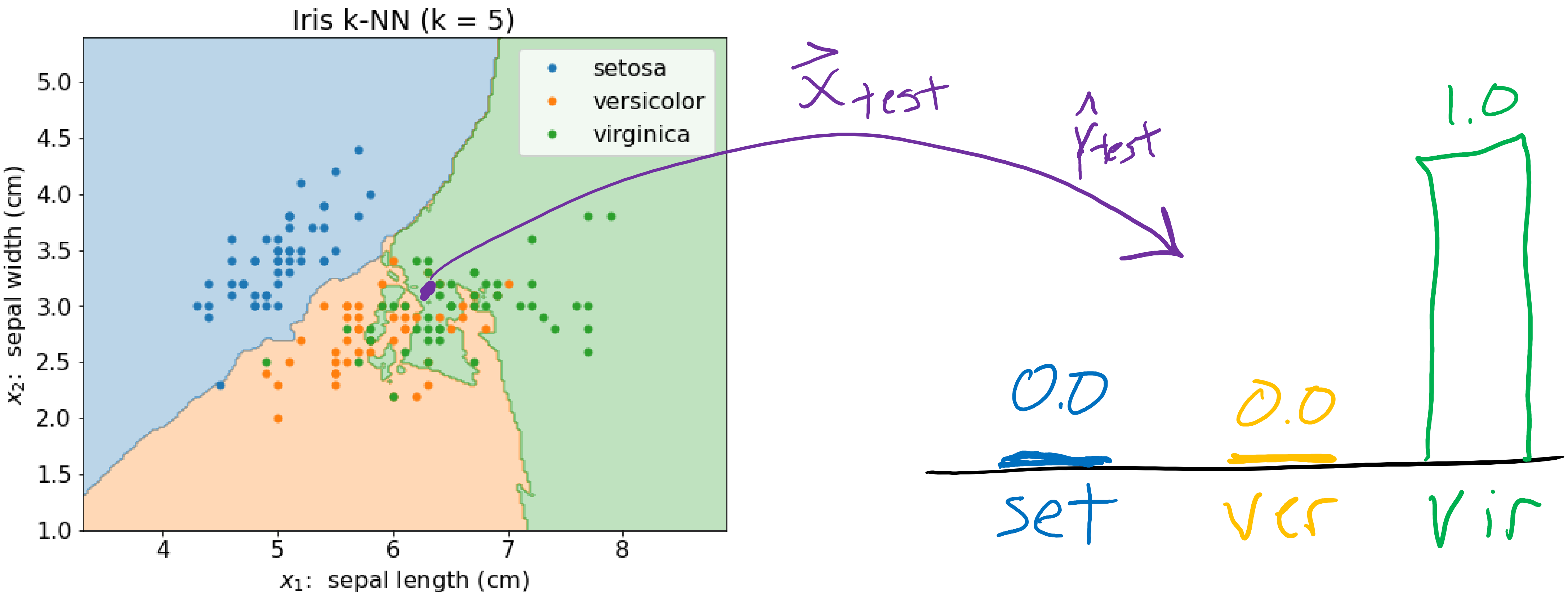
10-315
Introduction to ML

Logistic Regression

Instructor: Pat Virtue

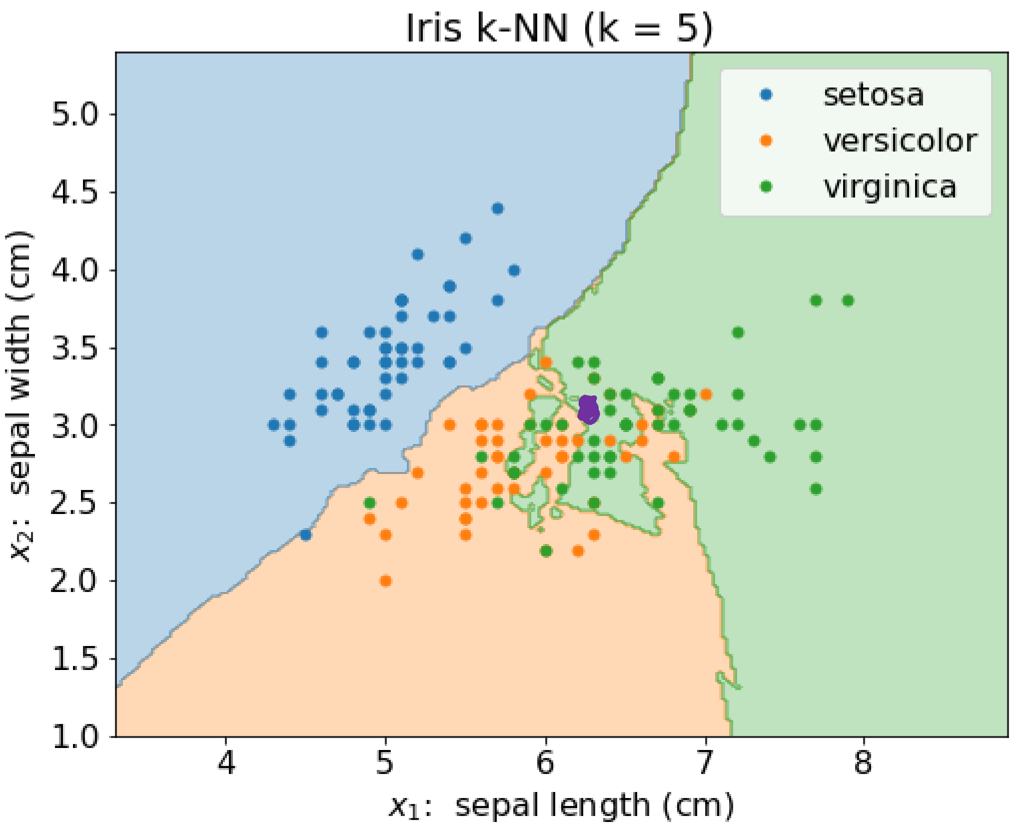
Classification Decisions

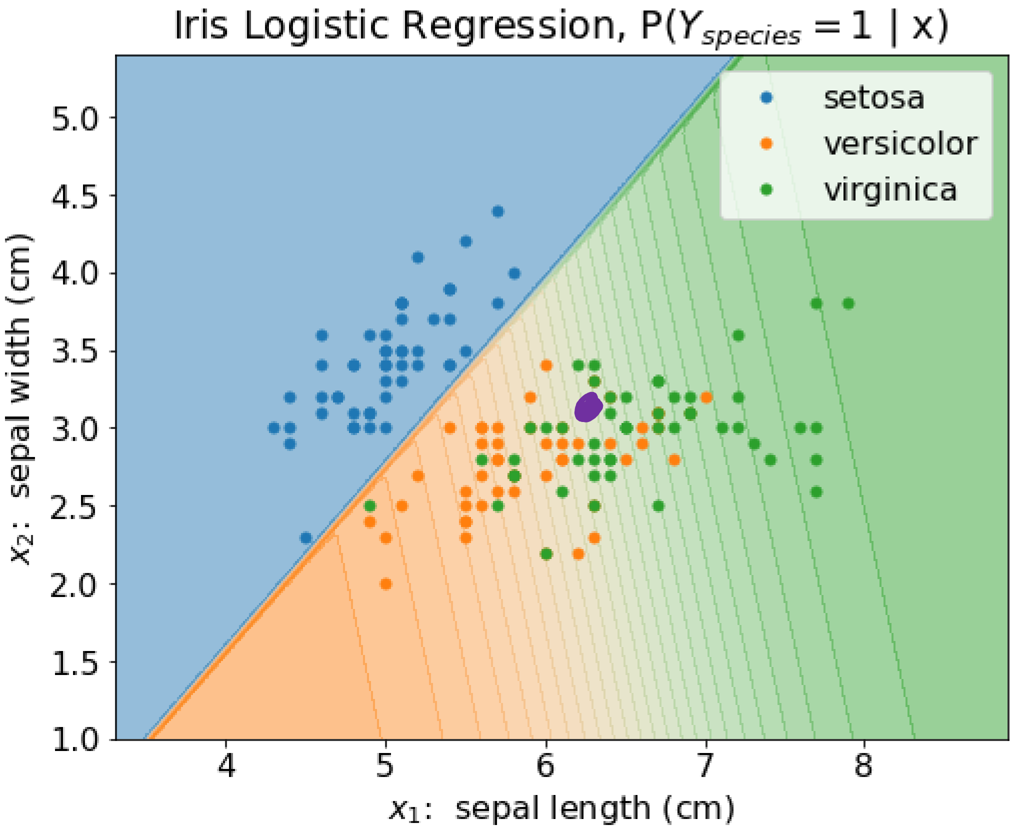
Predicting one specific class is troubling, especially when we know that there is some uncertainty in our prediction



Classification Probability

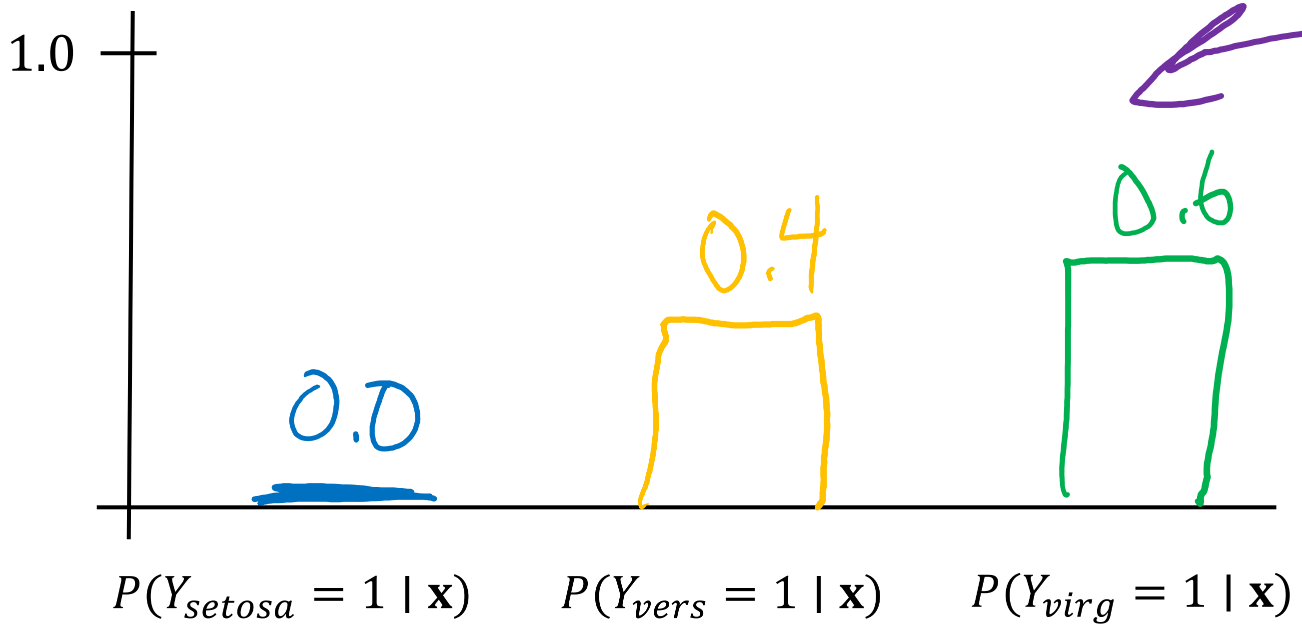
Constructing a model than can return the probability of the output being a specific class could be incredibly useful

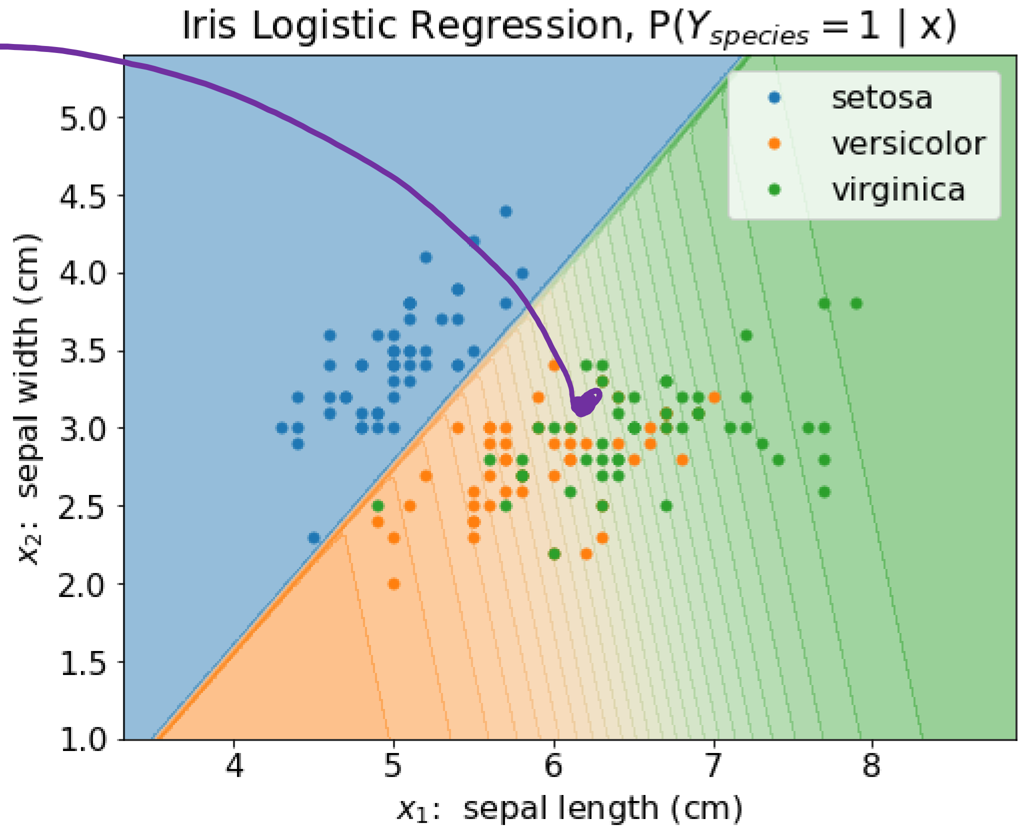




Classification Probability

Constructing a model than can return the probability of the output being a specific class could be incredibly useful





