10-315
Introduction to ML

Logistic Regression

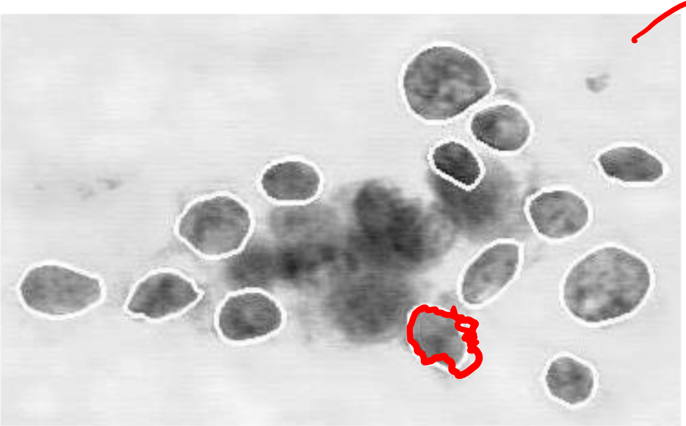
Instructor: Pat Virtue

Example: Breast cancer classification

Well-known classification example: using machine learning to diagnose whether a breast tumor is benign or malignant [Street et al., 1992]

Setting: doctor extracts a sample of fluid from tumor, stains cells, then outlines several of the cells (image processing refines outline)

feature eng.



$$\phi(x) \to \begin{bmatrix} \text{mean area} \\ \text{mean concavit} \\ \\ \end{bmatrix}$$

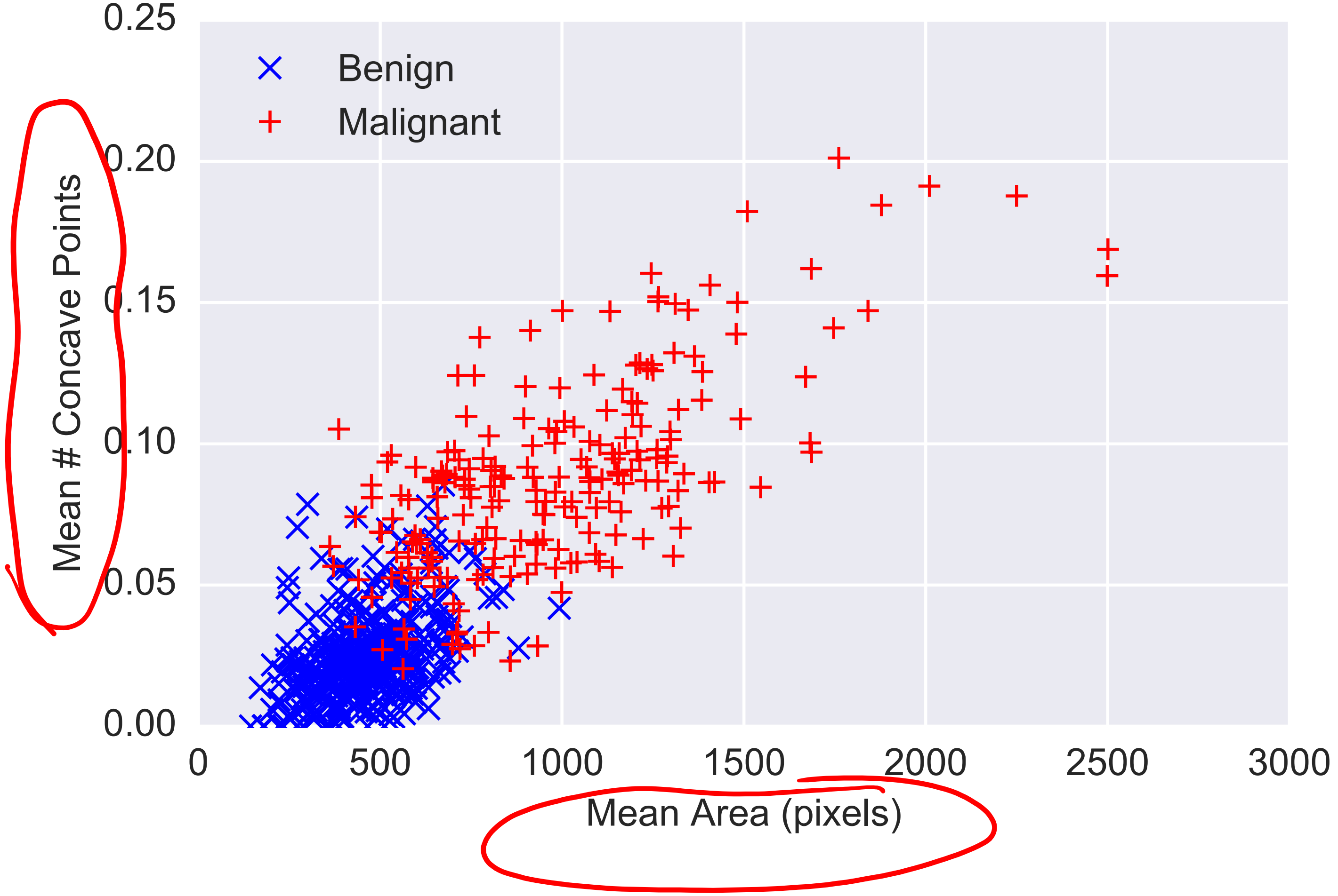
System computes features for each cell such as area, perimeter, concavity, texture (10 total); computes mean/std/max for all features

