\mathcal{D} = \{y^{(1)} = 1, y^{(2)} = 1, y^{(3)} = 0\}$$ Y=1 Y=0

$$L(z) = z \cdot z \cdot (1 - z) \qquad = \prod_n z^{y^{(n)}} (1 - z)^{(1 - y^{(n)})}$$

$$\ell(z) = \log z + \log z + \log(1 - z) \qquad = \sum_n y^{(n)} \log z + \left(1 - y^{(n)}\right) \log(1 - z)$$

M(C)LE for Logistic Regression

$p(Y \mid X, \theta)$

$p(Y \mid X, \underline{\mathbf{w}}) = \underline{\prod_{n=1}^{N} p(y^{(n)} \mid \boldsymbol{x}^{(n)}, \mathbf{w})}$

$p(D \mid \theta)$

$p(y^{(1)}, y^{(2)}, y^{(3)} \ldots \mid x, w)$

Model Y as a Bernoulli distribution, but the temporary z is now based on the logistic function of our linear model of input $\boldsymbol{x}$

$$Y \sim Bern(\mu), \qquad \mu = \underline{g(\mathbf{w}^T \boldsymbol{x})}, \qquad \underline{g(z) = \frac{1}{1+e^{-z}}}$$

if $y^{(n)} = 1$

if $y^{(n)} = 0$

What is the *conditional* log likelihood?

$L(\mathbf{w}) = \prod_n p(Y = y^{(n)} \mid X = \mathbf{x}^{(n)}, w) = \prod_n \mu^{(n)\, y^{(n)}} (1-\mu^{(n)})^{1-y^{(n)}}$

$\ell(\mathbf{w}) =$

M(C)LE for Logistic Regression

$p(Y \mid X, \theta)$

$p(Y \mid X, \boldsymbol{w}) = \prod_{n=1}^{N} p(y^{(n)} \mid \boldsymbol{x}^{(n)}, \boldsymbol{w})$

Model Y as a Bernoulli distribution, but the temporary z is now based on the logistic function of our linear model of input $\boldsymbol{x}$

$$Y \sim Bern(\mu), \qquad \mu = g(\boldsymbol{w}^T \boldsymbol{x}), \qquad g(z) = \frac{1}{1+e^{-z}}$$

What is the *conditional* log likelihood?

$$L(\boldsymbol{w}) = \prod_n g\left(\boldsymbol{w}^T \boldsymbol{x}^{(n)}\right)^{y^{(n)}} \left(1 - g\left(\boldsymbol{w}^T \boldsymbol{x}^{(n)}\right)\right)^{\left(1-y^{(n)}\right)}$$

$$\ell(\boldsymbol{w}) = \sum_n \left(y^{(n)} \log g\left(\boldsymbol{w}^T \boldsymbol{x}^{(n)}\right) + \left(1 - y^{(n)}\right) \log\left(1 - g\left(\boldsymbol{w}^T \boldsymbol{x}^{(n)}\right)\right)\right)$$

M(C)LE for Logistic Regression

$z = f(\boldsymbol{w}, \boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$

$\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x}) = \boldsymbol{x}$

$\mu = g(z) = \frac{1}{1+e^{-z}}$

$\frac{dg}{dz} = g(z)\big(1 - g(z)\big) = \mu(1-\mu)$

$$\ell(\boldsymbol{w}) = \sum_n \left(y^{(n)} \log \mu^{(n)} + \left(1 - y^{(n)}\right) \log\left(1 - \mu^{(n)}\right)\right)$$

$$\frac{\partial \ell}{\partial \boldsymbol{w}} = \sum_n \left(\frac{y^{(n)}}{\mu^{(n)}} - \frac{1-y^{(n)}}{1-\mu^{(n)}} \right) \frac{\partial g}{\partial z} \frac{\partial f}{\partial \vec{w}}$$

$$= \sum_n \left(\frac{y^{(n)} - \mu^{(n)}}{\mu^{(n)}\left(1-\mu^{(n)}\right)} \right) \mu^{(n)}\left(1-\mu^{(n)}\right) \vec{x}^{(n)T}$$