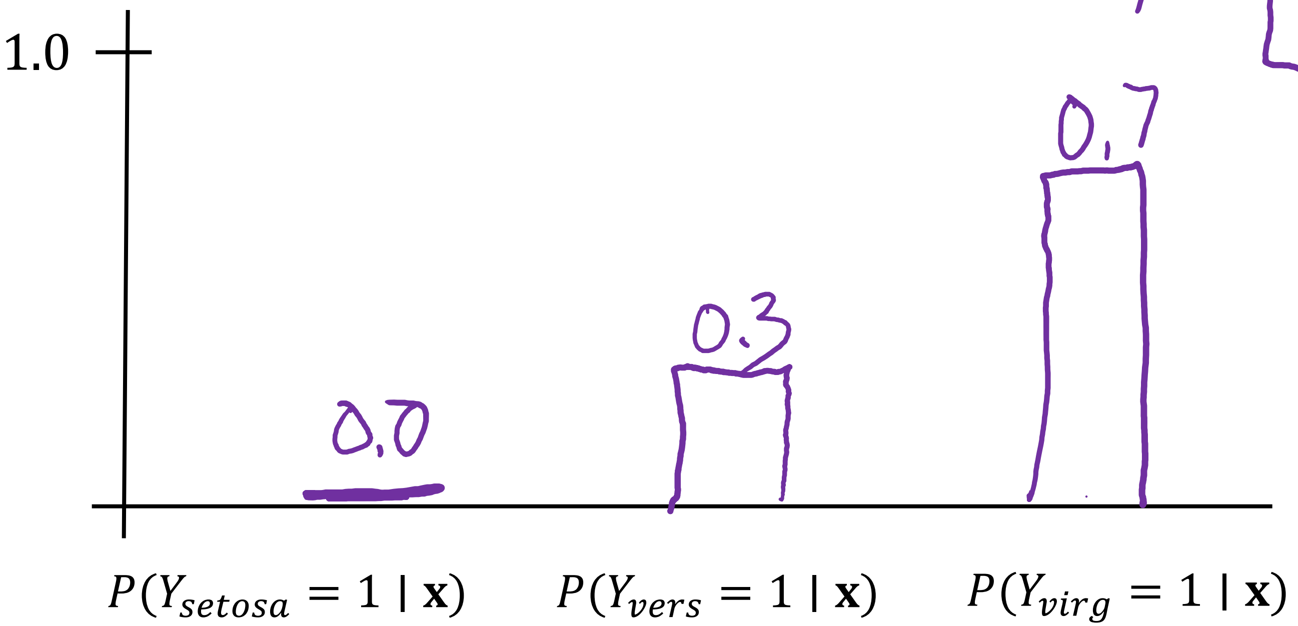
We can still make decisions, .e.g,

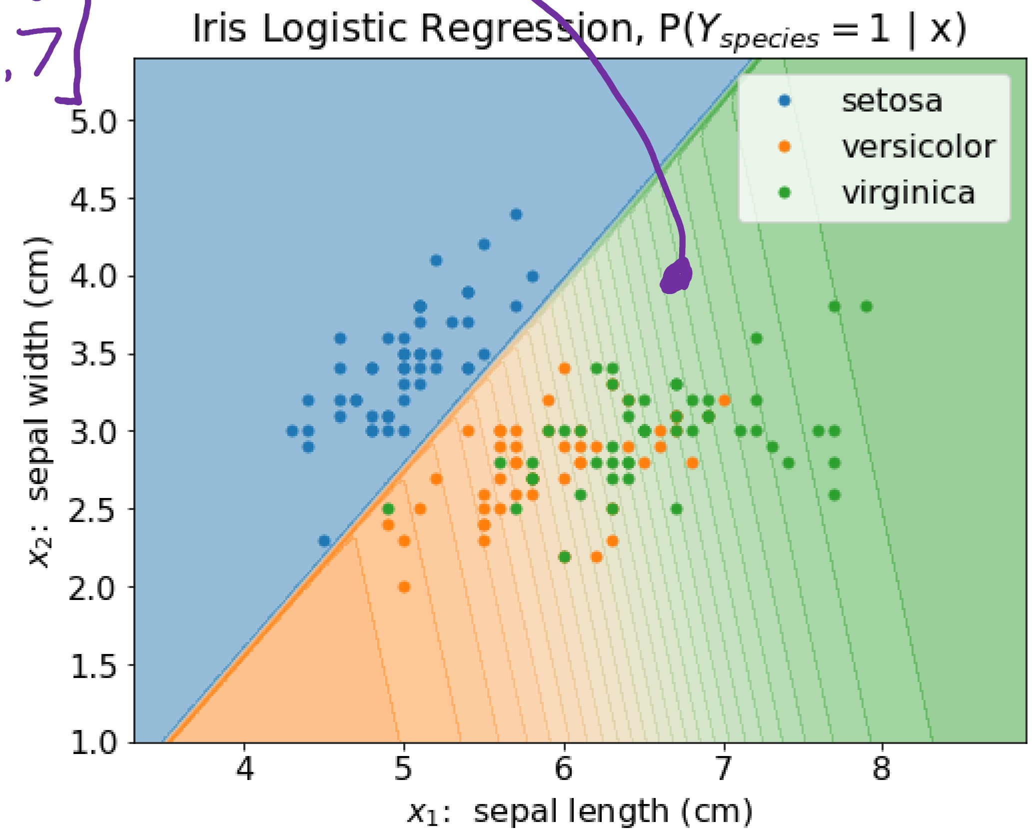
$$\underset{k}{\operatorname{argmax}}\, P(Y_k = 1 \mid \mathbf{x})$$

Loss for Probabilty Disributions

We need a way to compare how good/bad each prediction is

$\hat{y} = \begin{bmatrix} 0.0 \\ 0.3 \\ 0.7 \end{bmatrix}$





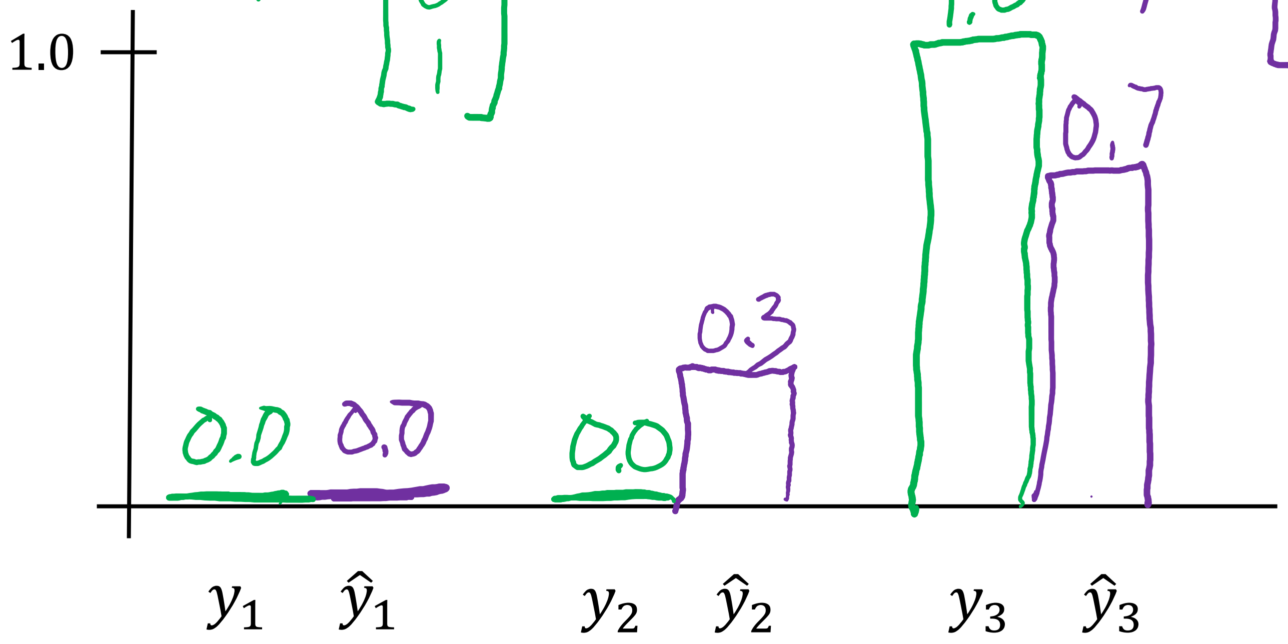
Cross-entropy loss

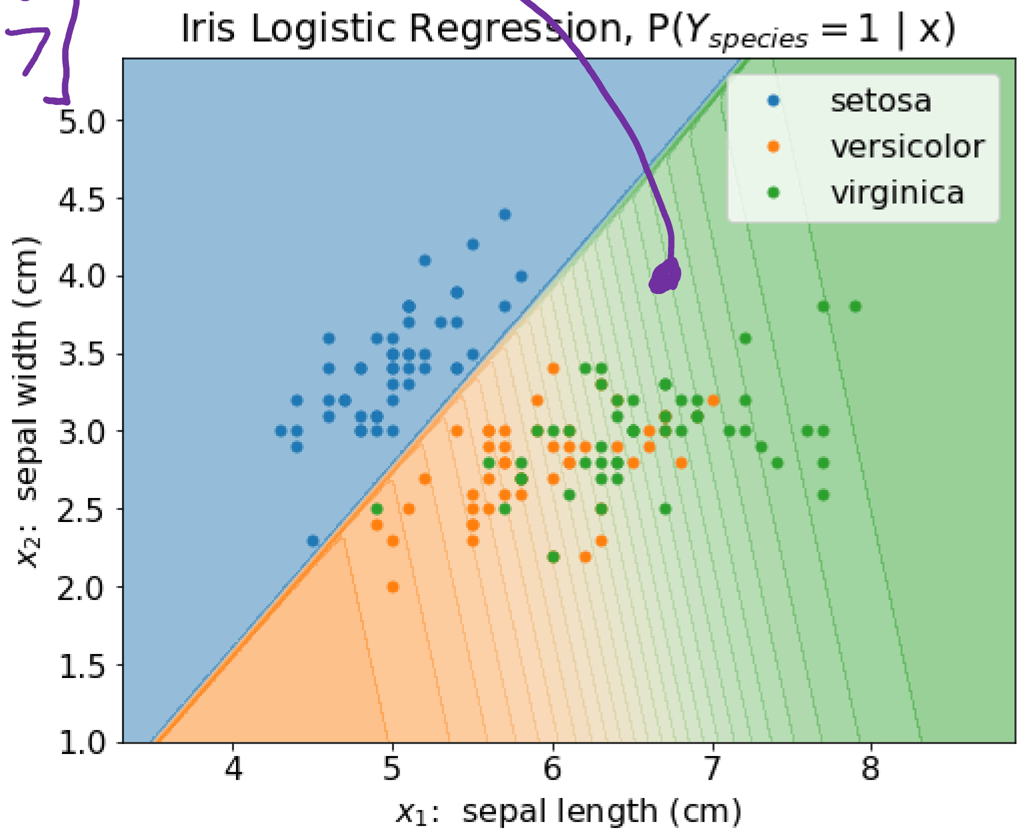
$$\ell(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{k=1}^{K} y_k \log \hat{y}_k$$

Loss for Probabilty Disributions

We need a way to compare how good/bad each prediction is

$y = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ $\hat{y} = \begin{bmatrix} 0.0 \\ 0.3 \\ 0.7 \end{bmatrix}$





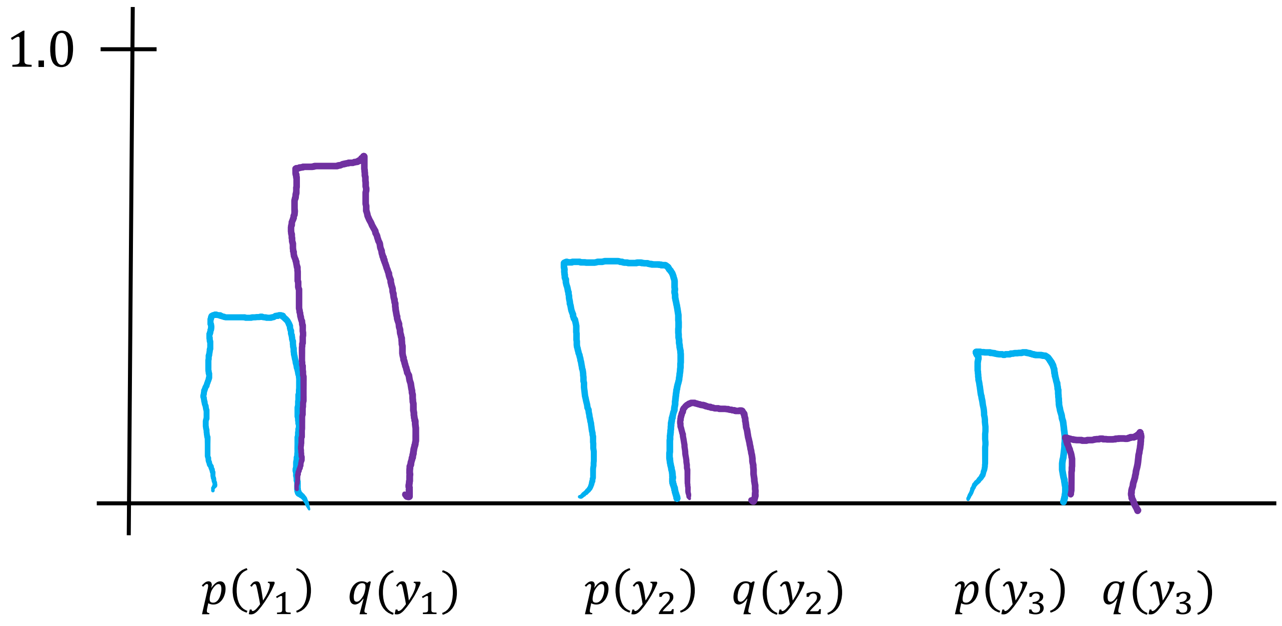
Cross-entropy loss

$$\ell(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{k=1}^{K} y_k \log \hat{y}_k$$