Slide credit: CMU AI Zico Kolter

Example: Breast cancer classification

Plot of two features: mean area vs. mean concave points, for two classes

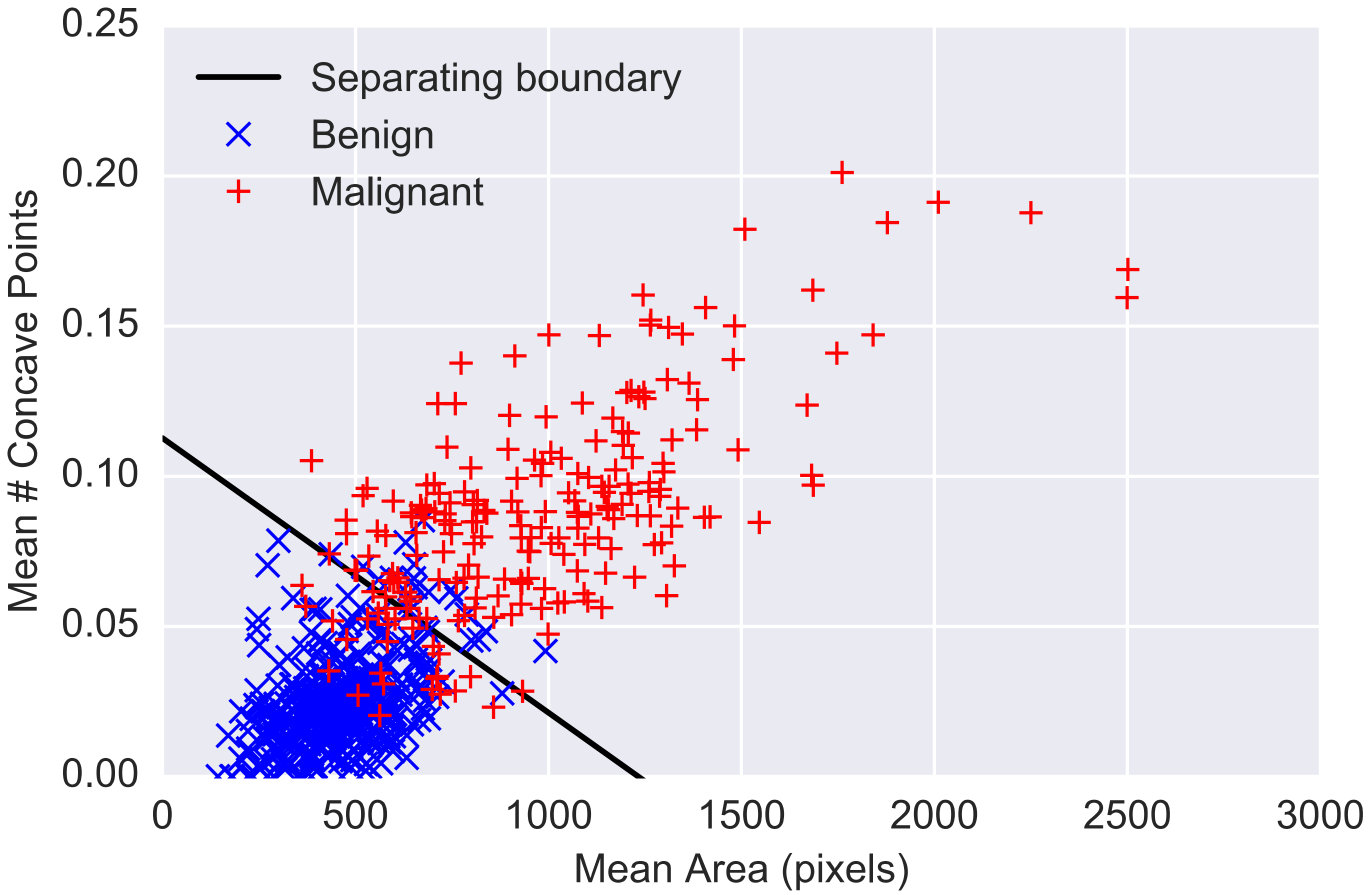
$\hat{y} = h(x_1, x_2)$



Slide credit: CMU AI Zico Kolter

Linear classification example

Linear classification: linear decision boundary

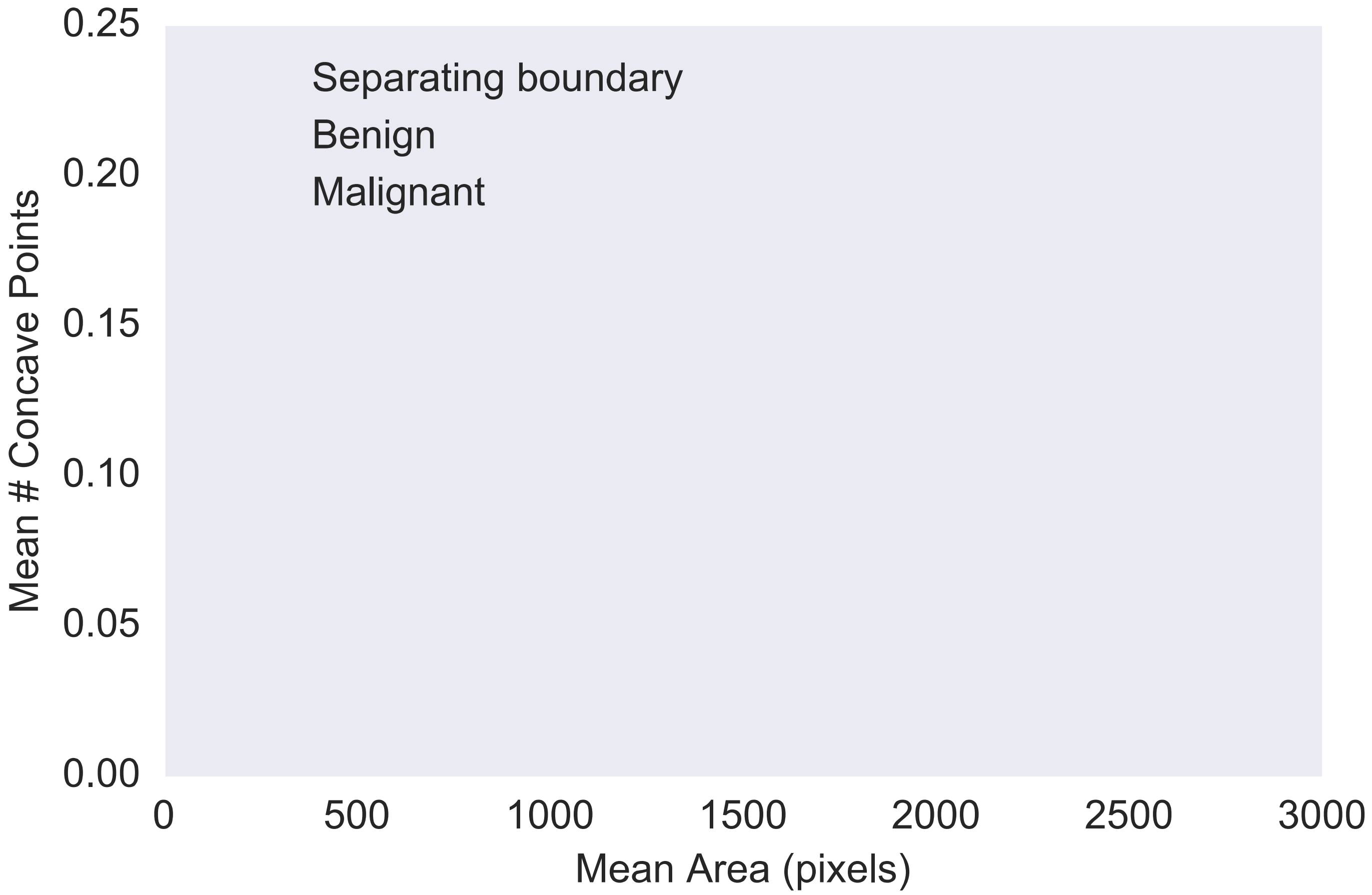


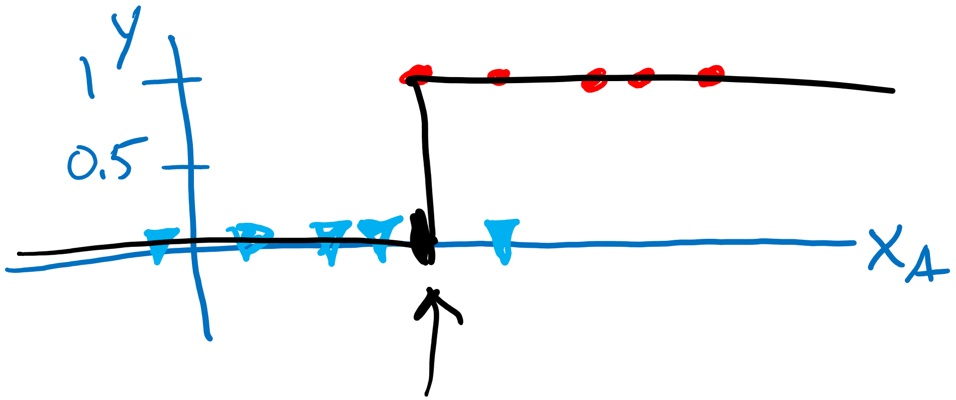
Slide credit: CMU AI Zico Kolter

Logistic regression for classification

Linear classification: linear decision boundary

Probabilistic classification: provide $P(Y = 1 \mid x)$ rather than just $\hat{y} \in \{0, 1\}$