M(C)LE for Logistic Regression

$$z = f(\boldsymbol{w}, \boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} \qquad \mu = g(z) = \frac{1}{1+e^{-z}}$$

$$\nabla_{\boldsymbol{w}} f(\boldsymbol{w}, \boldsymbol{x}) = \boldsymbol{x} \qquad \frac{dg}{dz} = g(z)\big(1 - g(z)\big) = \mu(1-\mu)$$

$$\ell(\boldsymbol{w}) = \sum_n \left(y^{(n)} \log \mu^{(n)} + \left(1 - y^{(n)}\right) \log\left(1 - \mu^{(n)}\right)\right)$$

$$\frac{\partial \ell}{\partial \boldsymbol{w}} = \sum_n \left(\frac{y^{(n)}}{\mu^{(n)}} - \frac{1-y^{(n)}}{1-\mu^{(n)}}\right) \frac{\partial g}{\partial f} \frac{\partial f}{\partial \boldsymbol{w}}$$

$$= \sum_n \left(\frac{y^{(n)} - \mu^{(n)}}{\mu^{(n)}\left(1-\mu^{(n)}\right)}\right) \cancel{\mu^{(n)}\left(1 - \mu^{(n)}\right)}\, \boldsymbol{x}^{(n)^T}$$

$$= \sum_n \left(y^{(n)} - \mu^{(n)}\right) \boldsymbol{x}^{(n)^T}$$

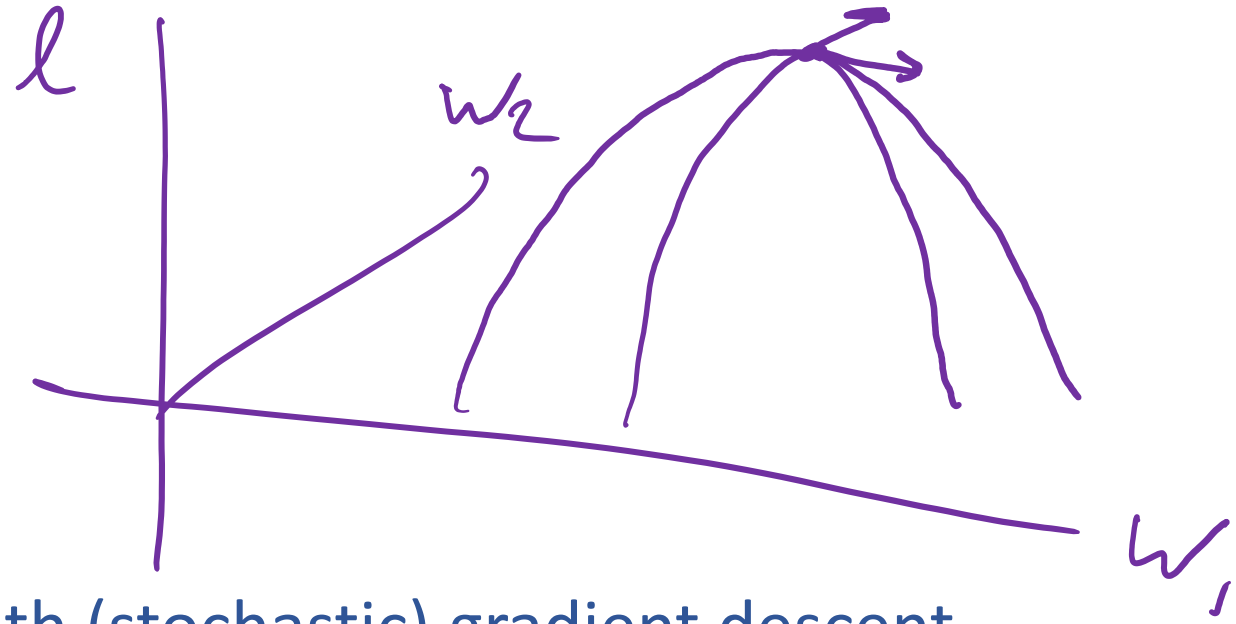
M(C)LE for Logistic Regression

$z = f(\boldsymbol{w}, \boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$ $\mu = g(z) = \frac{1}{1+e^{-z}}$

$\ell(\boldsymbol{w}) = \sum_n \left(y^{(n)} \log \mu^{(n)} + \left(1 - y^{(n)}\right) \log\left(1 - \mu^{(n)}\right)\right)$

$\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}) = \sum_n \left(y^{(n)} - \underline{\mu^{(n)}}\right) \boldsymbol{x}^{(n)}$

$\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}) = 0?$



No closed form solution ☹

Back to iterative methods. Solve with (stochastic) gradient descent, Newton's method, or Iteratively Reweighted Least Squares (IRLS)

Logistic Function

Cool note: Logistic function is related the invers of logit function!