Loss for Probabilty Disributions

Cross-entropy more generally is a way to compare any to probability distributions*

*when used in logistic regression
$\mathbf{y}$ is always a one-hot vector



Cross-entropy loss

$$H(P, Q) = -\sum_{k=1}^{K} p(y_k) \log q(y_k)$$

Empirical Risk Minimization

Still doing empirical risk minimization, just with a cross-enropy loss

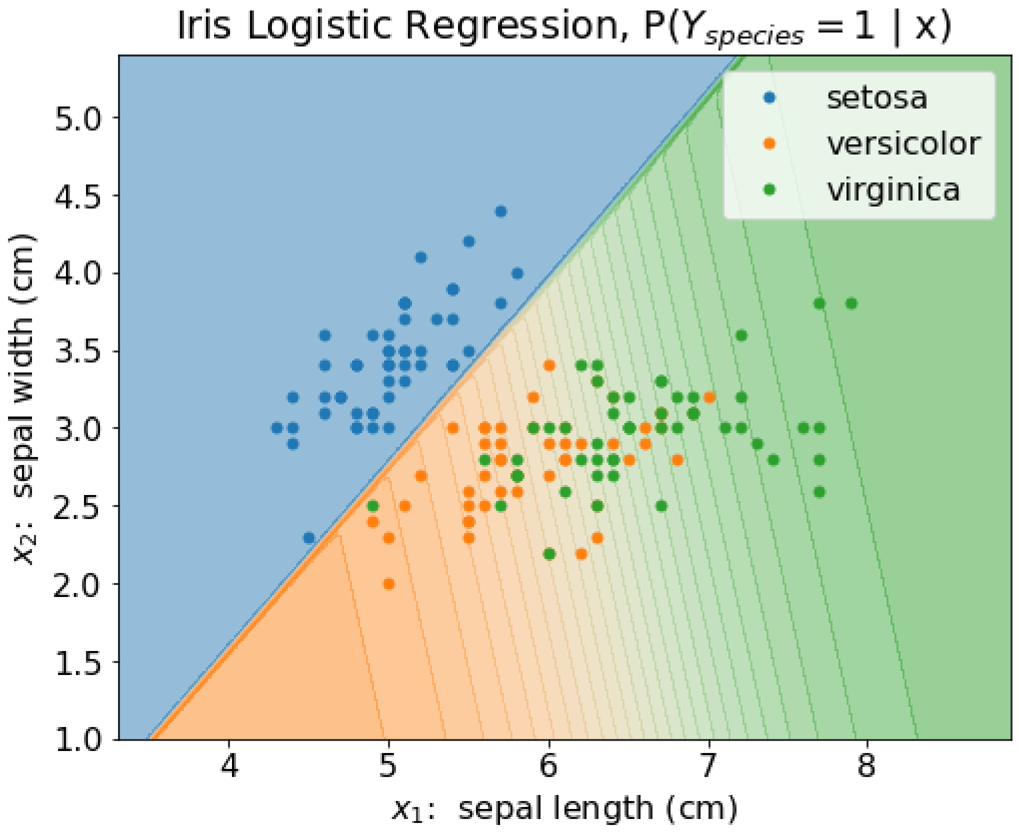
$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \hat{R}(h)$$

$$\hat{R}(h) = \frac{1}{N} \sum_{i=1}^{N} \ell\left(y^{(i)}, h\left(x^{(i)}\right)\right)$$

Cross-entropy loss

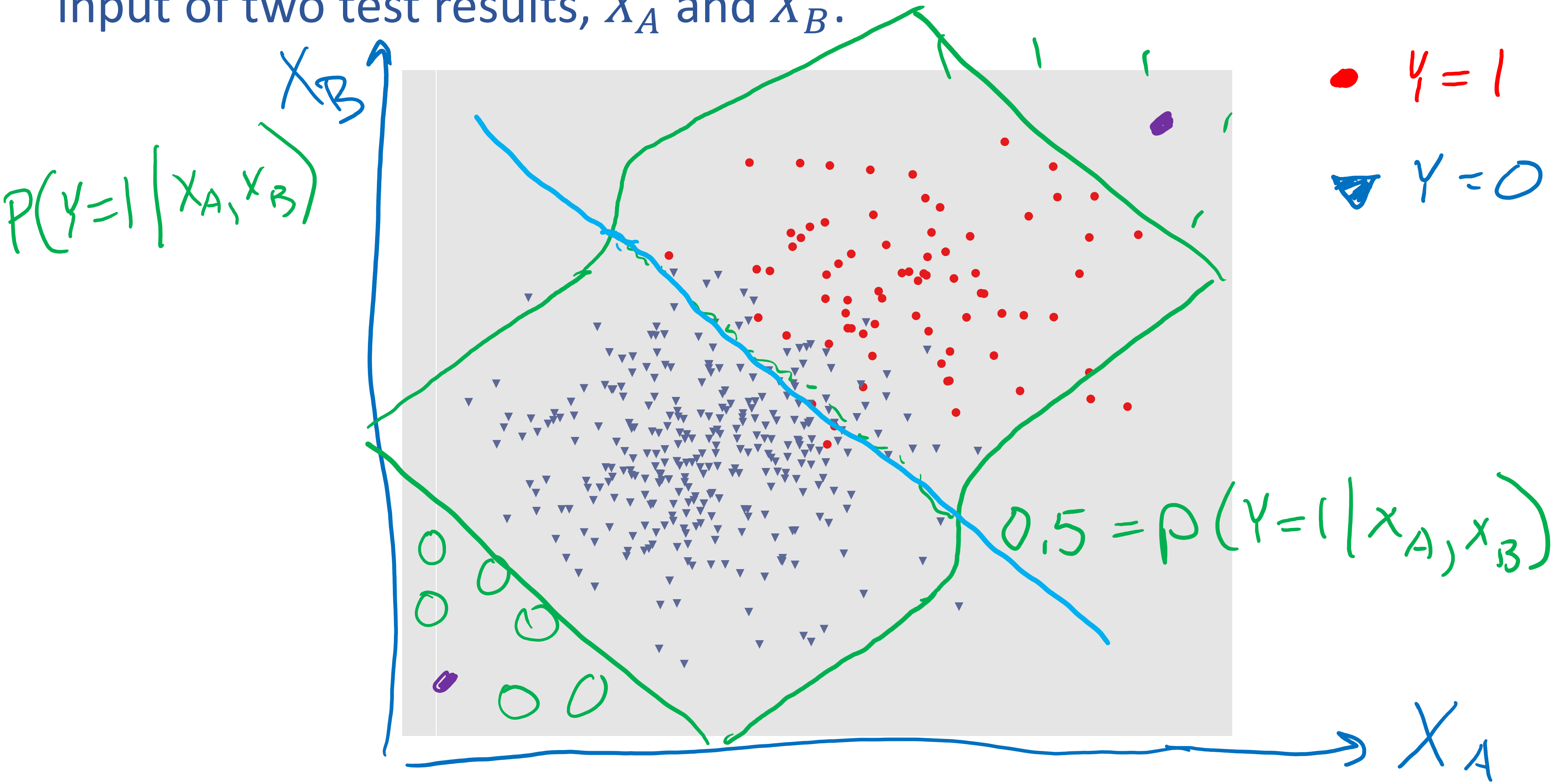
$\ell(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{k=1}^{K} y_k \log \hat{y}_k$

But now we need a model $h_{\boldsymbol{\theta}}(\mathbf{x})$ that returns values that look like probabilities



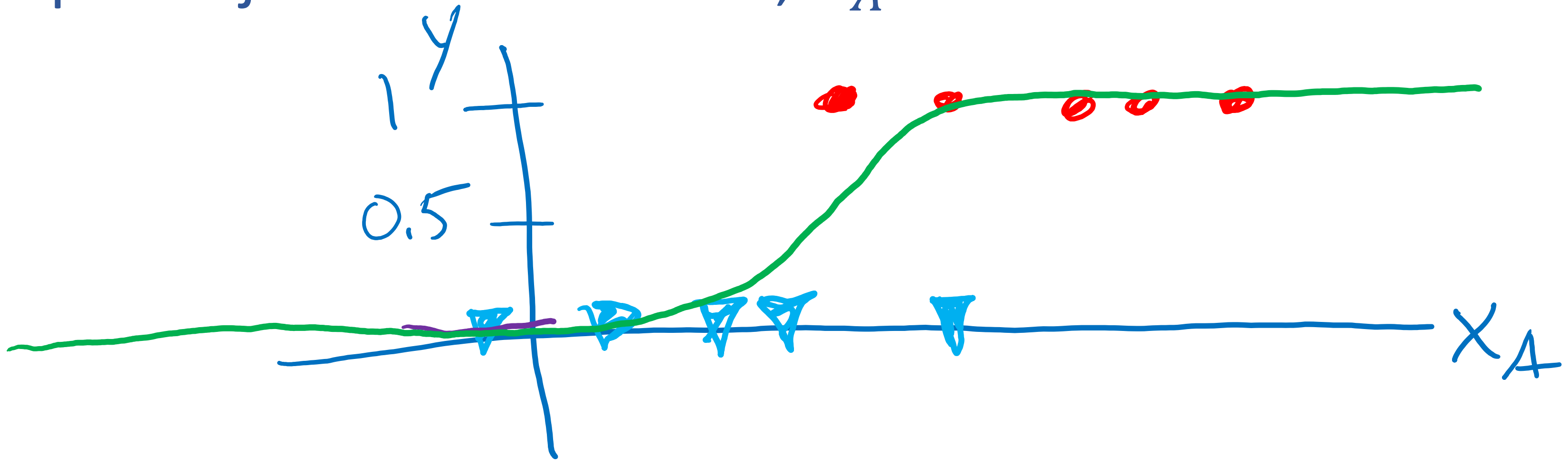
Prediction for Cancer Diagnosis

Learn to predict if a patient has cancer ($Y = 1$) or not ($Y = 0$) given the input of two test results, X_A and X_B.



Prediction for Cancer Diagnosis

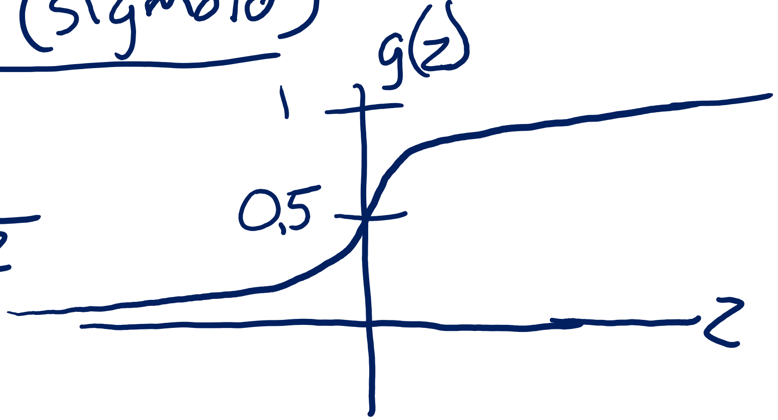
Learn to predict if a patient has cancer ($Y = 1$) or not ($Y = 0$) given the input of just one test result, X_A.