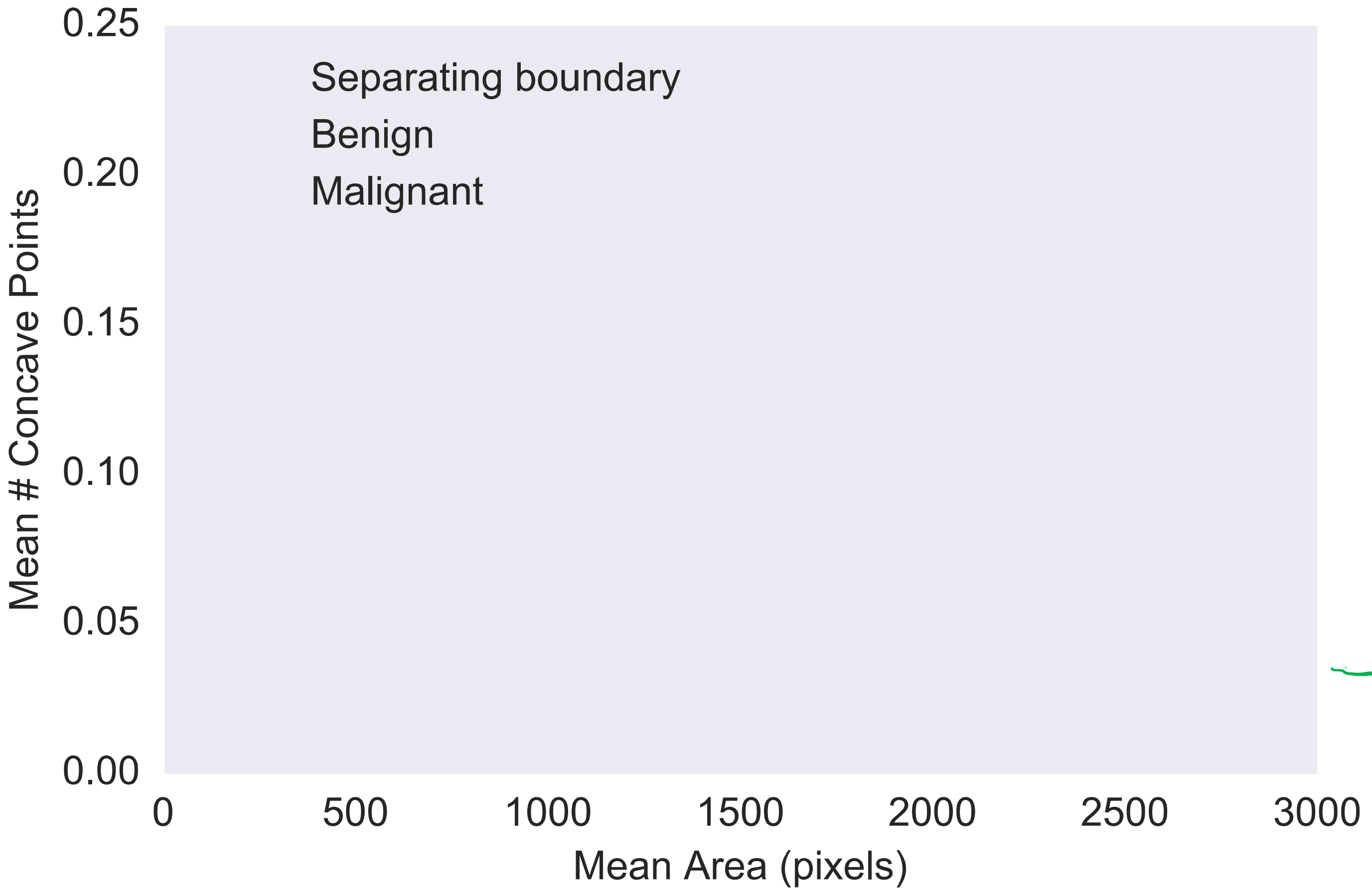


Slide credit: CMU AI Zico Kolter

Logistic regression for classification

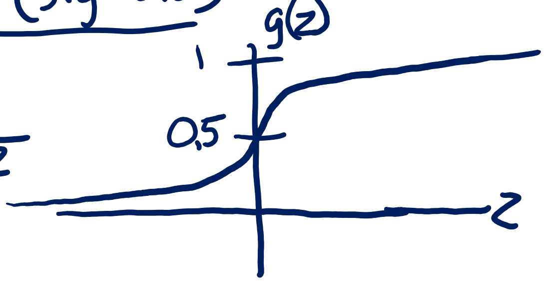
Linear classification: linear decision boundary

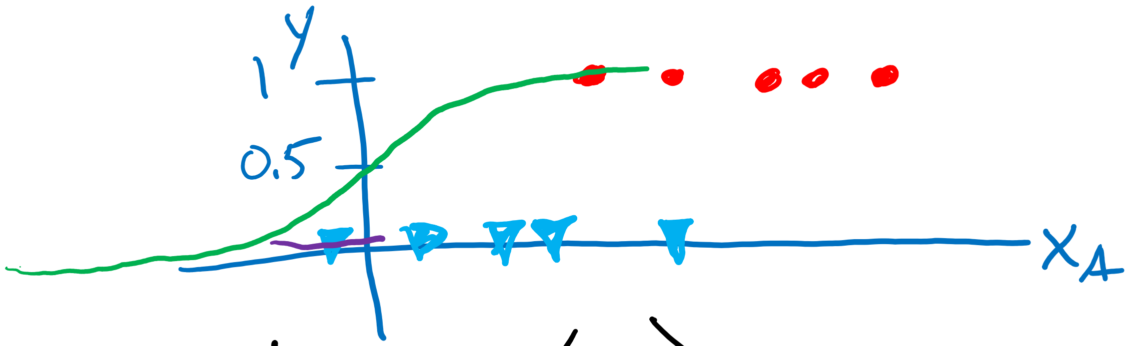
Probabilistic classification: provide $P(Y = 1 \mid x)$ rather than just $\hat{y} \in \{0, 1\}$



logistic function (sigmoid)

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





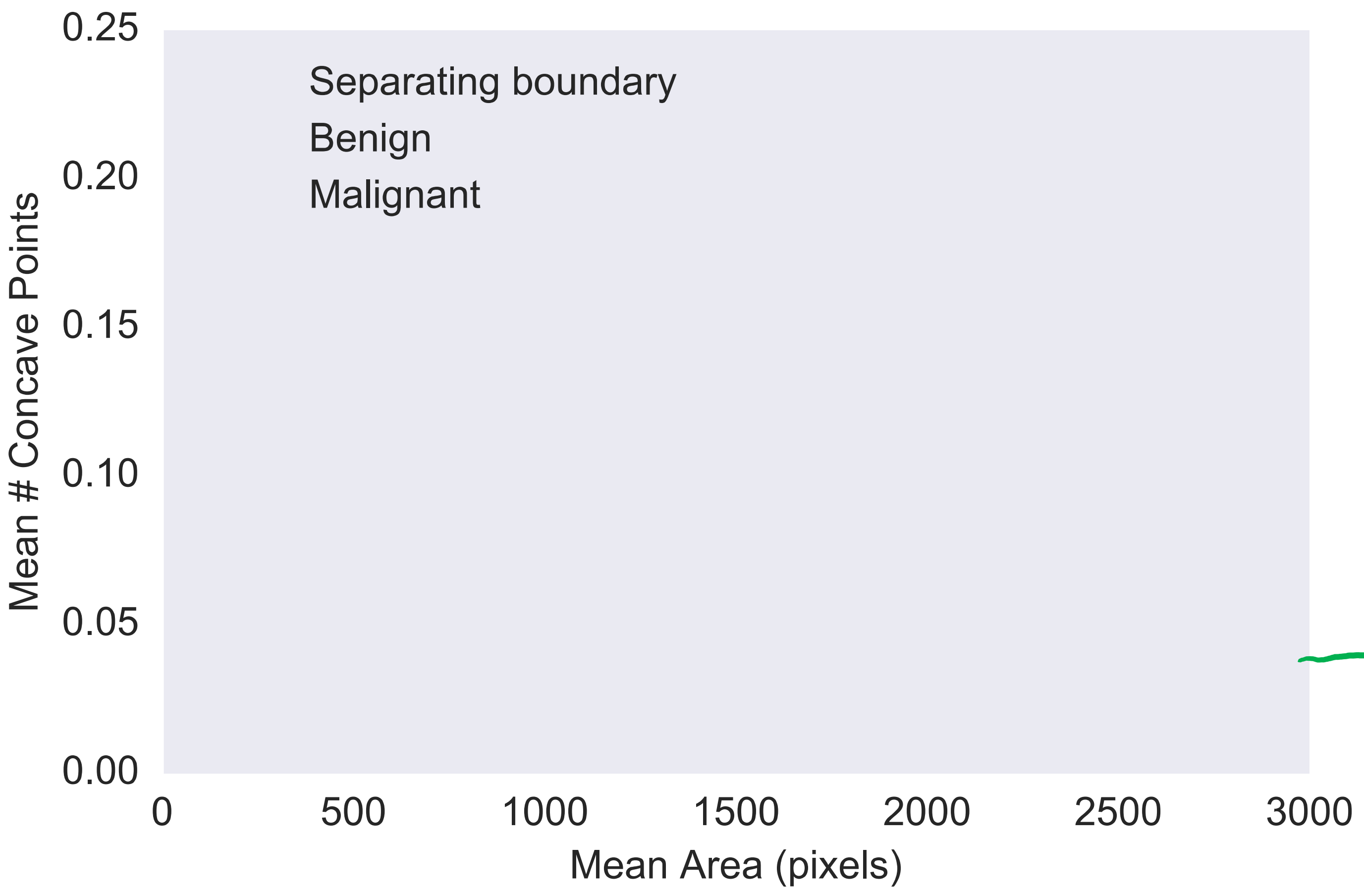
$$\hat{y} = g(x_A)$$

Slide credit: CMU AI Zico Kolter

Logistic regression for classification

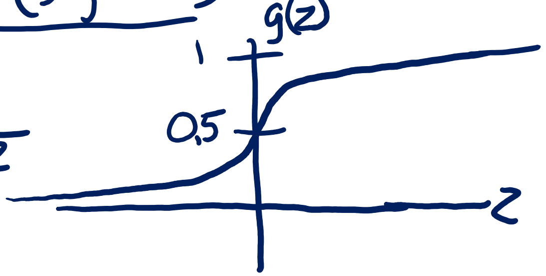
Linear classification: linear decision boundary

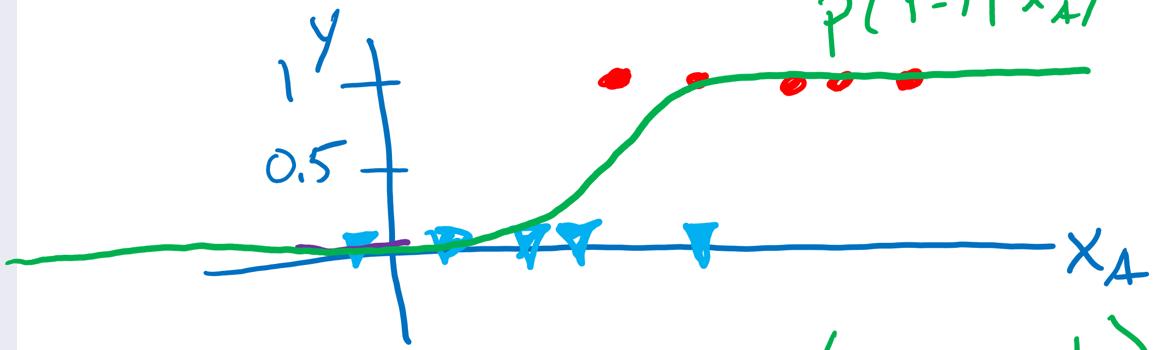
Probabilistic classification: provide $P(Y = 1 \mid x)$ rather than just $\hat{y} \in \{0, 1\}$



logistic function (sigmoid)