Odds: Ratio of two probabilities. For $Y \sim Bern(p)$, $\frac{p(Y=1)}{p(Y=0)} = \frac{p}{1-p}$

Logit function: Log odds. $\log \frac{p(Y=1)}{p(Y=0)} = \log \frac{p}{1-p}$

$$z = logit(p) \stackrel{\Delta}{=} \log \frac{p}{1-p}$$

$$p = logit^{-1}(z) = \frac{1}{1+e^{-z}}$$

Log Odds and Logistic Regression

Formulate log odds as linear model of X:

$$\log \frac{p(Y=1 \mid X=\boldsymbol{x}, \boldsymbol{w})}{p(Y=0 \mid X=\boldsymbol{x}, \boldsymbol{w})} = \boldsymbol{w}^T \boldsymbol{x}$$

Equivalent to logistic representation:

$$p(Y=1 \mid X=\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1+e^{-\boldsymbol{w}^T \boldsymbol{x}}}$$

Log Odds and Logistic Regression (Multi-class!)

Formulate log odds as linear model of X:

W

$$\log \frac{p(Y=1 \mid X=\boldsymbol{x}, \boldsymbol{W})}{p(Y=K \mid X=\boldsymbol{x}, \boldsymbol{W})} = \boldsymbol{w}_1^T \boldsymbol{x}$$

$$\log \frac{p(Y=2 \mid X=\boldsymbol{x}, \boldsymbol{W})}{p(Y=K \mid X=\boldsymbol{x}, \boldsymbol{W})} = \boldsymbol{w}_2^T \boldsymbol{x}$$

$$\vdots$$

$$\log \frac{p(Y=K-1 \mid X=\boldsymbol{x}, \boldsymbol{W})}{p(Y=K \mid X=\boldsymbol{x}, \boldsymbol{W})} = \boldsymbol{w}_{K-1}^T \boldsymbol{x}$$

Equivalent to softmax representation:

$$p(Y=k \mid X=\boldsymbol{x}, W) = \frac{e^{\boldsymbol{w}_k^T \boldsymbol{x}}}{1+\sum_{j=1}^{K-1} e^{\boldsymbol{w}_j^T \boldsymbol{x}}}$$

$k \in \{1, \ldots K-1\}$

$$p(Y=K \mid X=\boldsymbol{x}, W) = \frac{1}{1+\sum_{j=1}^{K-1} e^{\boldsymbol{w}_j^T \boldsymbol{x}}}$$

OR

$$p(Y=k \mid X=\boldsymbol{x}, W) = \frac{e^{\boldsymbol{w}_k^T \boldsymbol{x}}}{\sum_{j=1}^{K} e^{\boldsymbol{w}_j^T \boldsymbol{x}}}$$

Multi-class Logistic Regression

$p(Y \mid X, \theta)$

$p(Y \mid X, \boldsymbol{W}) = \prod_{n=1}^{N} p(y^{(n)} \mid \boldsymbol{x}^{(n)}, \boldsymbol{W})$

$$p(y^{(n)} = k \mid X = \boldsymbol{x}^{(n)}, W) = \frac{e^{\boldsymbol{w}_k^T \boldsymbol{x}^{(n)}}}{\sum_{j=1}^{K} e^{\boldsymbol{w}_j^T \boldsymbol{x}^{(n)}}}$$

What is the *conditional* likelihood?

$$L(\boldsymbol{w}) = \prod_n \frac{e^{\boldsymbol{w}_k^T \boldsymbol{x}^{(n)}}}{\sum_{j=1}^{K} e^{\boldsymbol{w}_j^T \boldsymbol{x}^{(n)}}}$$

What is the hypothesis function?

$$\hat{y} = h_W(\boldsymbol{x}) = \operatorname*{argmax}_{k} \ \mathrm{softmax}(x, W)$$