$p(Y=1 | x_A)$

logistic function (sigmoid)

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

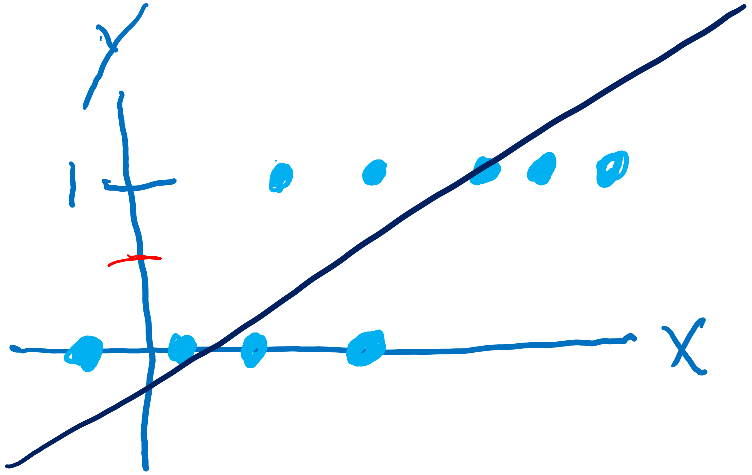


logistic regression

$$p(Y=1 | \vec{x}, \vec{\theta}) = g(\theta^T x)$$

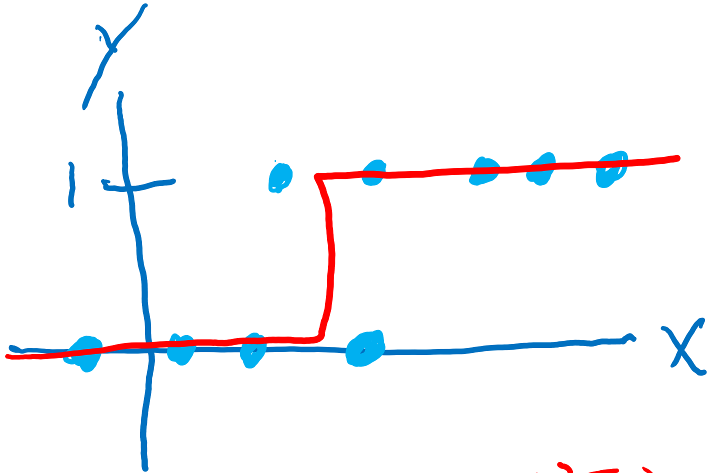
Building on a Linear Model

Linear vs Thresholded Linear vs Logistic Linear



$\hat{y} = \vec{\theta}^T \vec{x}$

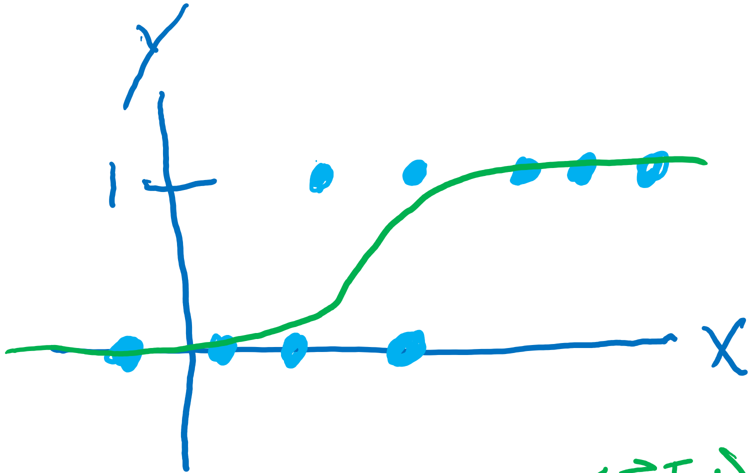
☹ not classification



$\hat{y} = g_{thresh}(\vec{\theta}^T \vec{x})$

☹ classification only (0/1)

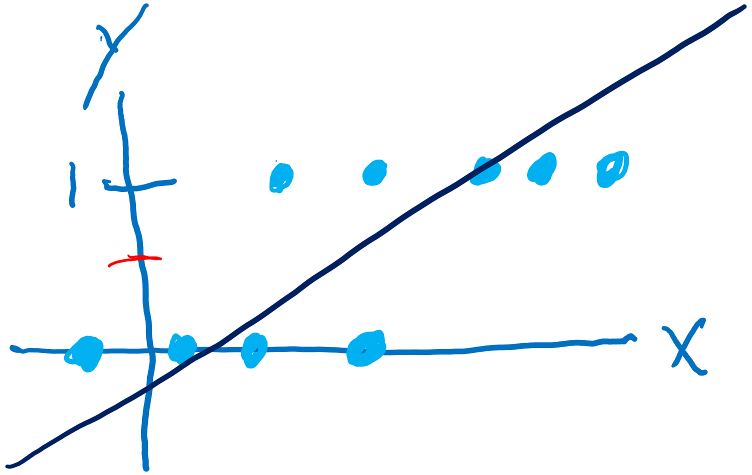
☹ zero derivatives



$\hat{y} = g_{logistic}(\vec{\theta}^T \vec{x})$

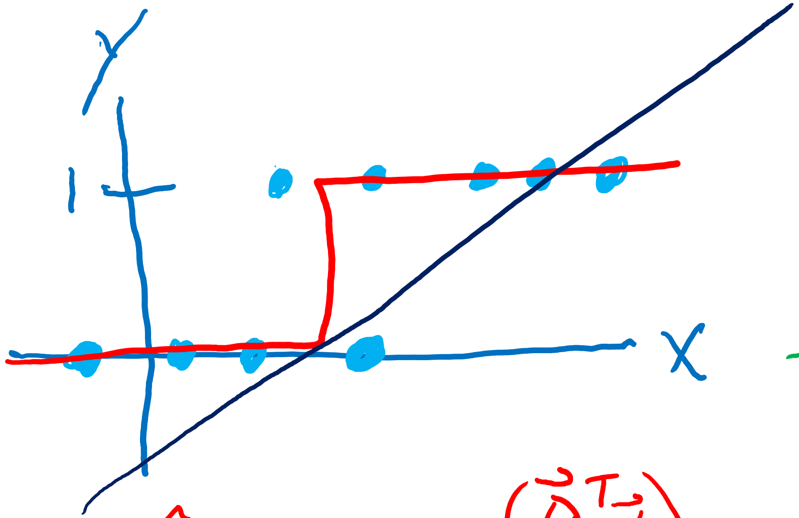
Building on a Linear Model

Linear vs Thresholded Linear vs Logistic Linear



$\hat{y} = \vec{\theta}^T x$

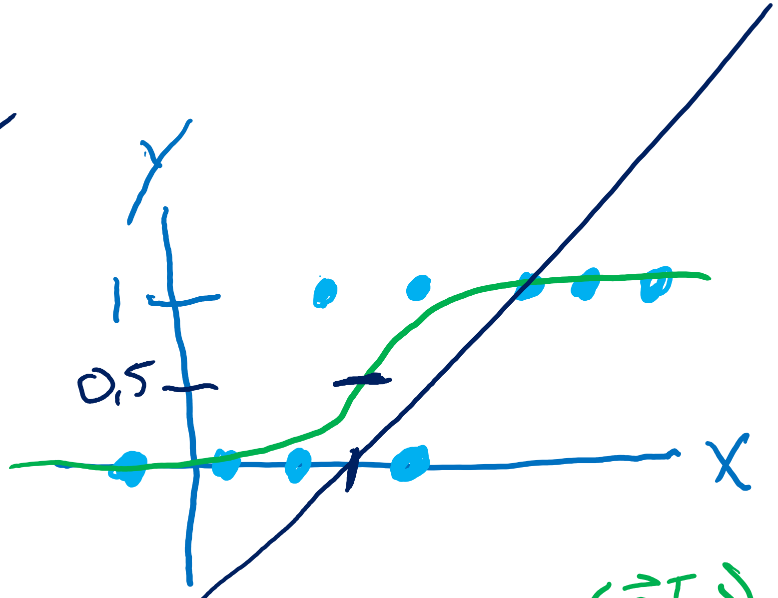
☹ not classification



$\hat{y} = g_{thresh}(\vec{\theta}^T \vec{x})$

☹ classification only (0/1)

☹ zero derivatives



$\hat{y} = g_{logistic}(\vec{\theta}^T \vec{x})$

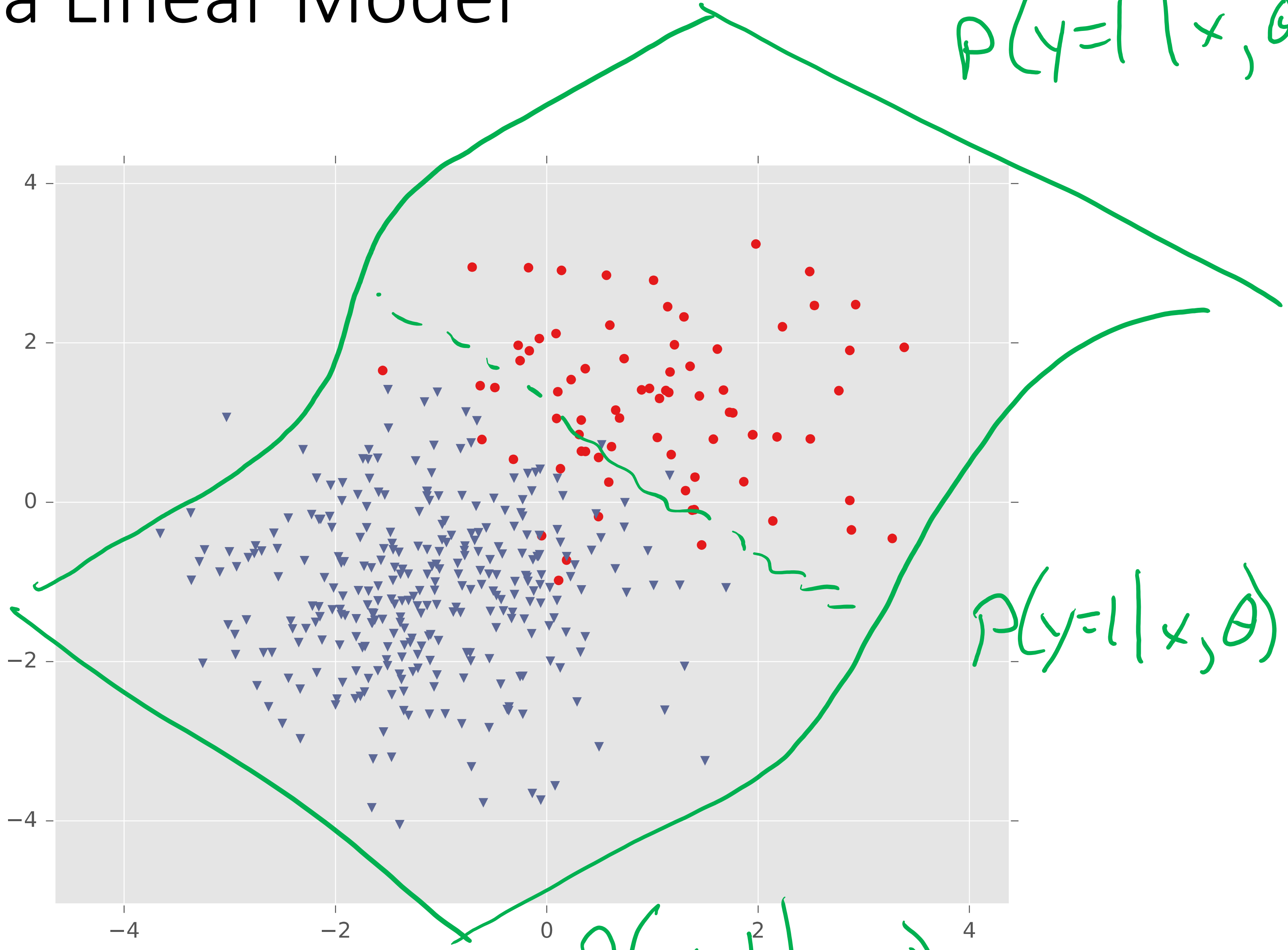
↑

$\sigma(\vec{\theta}^T \vec{x})$

Building on a Linear Model

$P(y=1 \mid x, \theta) = 1$

$$\vec{x} = \begin{bmatrix} 1 \\ x_A \\ x_B \end{bmatrix}$$



$P(y=1 \mid x, \theta) = 0{,}5$

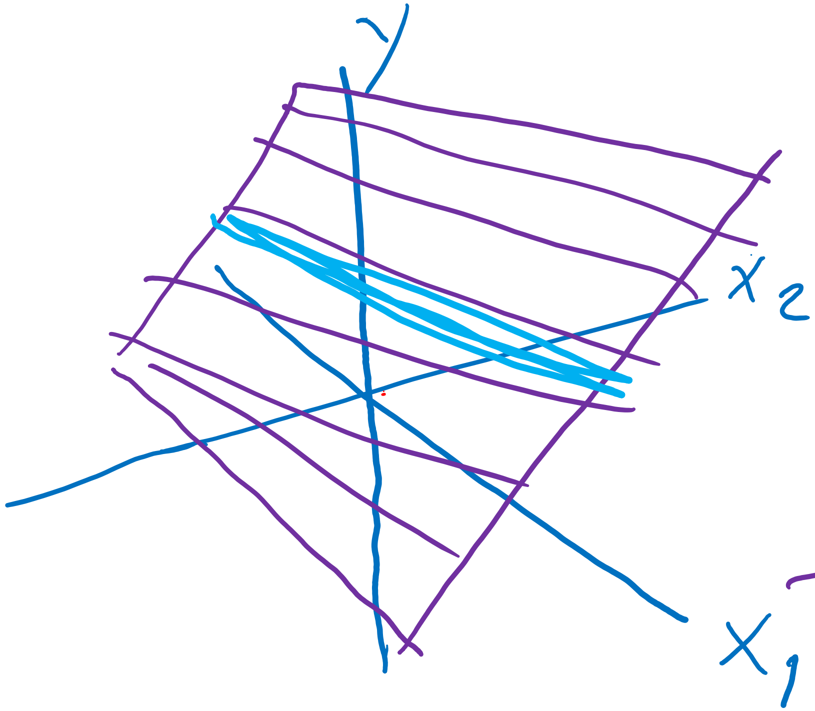
$$\vec{\theta} = \begin{bmatrix} b \\ w_1 \\ w_2 \end{bmatrix}$$