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





$$P(Y=1 \mid x_A) = g(w x_A + b)$$

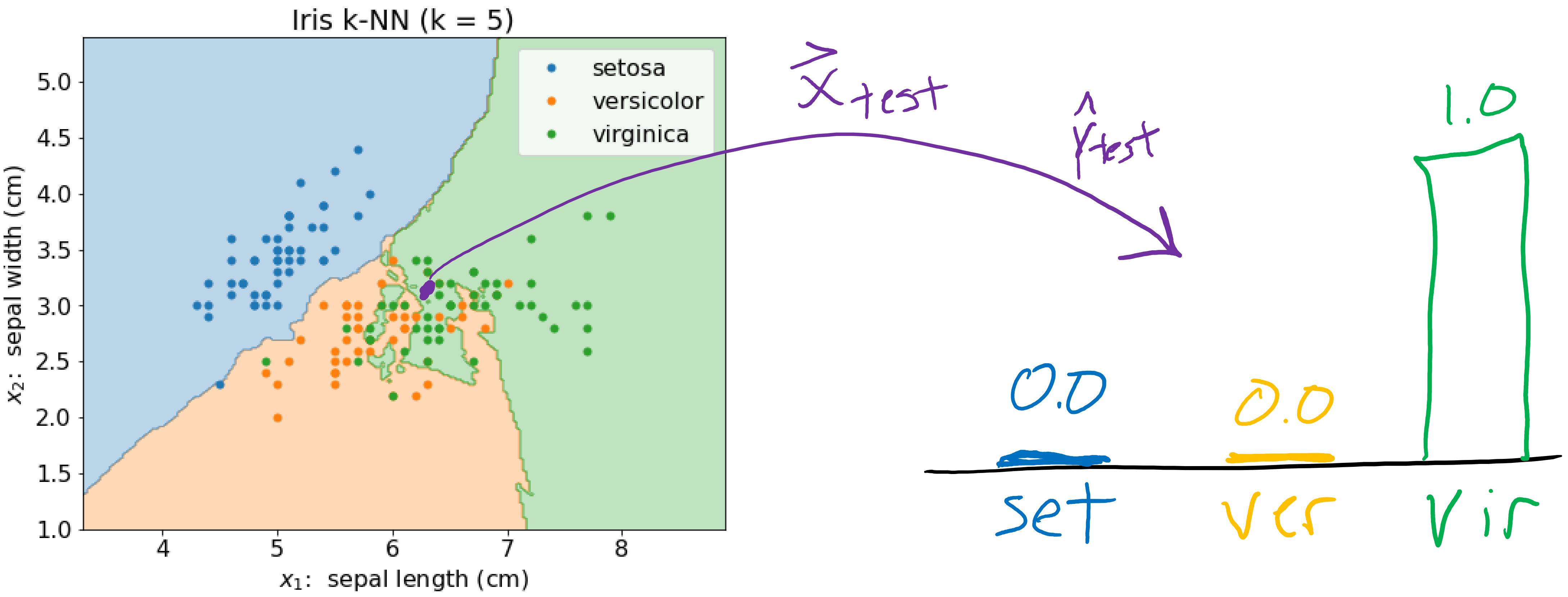
Slide credit: CMU AI Zico Kolter

Pre-reading

Cross-entropy loss

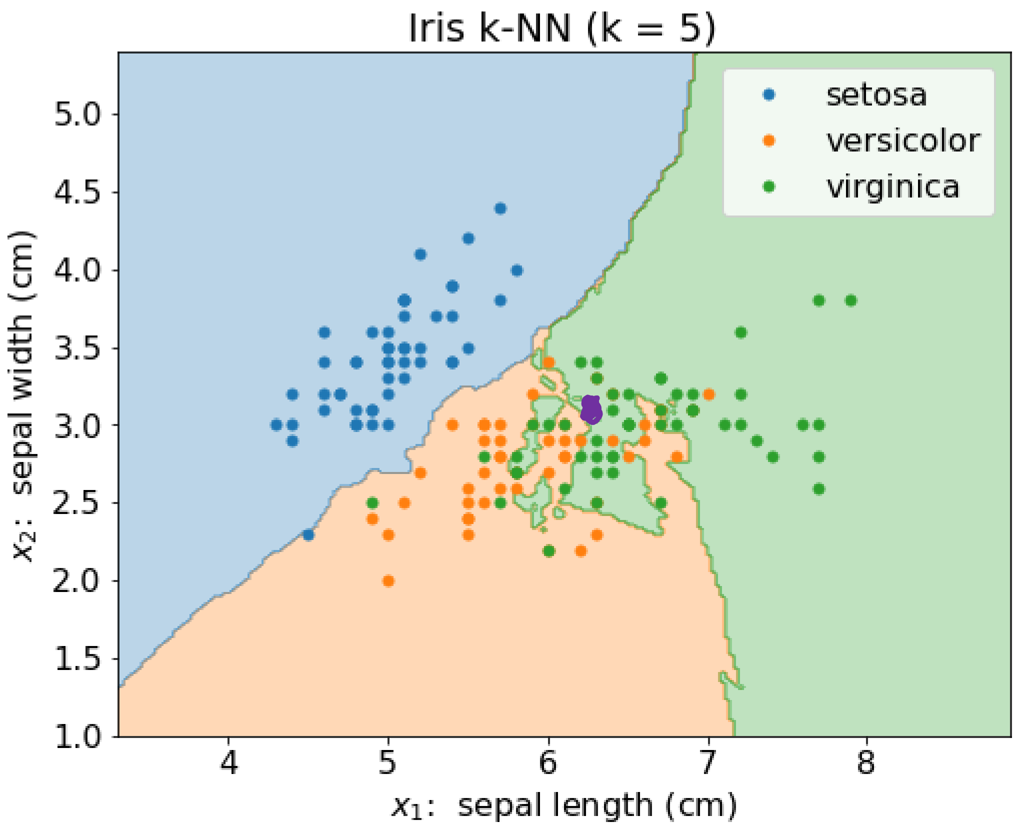
Classification Decisions

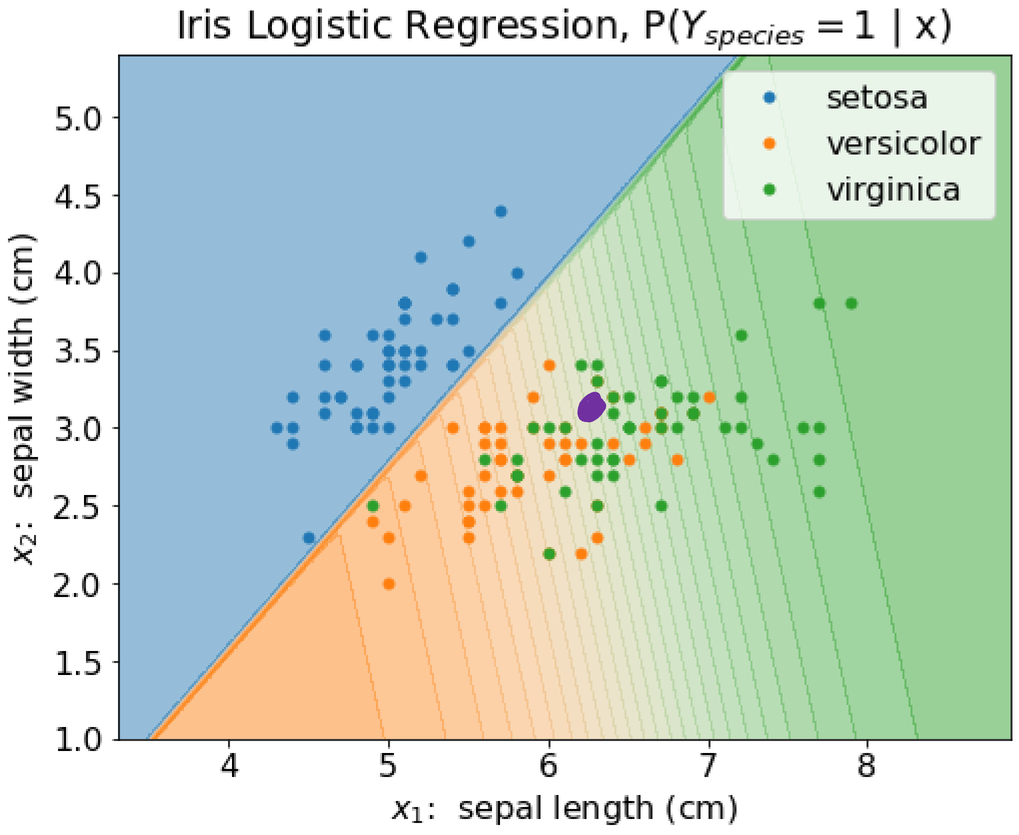
Predicting one specific class is troubling, especially when we know that there is some uncertainty in our prediction



Classification Probability

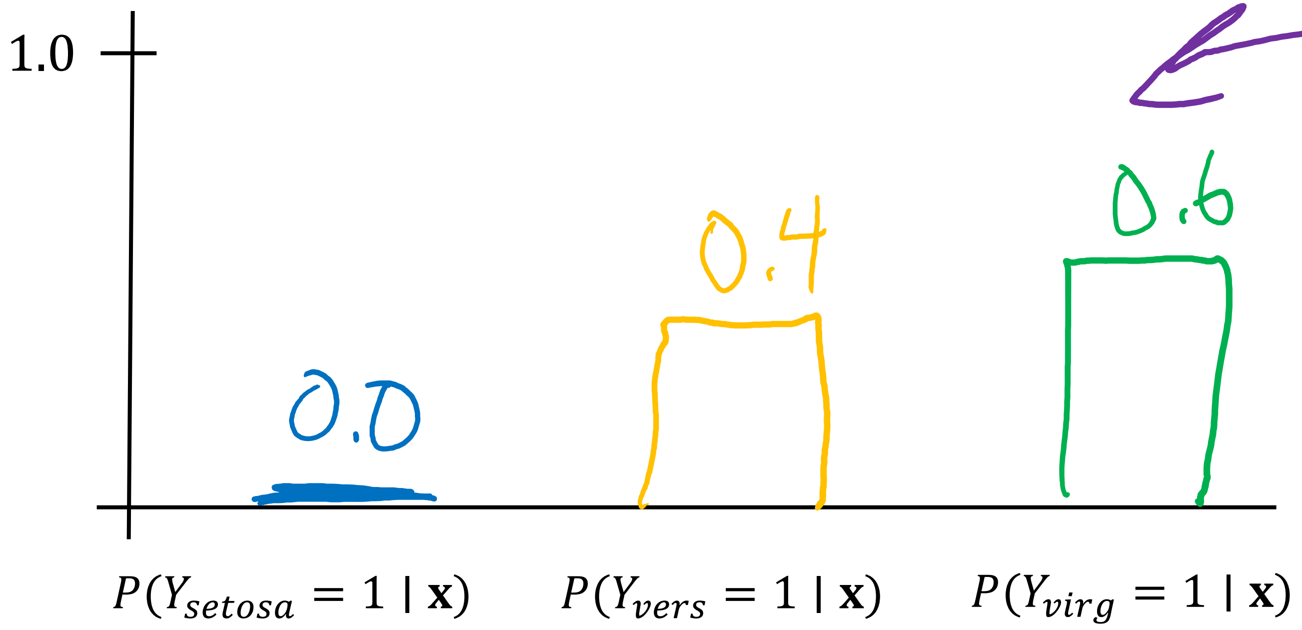
Constructing a model than can return the probability of the output being a specific class could be incredibly useful