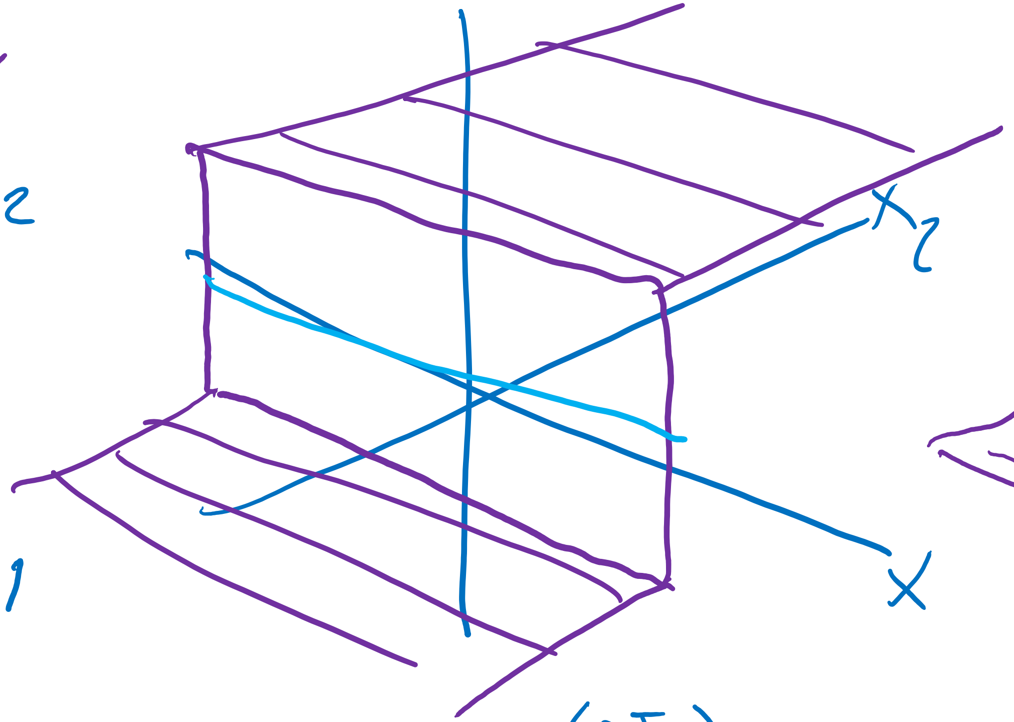
$P(y=1 \mid x, \theta) = 0$

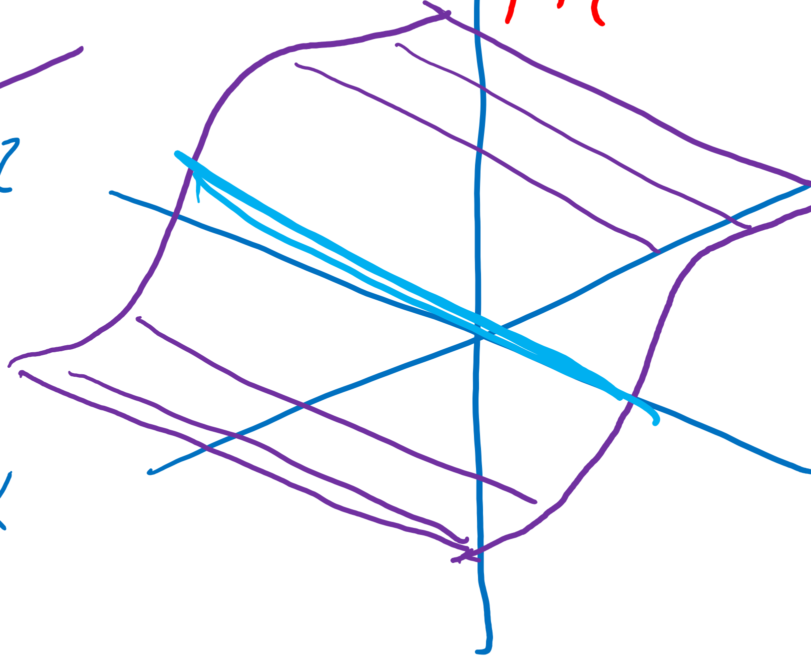
Building on a Linear Model

$$\frac{d\sigma}{dz} = \sigma(z)(1-\sigma(z))$$

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







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

$y = \text{sign}(\theta^T x)$

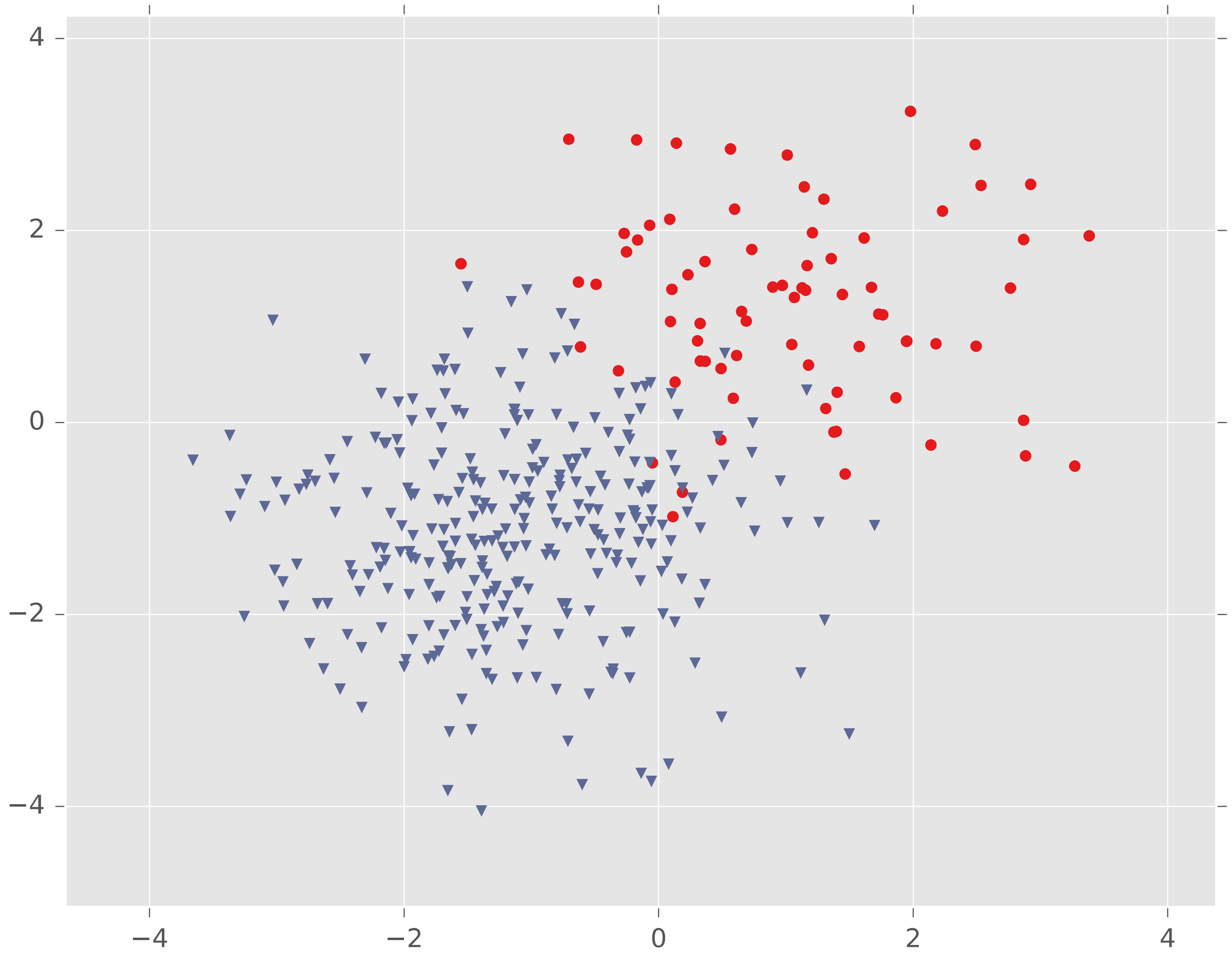
$y = \sigma(\theta^T x)$

logistic function

$p(y|x)$

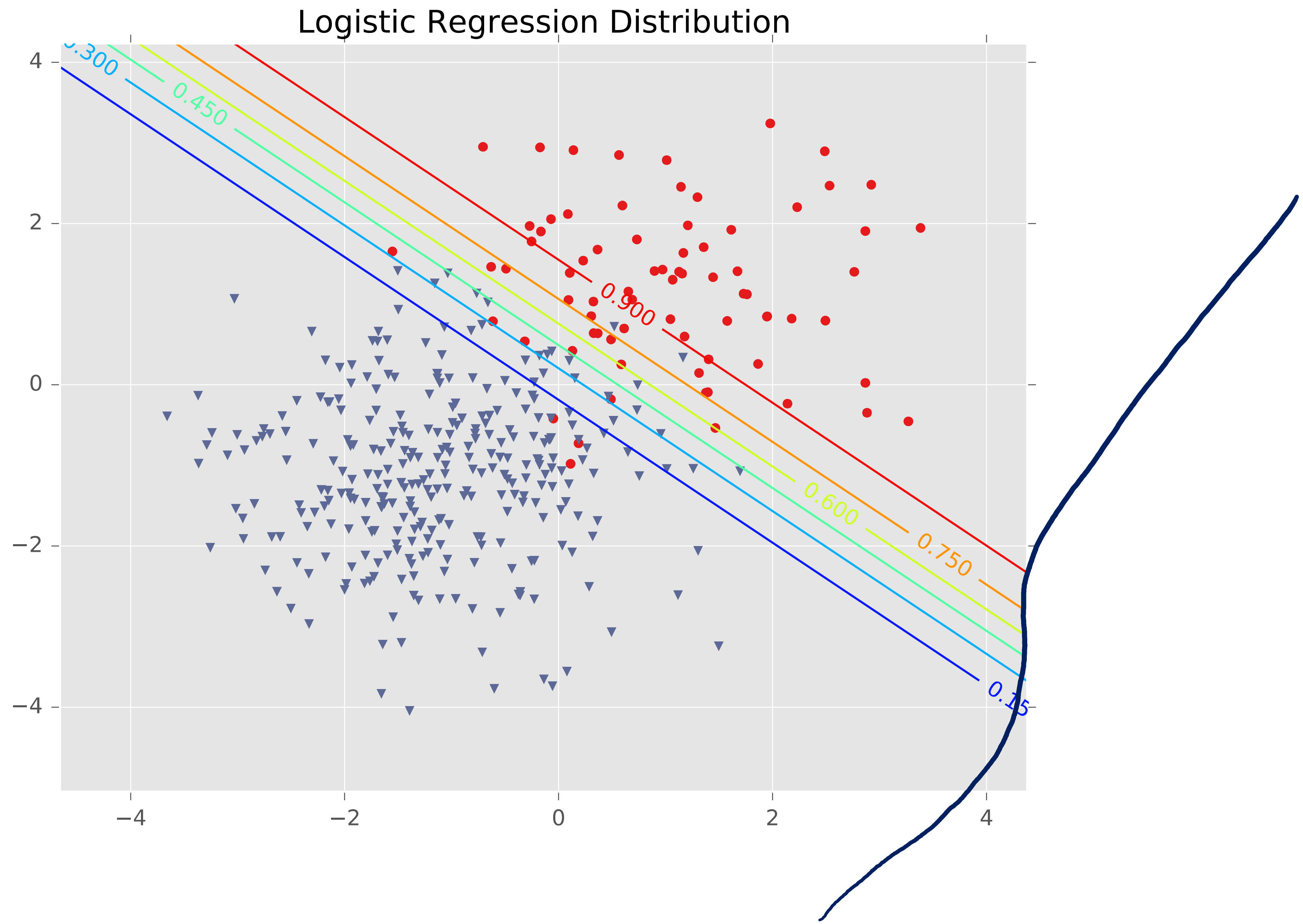
$y=1$

Logistic Regression



Slide credit: CMU MLD Matt Gormley

Logistic Regression



Slide credit: CMU MLD Matt Gormley

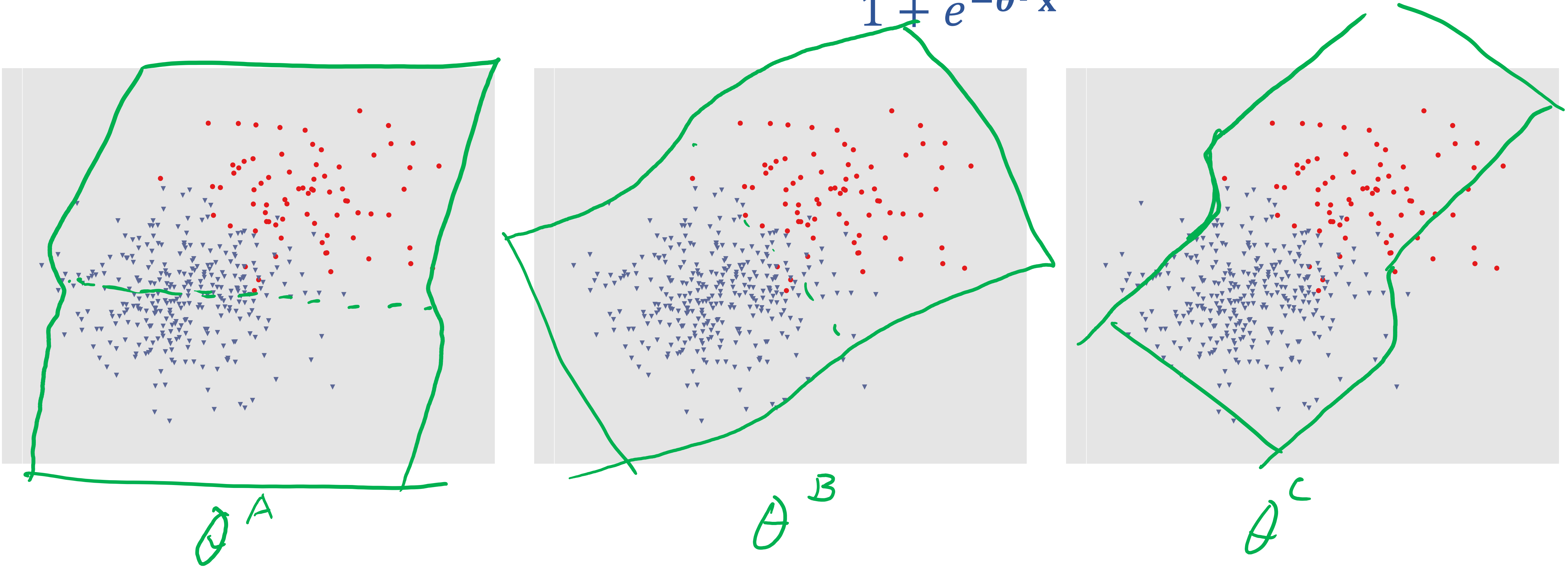
Logistic Regression Decision Boundary

Slide credit: CMU MLD Matt Gormley

Optimizing a Model for Cancer Diagnosis

Learn to predict if a patient has cancer ($Y = 1$) or not ($Y = 0$) given the input of two test results, X_A, X_B. Note: bias term included in $\mathbf{x}$.