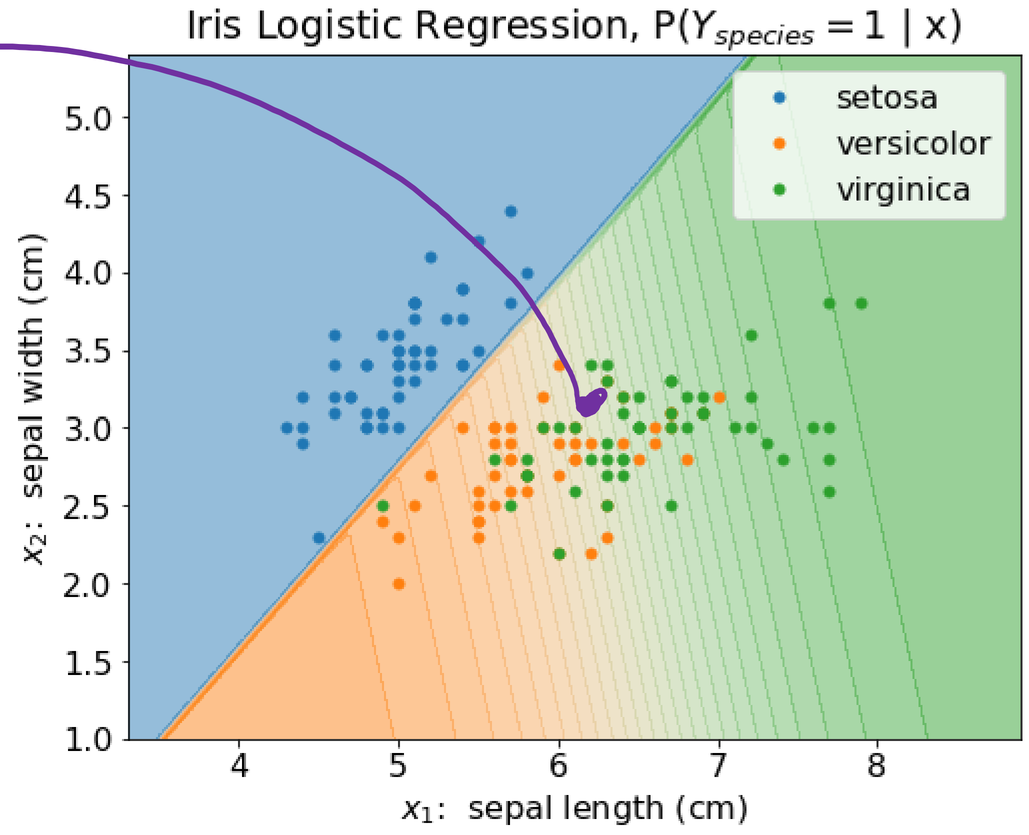




Classification Probability

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





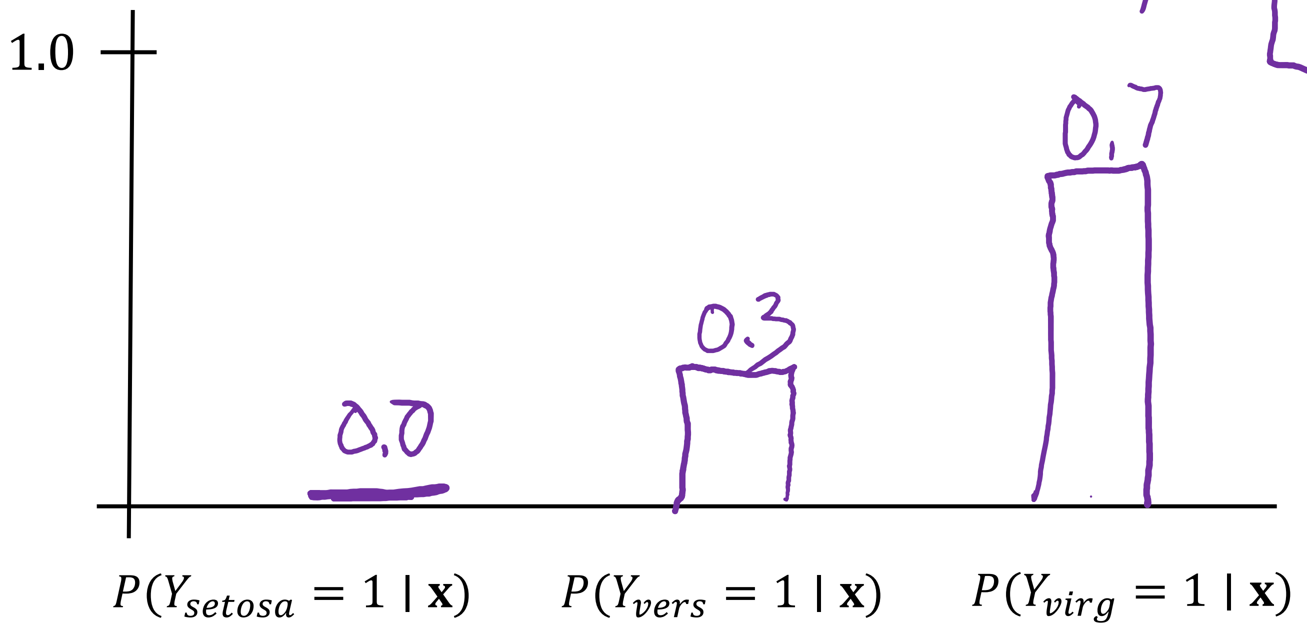
We can still make decisions, .e.g,

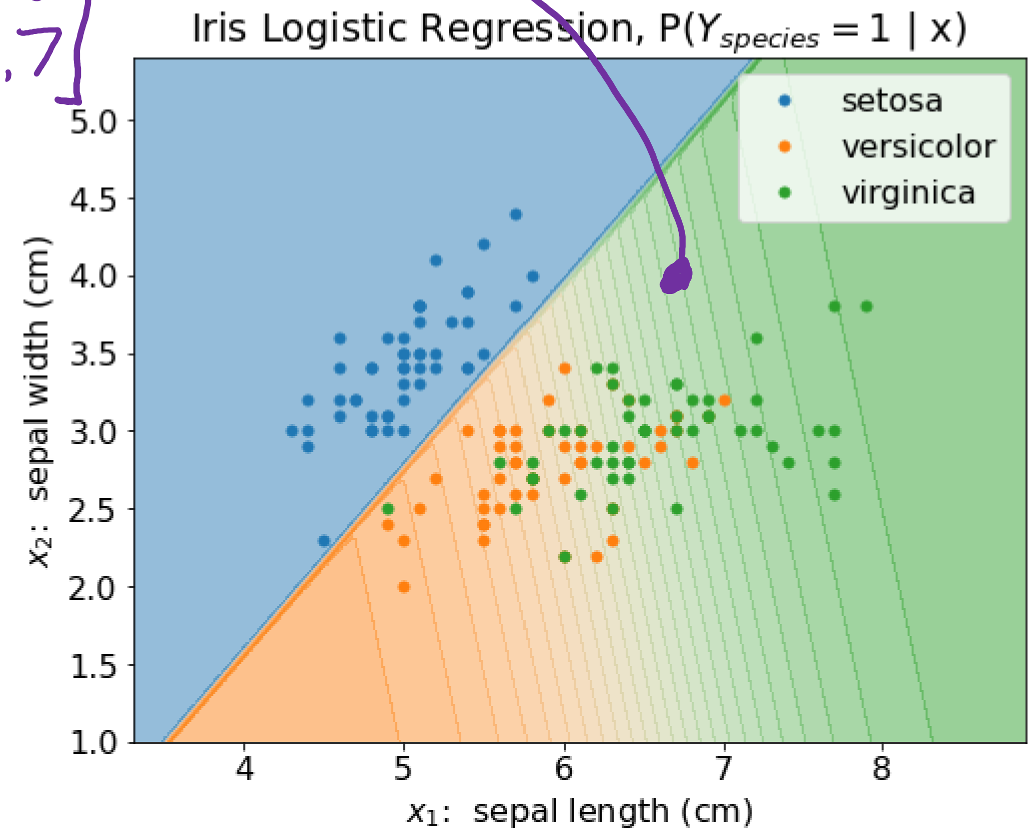
$$\underset{k}{\operatorname{argmax}}\, P(Y_k = 1 \mid \mathbf{x})$$

Loss for Probabilty Disributions

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

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





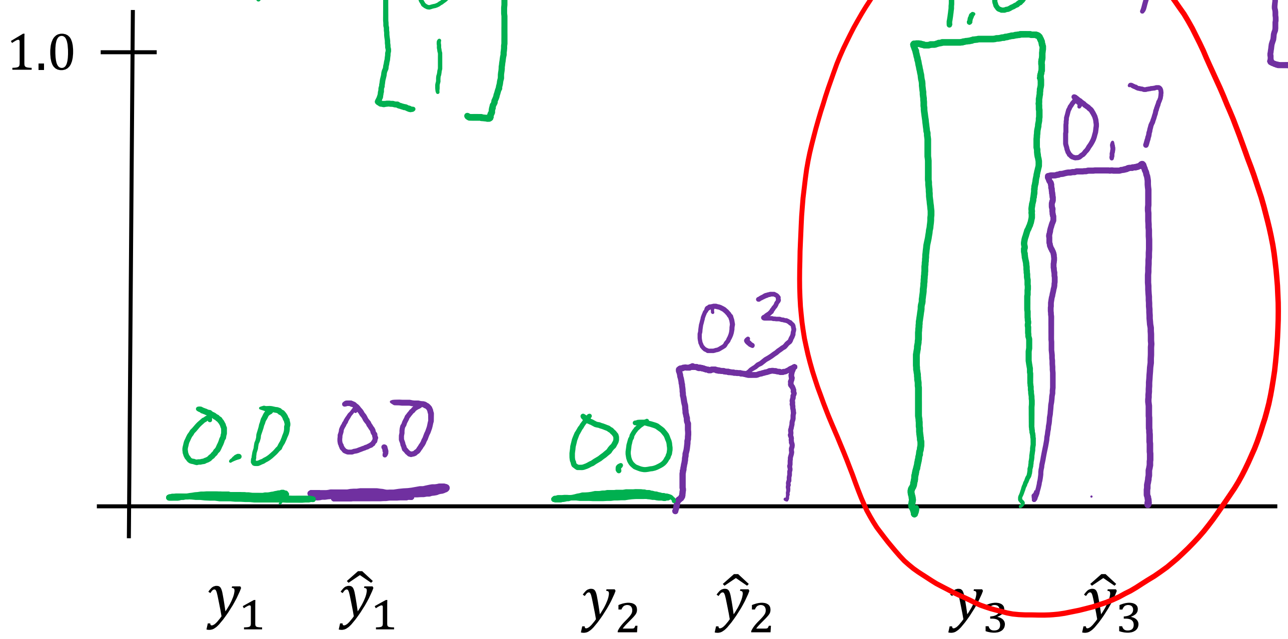
Cross-entropy loss

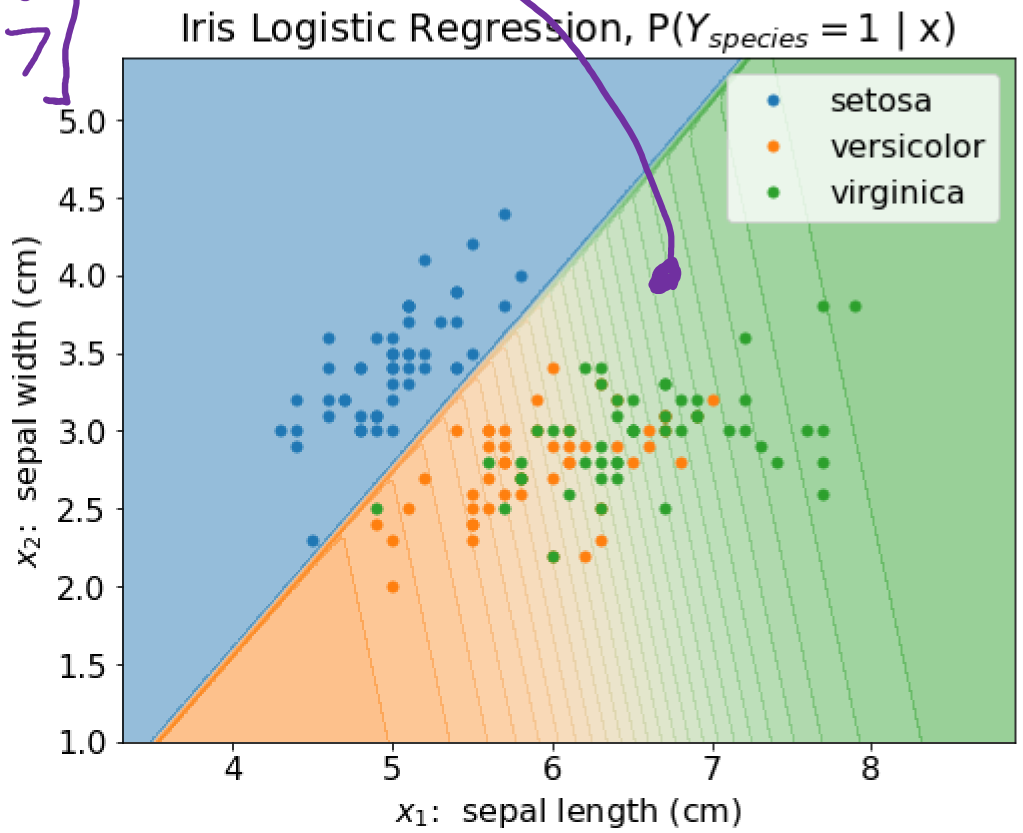
$$\ell(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{k=1}^{K} y_k \log \hat{y}_k$$

Loss for Probabilty Disributions

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

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





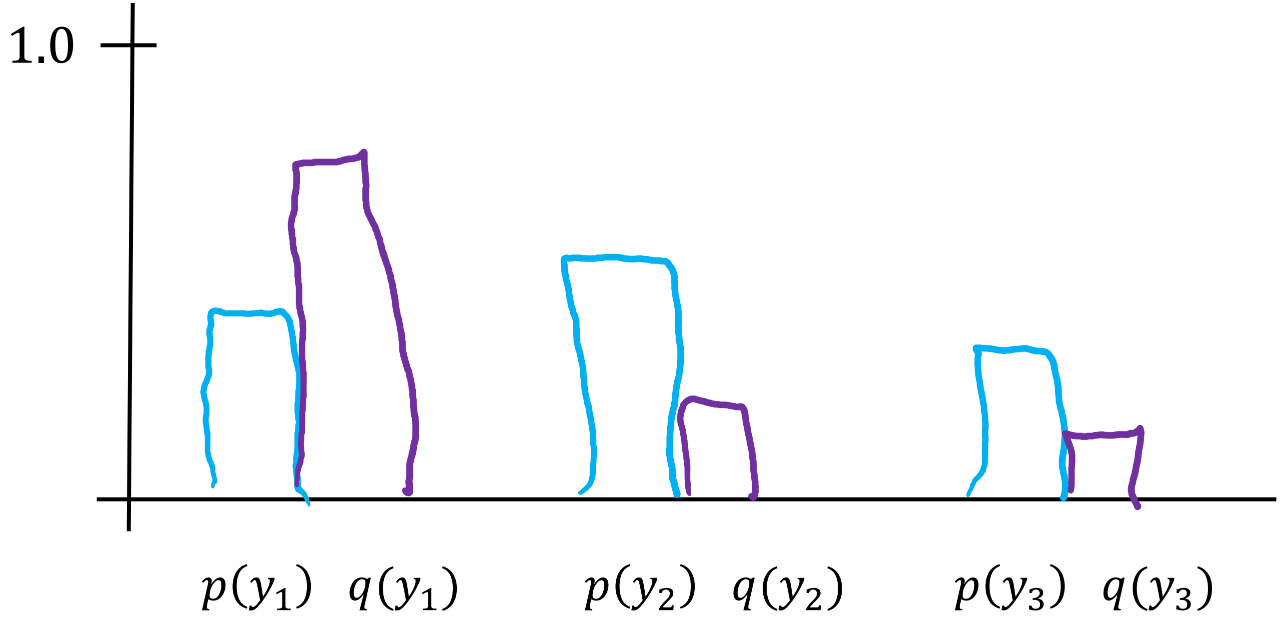
Cross-entropy loss

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

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

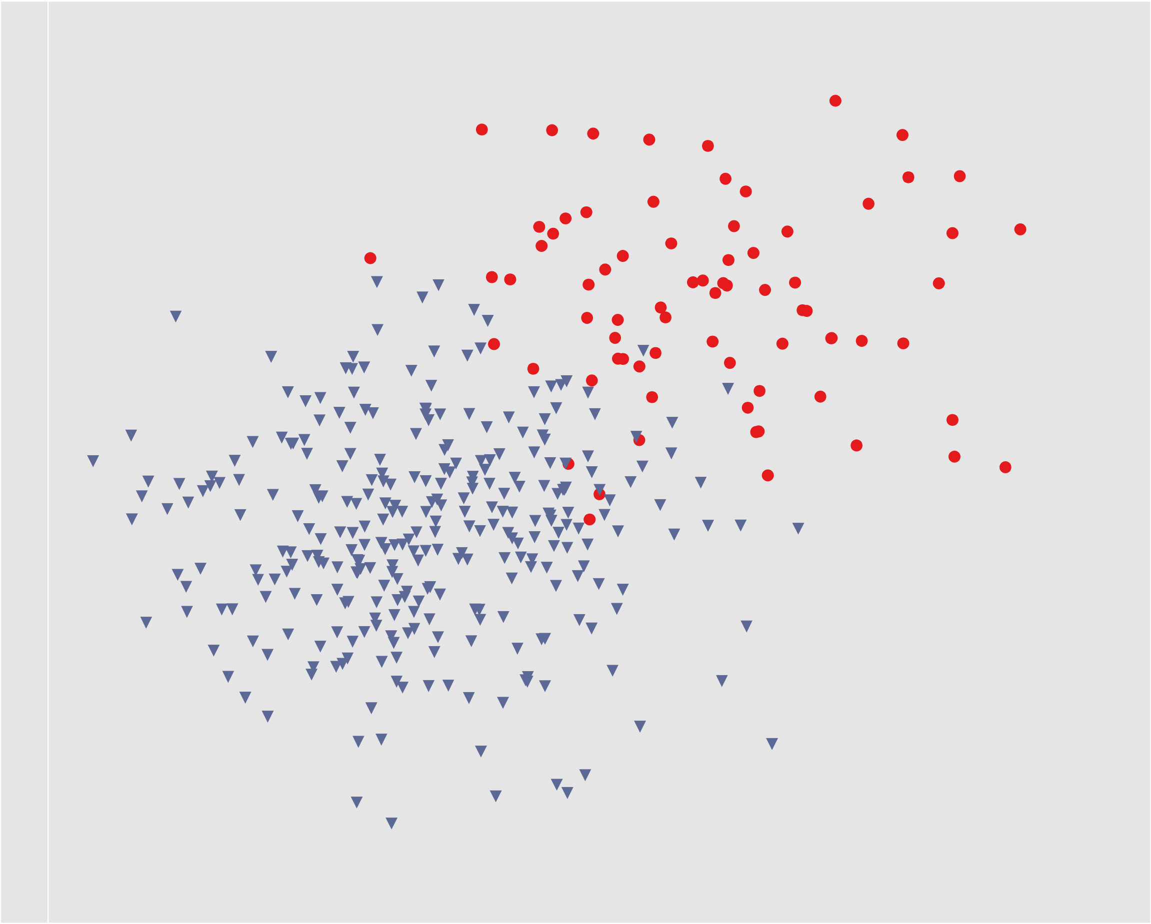
$$H(P, Q) = -\sum_{k=1}^{K} p(y_k) \log q(y_k)$$

Pre-reading

Linear model for classification

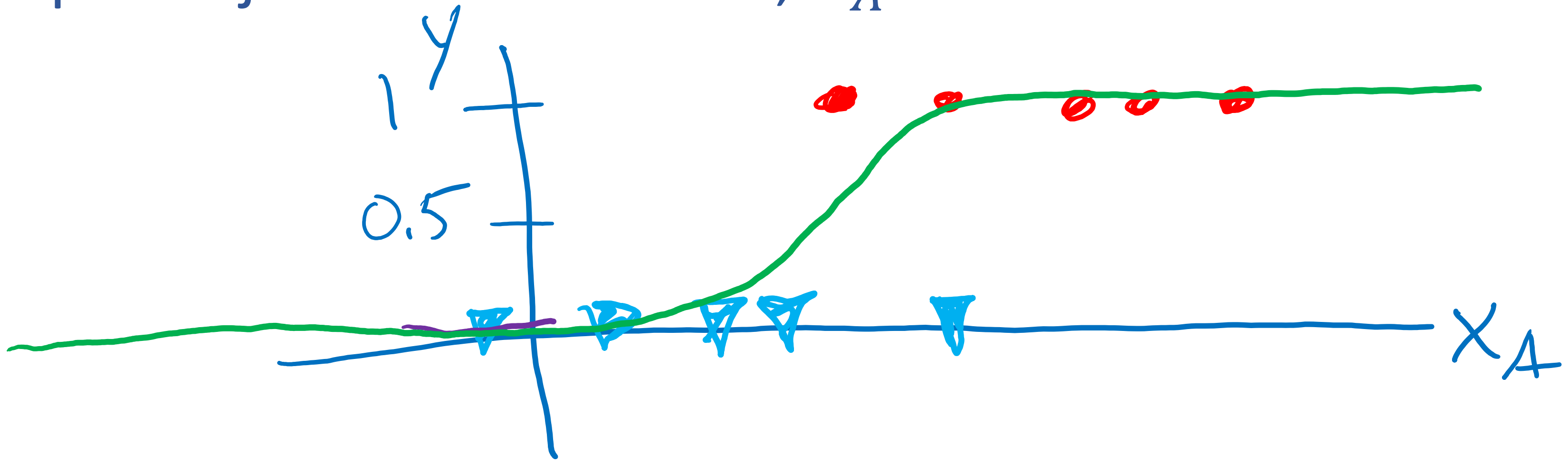
Prediction for Cancer Diagnosis

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

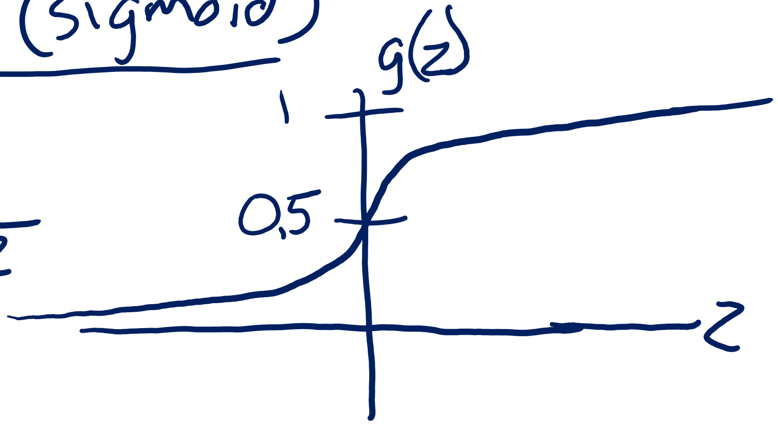
Learn to predict if a patient has cancer ($Y = 1$) or not ($Y = 0$) given the input of just one test result, X_A.