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



Binary Logistic Regression

$\hat{y} \in [0,1]$

1) Model

$$\hat{y} = h(\vec{x}) = P(y=1|\vec{x}) = \frac{1}{1+e^{-\theta^T \vec{x}}}$$

$$P(y=0|\vec{x}) = 1 - \frac{1}{1+e^{-\theta^T \vec{x}}}$$

2) Objective function

$$\frac{1}{N}\sum_{i=1}^{N} \ell(y^{(i)}, \hat{y}^{(i)}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_k^{(i)} \log \hat{y}_k^{(i)}$$

$$-\frac{1}{N}\sum \left[y_1^{(i)} \log \hat{y}_1 + y_2^{(i)} \log \hat{y}_2 \right]$$

3) Solve for $\hat{\boldsymbol{\theta}}$

$$-\frac{1}{N}\sum \left[y^{(i)} \log \hat{y}^{(i)} + (1-y^{(i)}) \log (1-\hat{y}^{(i)}) \right]$$

Binary Logistic Regression

Gradient

$\vec{x} \in \mathbb{R}^M \quad \vec{\theta} \in \mathbb{R}^M$

$\nabla_\theta J^{(i)}(\theta): \mathbb{R}^M \to \mathbb{R}^M$

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \underbrace{\log p(y^{(i)} \mid \vec{x}^{(i)}, \vec{\theta})}_{J^{(i)}(\vec{\theta})}$$

$$\nabla_\theta J^{(i)} = \begin{bmatrix} \frac{\partial J^{(i)}}{\partial \theta_1} \\ \vdots \\ \frac{\partial J^{(i)}}{\partial \theta_m} \\ \vdots \end{bmatrix}$$

$\vec{x}^{(i)} = \begin{bmatrix} x_m^{(i)} \end{bmatrix}$

$$\frac{\partial J^{(i)}}{\partial \theta_m} = -\left(y^{(i)} - \sigma(\theta^T x^{(i)})\right) x_m^{(i)}$$

SGD

$$\nabla_\theta J^{(i)} = -\left(y^{(i)} - \sigma(\theta^T x^{(i)})\right) \vec{x}^{(i)}$$

Solve Logistic Regression

$J(\theta)$ is convex

☺

$\hat{y} = g(\boldsymbol{\theta}^T \mathbf{x})$ $\qquad$ $g(z) = \frac{1}{1+e^{-z}}$

$$J(\boldsymbol{\theta}) = -\frac{1}{N}\sum_i \left(y^{(i)} \log \hat{y}^{(i)} + \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right)\right)$$

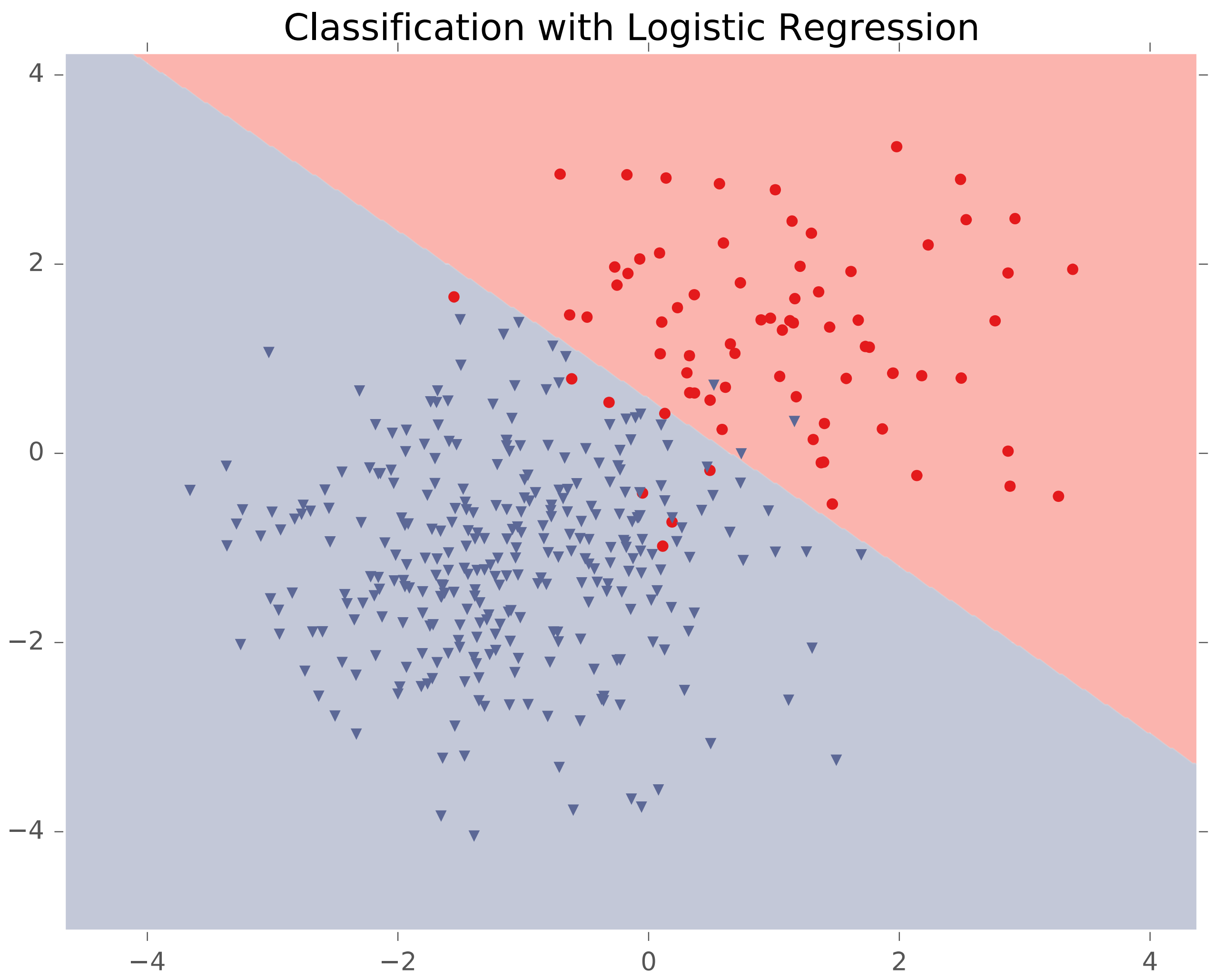
$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = -\frac{1}{N}\sum_i \left(y^{(i)} - \hat{y}^{(i)}\right) \mathbf{x}^{(i)}$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = 0?$$

No closed form solution ☹

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

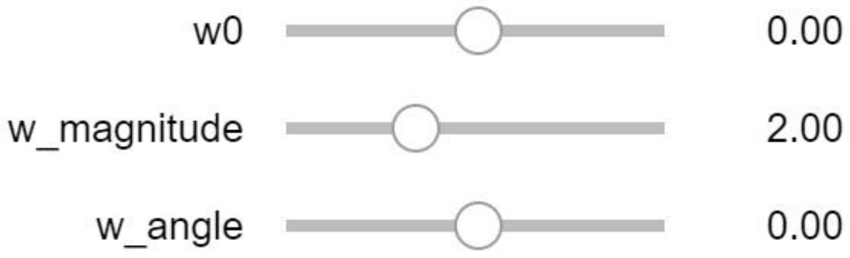
Logistic Regression Decision Boundary



Slide credit: CMU MLD Matt Gormley

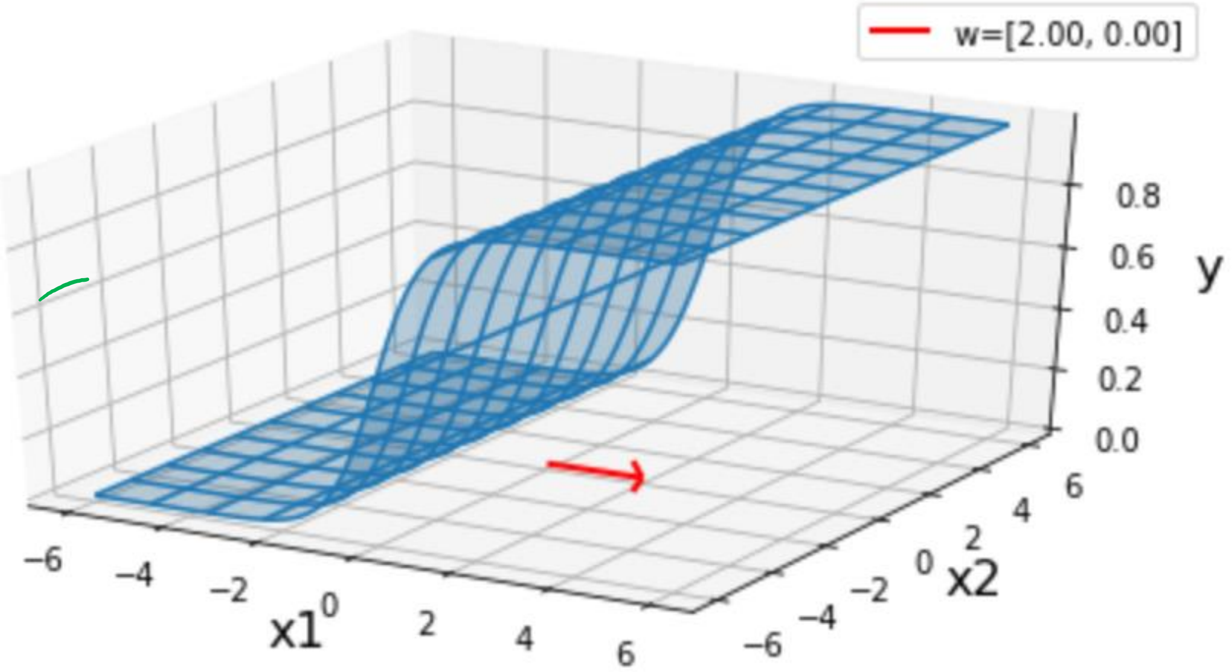
Exercise

Interact with the linear_logistic.ipynb posted on the course website schedule



b

$\{ w_1, w_2$



Linear in Higher Dimensions

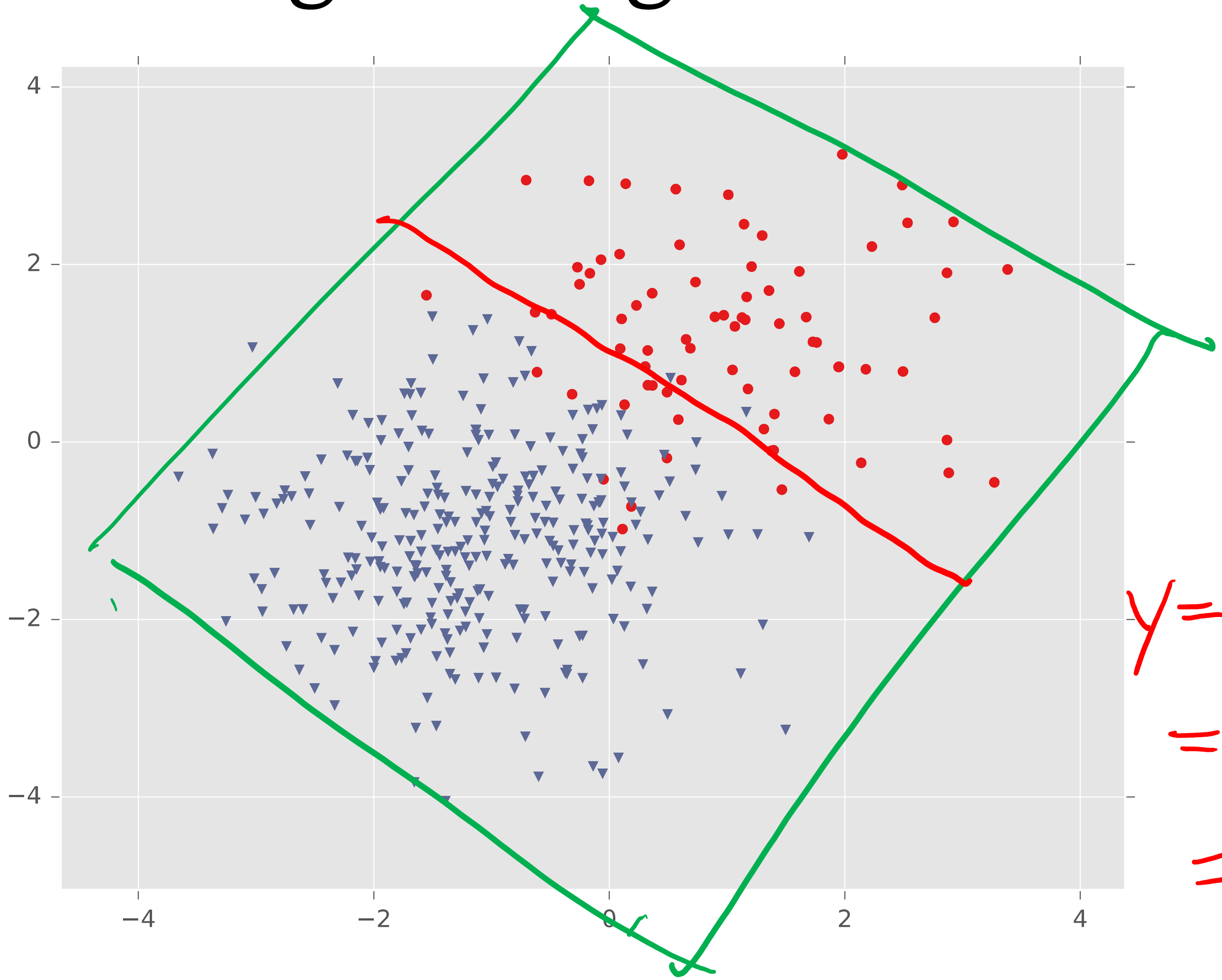
1-D $y = wx + b$

2-D $y = w_1 x_1 + w_2 x_2 + b$

What are these linear shapes called for 1-D, 2-D, 3-D, M-D input?

	$\boldsymbol{x} \in \mathbb{R}$	$\boldsymbol{x} \in \mathbb{R}^2$	$\boldsymbol{x} \in \mathbb{R}^3$	$\boldsymbol{x} \in \mathbb{R}^M$
$y = \boldsymbol{w}^T \boldsymbol{x} + b$	line	plane	hyperplane	hyperplane
$\boldsymbol{w}^T \boldsymbol{x} + b = 0$	point	line	plane	hyperplane
$\boldsymbol{w}^T \boldsymbol{x} + b \geq 0$	halfline	halfplane	halfspace	halfspace

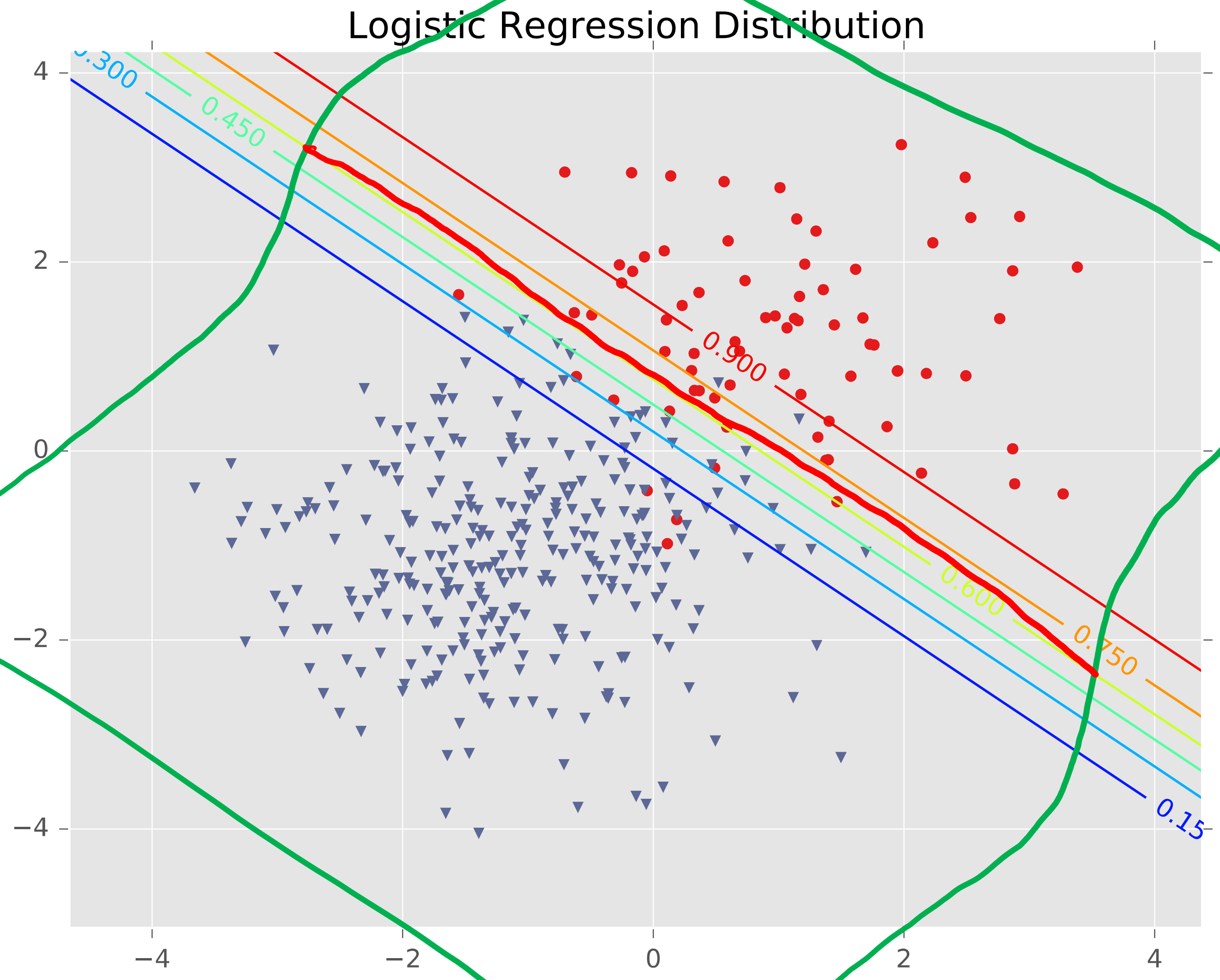
Logistic Regression



$$y = \theta^T x$$
$$= w^T x + b$$
$$= 0$$

Slide credit: CMU MLD Matt Gormley

Logistic Regression



$g(w^T x + b)$

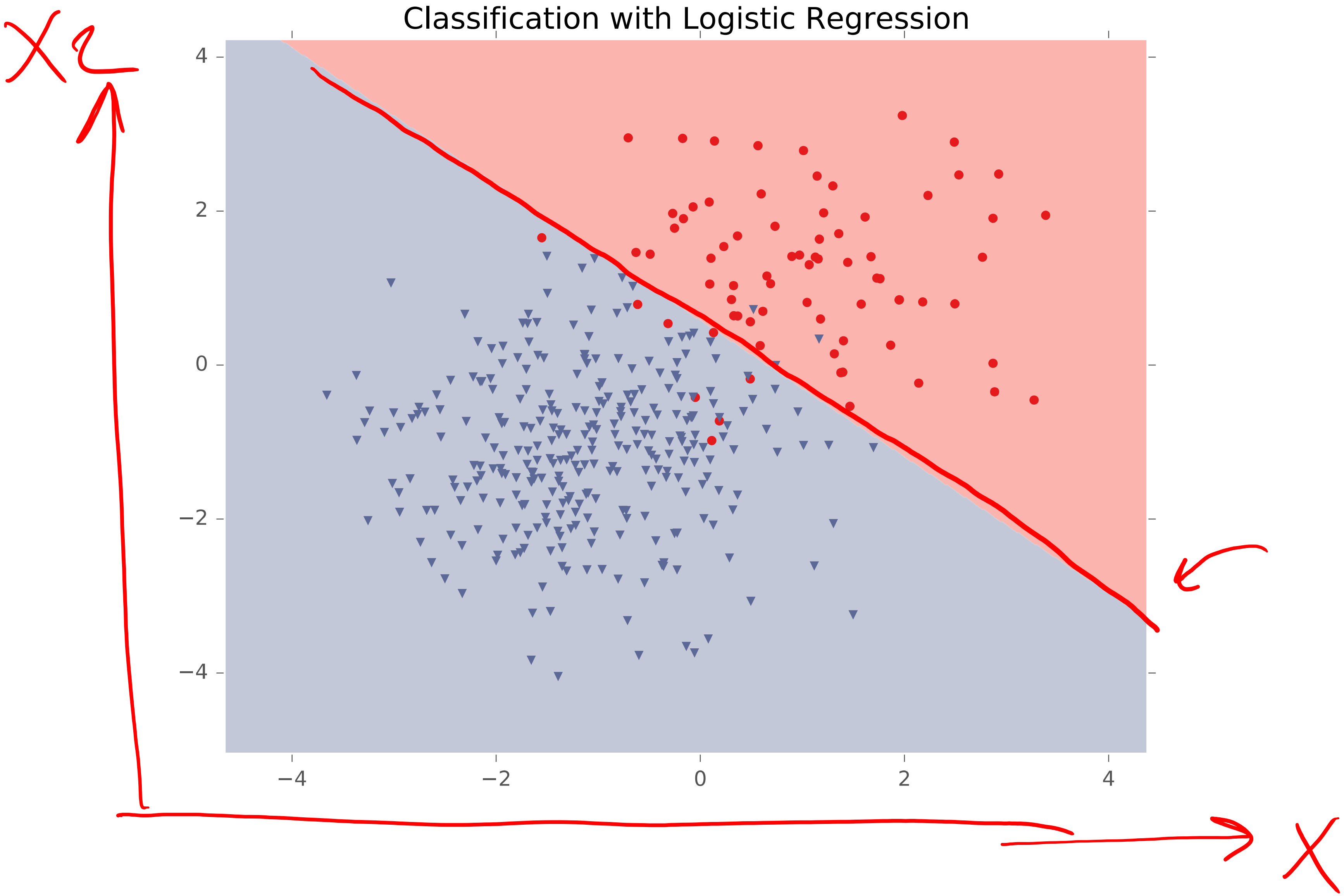
$\frac{1}{1+e^{-\theta^T x}}$

$p(Y=1|x,\theta)$

$= p(Y=0|x,\theta)$

$= 0.5$

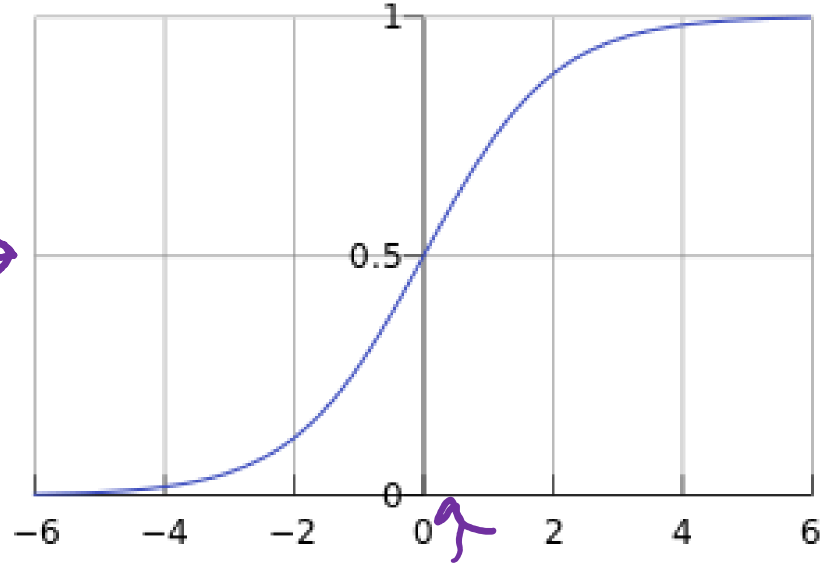
Logistic Regression



Poll 1

For a point $\boldsymbol{x}$ on the decision boundary of logistic regression, does $\underbrace{g\left(\mathbf{w}^T\mathbf{x}+b\right)}_{0.5} = \underbrace{\mathbf{w}^T\mathbf{x}+b}_{0}$?

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

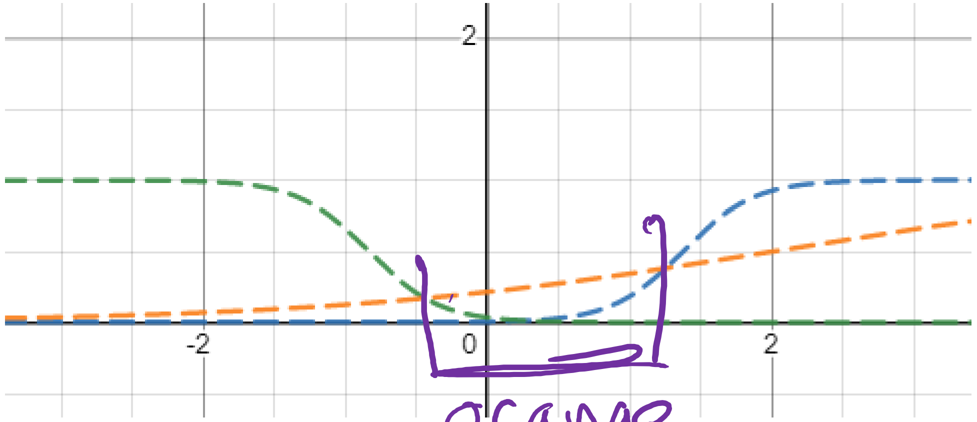


Multi-class Logistic Regression

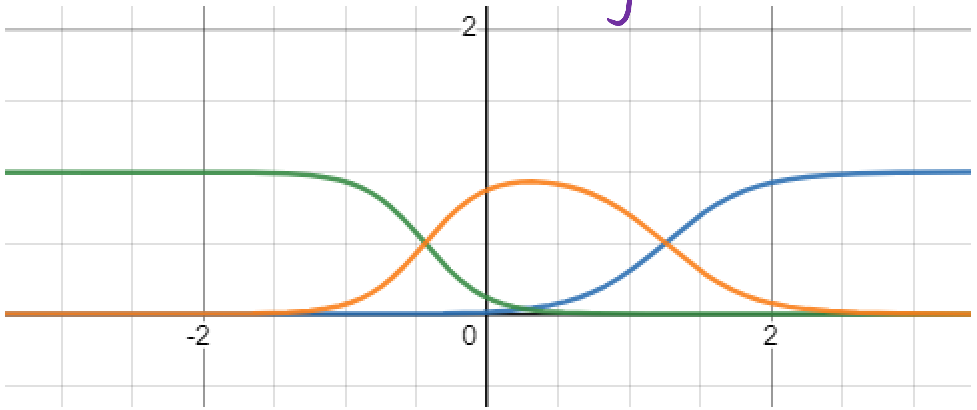
Multi-class Logistic Regression

Desmos Demo:

https://www.desmos.com/calculator/53bautbxjp



orange



$$g_\ell(\theta_1^T x)$$
$$g_\ell(\theta_2^T x)$$
$$g_\ell(\theta_3^T x)$$

$$\hat{y} = g_{sm}\left(\begin{bmatrix}\theta_1^T x \\ \theta_2^T x \\ \theta_3^T x\end{bmatrix}\right)$$

Multi-class Logistic Regression

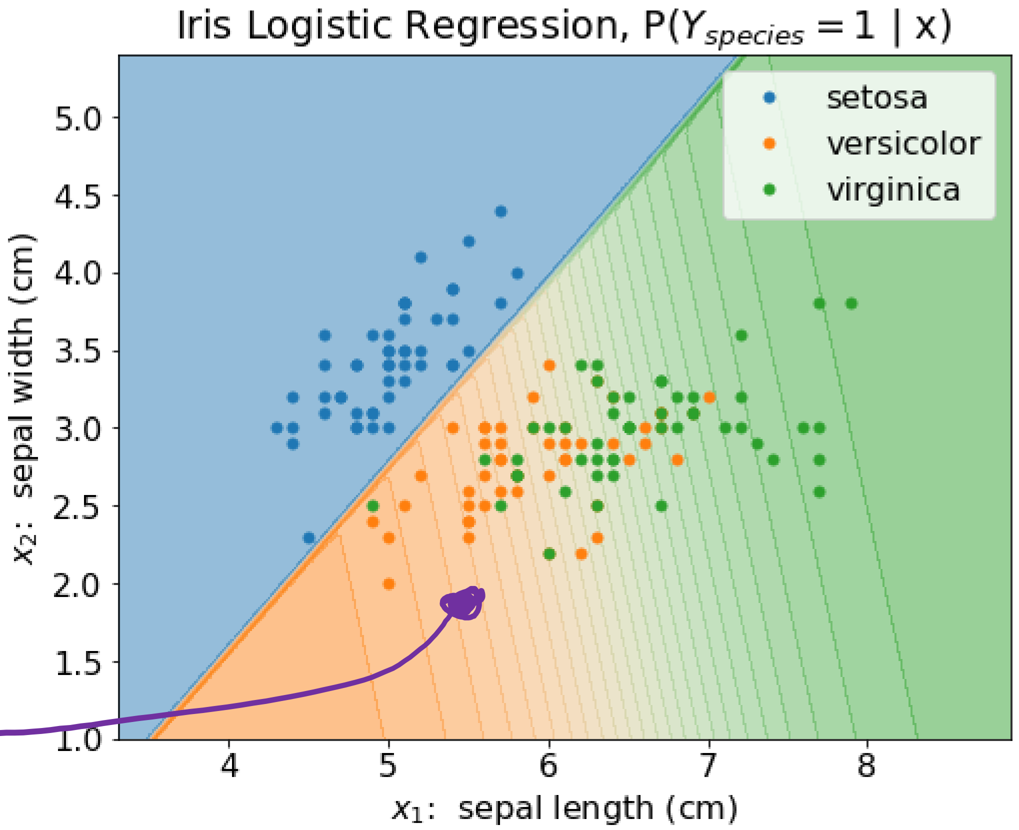
Cross-entropy loss

$$\ell(\boldsymbol{y}, \hat{\mathbf{y}}) = -\sum_{k=1}^{K} y_k \log \hat{y}_k$$

Model

$$\hat{y} = g_{sm}\left(\quad \right)$$

$$\hat{y} = \begin{bmatrix} 0 \\ .8 \\ .2 \end{bmatrix}$$



Logistic Function

Logistic (sigmoid) function converts value from $(-\infty, \infty) \rightarrow (0, 1)$

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

$g(z)$ and $1 - g(z)$ sum to one

Example 2 $\rightarrow$ $g(2) = 0.88$, 1-$g(2) = 0.12$

Softmax Function

Softmax function convert each value in a vector of values from $(-\infty, \infty) \rightarrow (0, 1)$, such that they all sum to one.

$$g(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_K \end{bmatrix} \rightarrow \begin{bmatrix} e^{z_1} \\ e^{z_2} \\ \vdots \\ e^{z_K} \end{bmatrix} \cdot \frac{1}{\sum_{k=1}^{K} e^{z_k}}$$

Example $\begin{bmatrix} -1 \\ 4 \\ 1 \\ -2 \\ 3 \end{bmatrix} \rightarrow \begin{bmatrix} 0.0047 \\ 0.7008 \\ 0.0349 \\ 0.0017 \\ 0.2578 \end{bmatrix}$

Multiclass Predicted Probability

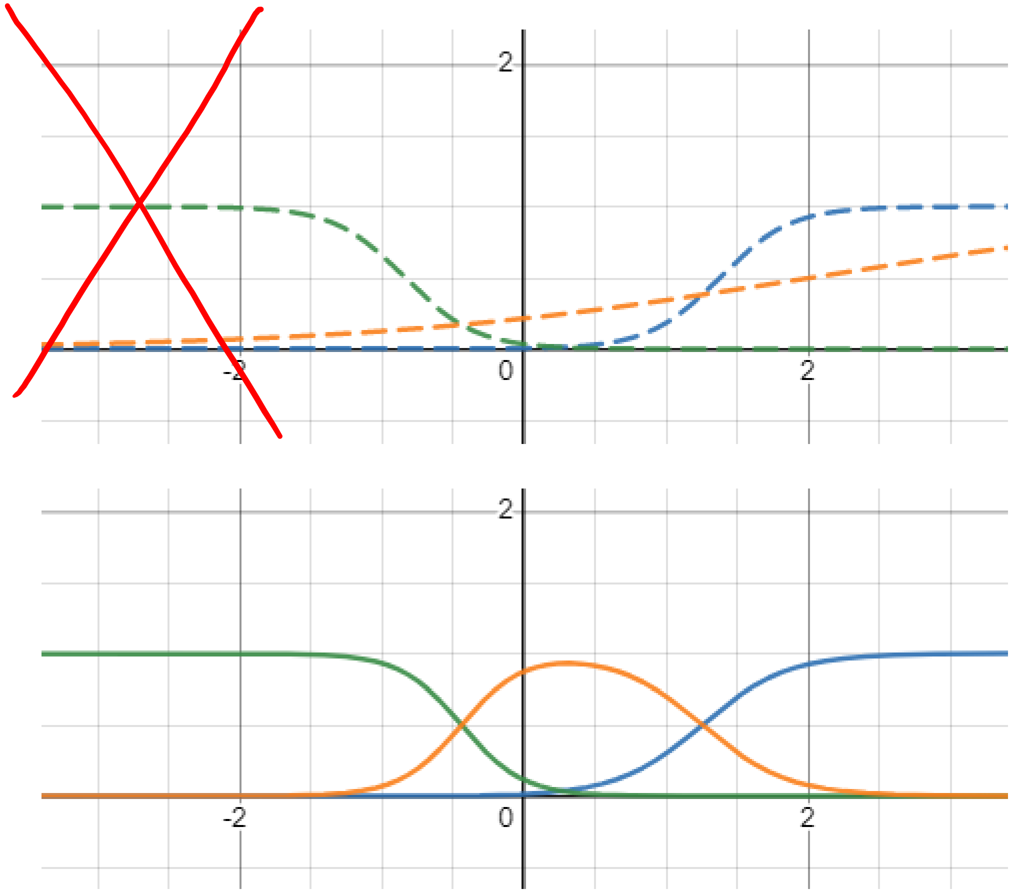
Multiclass logistic regression uses the parameters learned across all K classes to predict the discrete conditional probability distribution of the output Y given a specific input vector $\mathbf{x}$

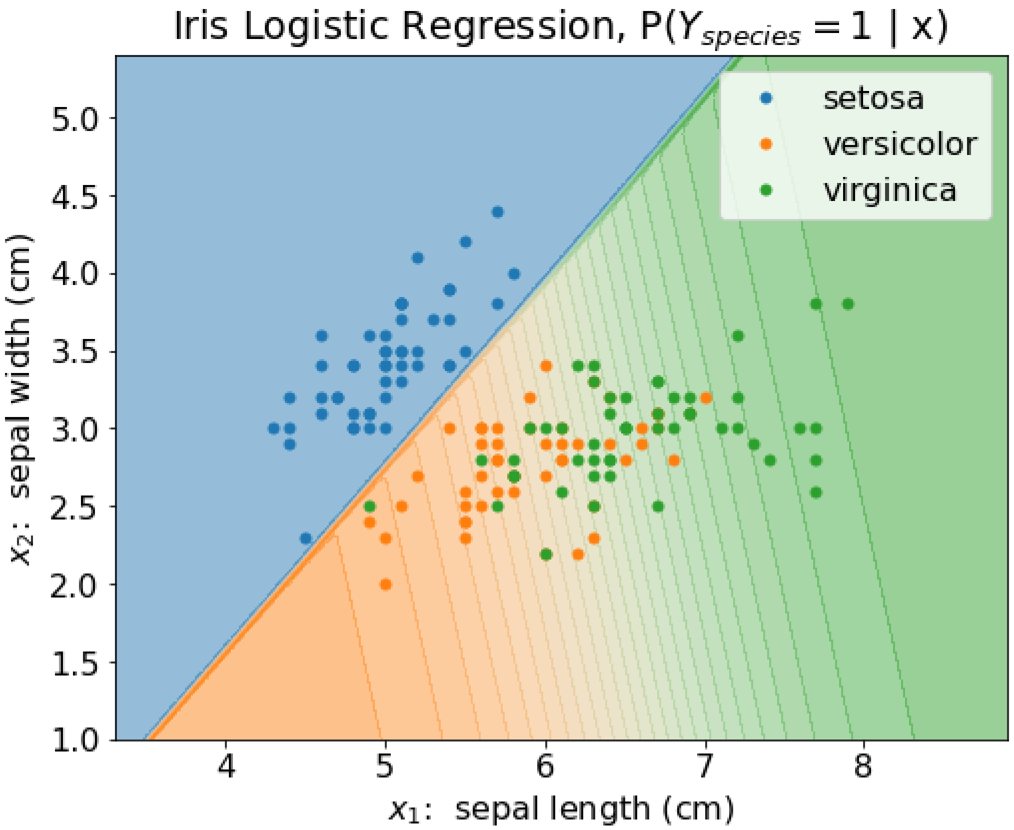
$$\begin{bmatrix} p(Y=1 \mid \mathbf{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) \\ p(Y=2 \mid \mathbf{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) \\ p(Y=3 \mid \mathbf{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) \end{bmatrix} = \begin{bmatrix} e^{\boldsymbol{\theta}_1^T \mathbf{x}} \\ e^{\boldsymbol{\theta}_2^T \mathbf{x}} \\ e^{\boldsymbol{\theta}_3^T \mathbf{x}} \end{bmatrix} \cdot \frac{1}{\sum_{k=1}^{K} e^{\boldsymbol{\theta}_k^T \mathbf{x}}}$$

each of these is a vector (confusing notation)

Multiclass Predicted Probability

Multiclass logistic regression uses the parameters learned across all K classes to predict the discrete conditional probability distribution of the output Y given a specific input vector $\mathbf{x}$

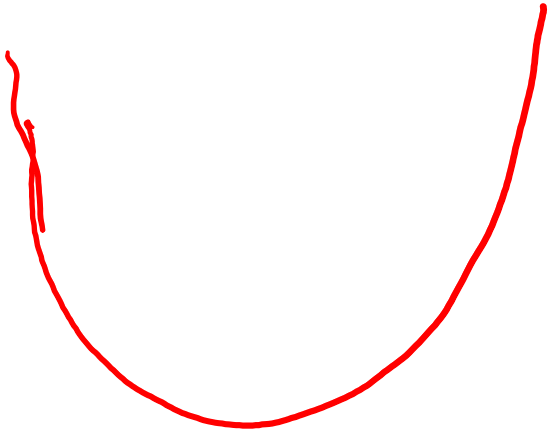




Logistic Regression with Polynomial Features

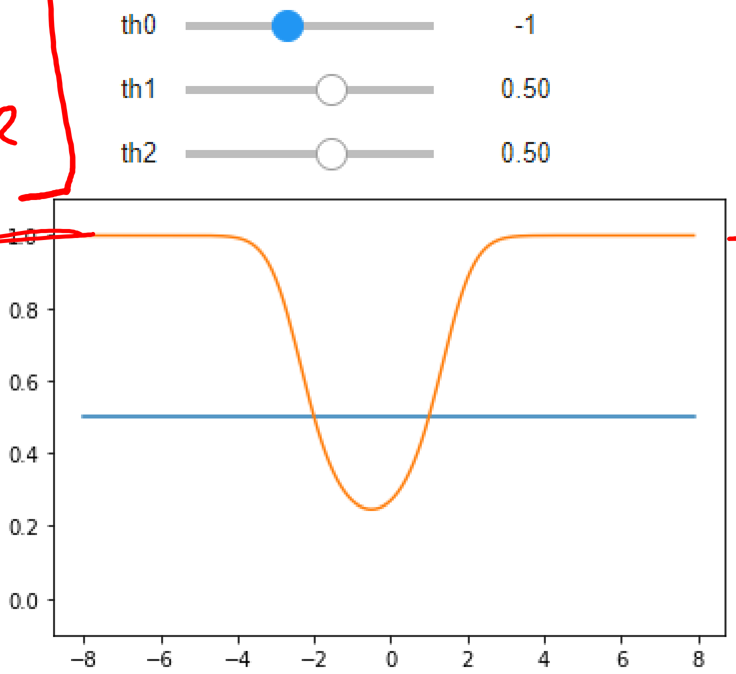
Exercise

Interact with the logistic_quadratic.ipynb posted on the course website schedule



$\vec{x}$ $\vec{\theta}$

$$\begin{bmatrix} 1 \\ x_1 \\ x_1^2 \end{bmatrix}$$



Exercise

Interact with the logistic_quadratic.ipynb posted on the course website schedule