$p(Y=1 \mid x_A)$

logistic function (sigmoid)

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



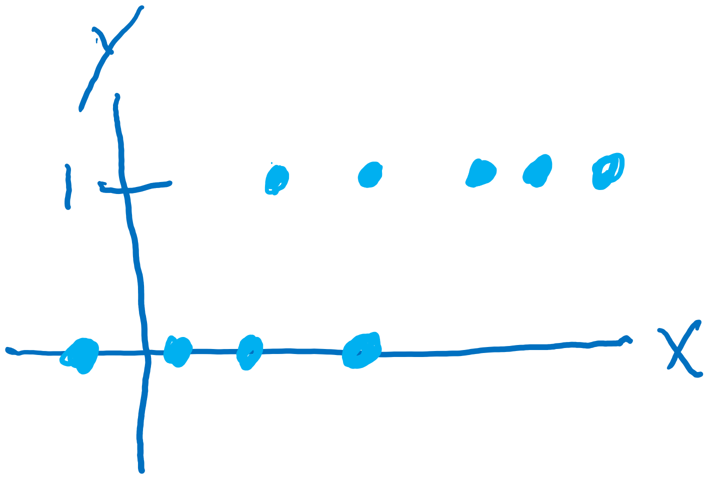
logistic regression

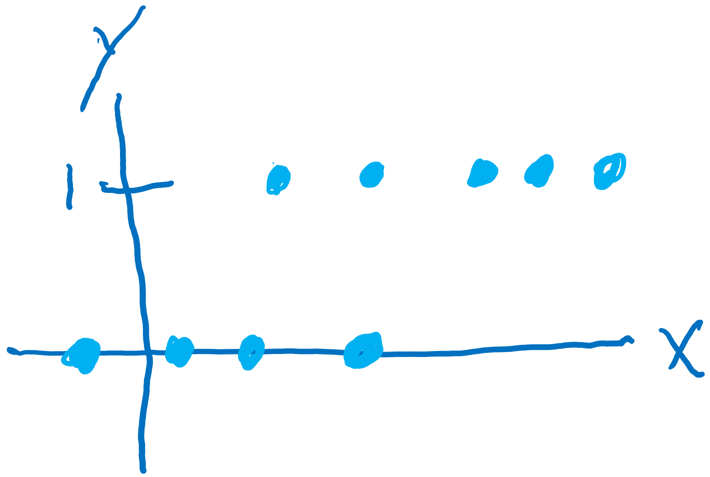
$$p(Y=1 \mid \vec{x}, \vec{\theta}) = g(\theta^T x)$$

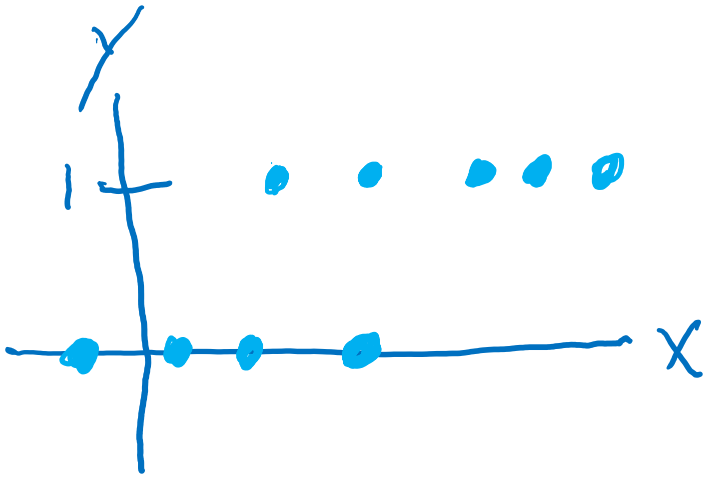
$$g(wx+b)$$

Building on a Linear Model

Linear vs Thresholded Linear vs Logistic Linear

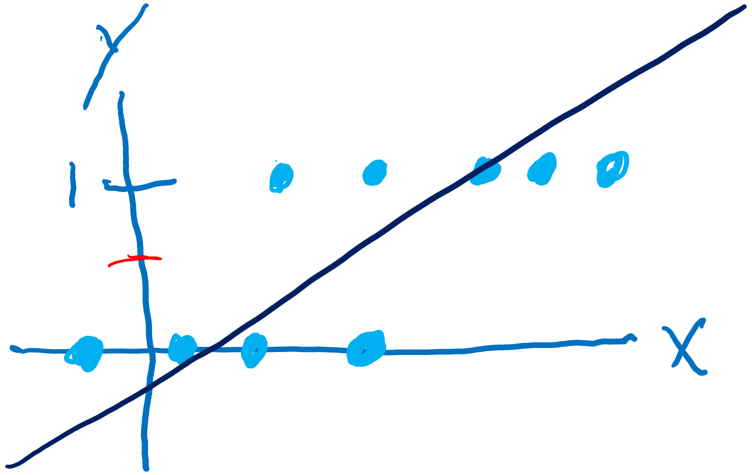






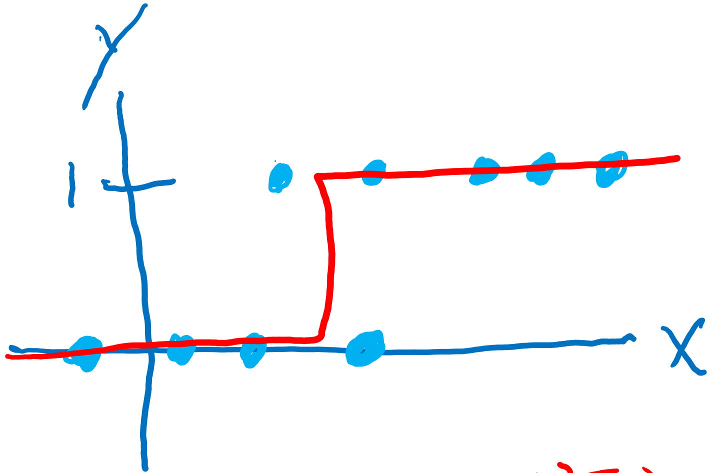
Building on a Linear Model

Linear vs Thresholded Linear vs Logistic Linear



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

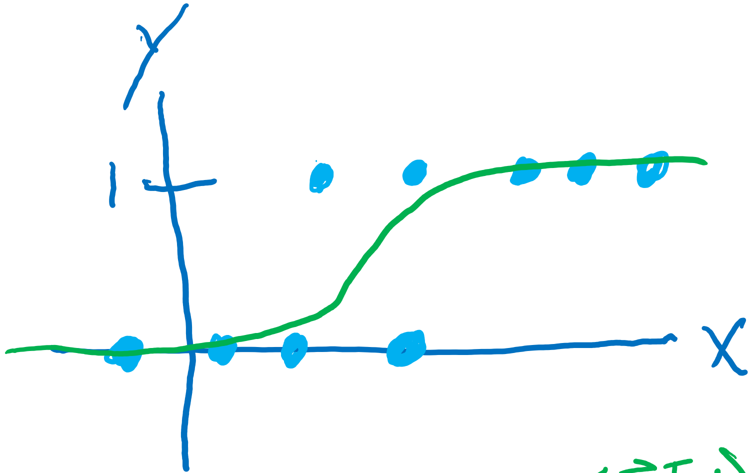
☹ not classification



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

☹ classification only (0/1)

☹ zero derivatives



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

Logistic Regression

Linear model for classification

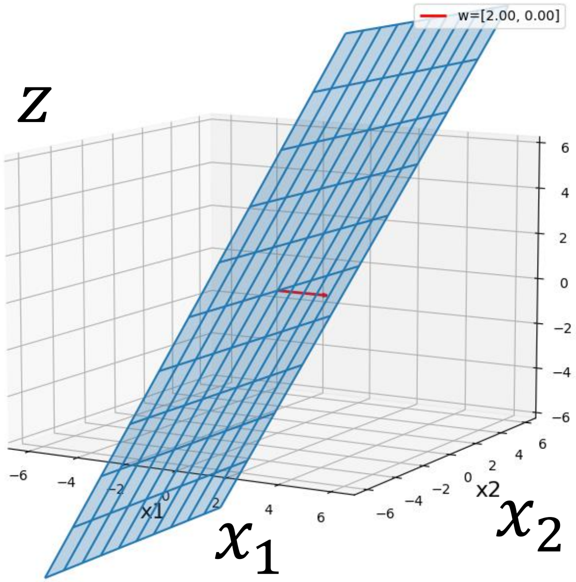
(now with 2+ input features)

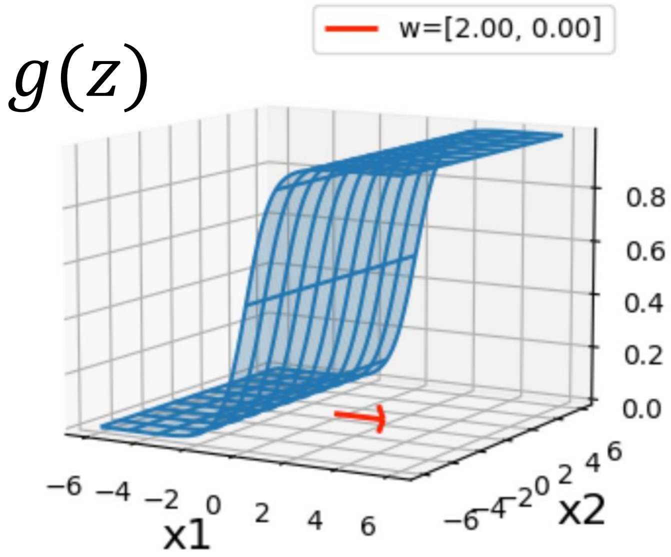
Building on a Linear Model

With two input features, $\mathbf{x} = [x_1 \;\; x_2]^\top$, we have two weight parameters and one bias parameter, $\mathbf{w} = [w_1 \;\; w_2]^\top$ and b, that control the slope and vertical offset of the following plane:

$$z = \mathbf{w}^\top \mathbf{x} + b$$

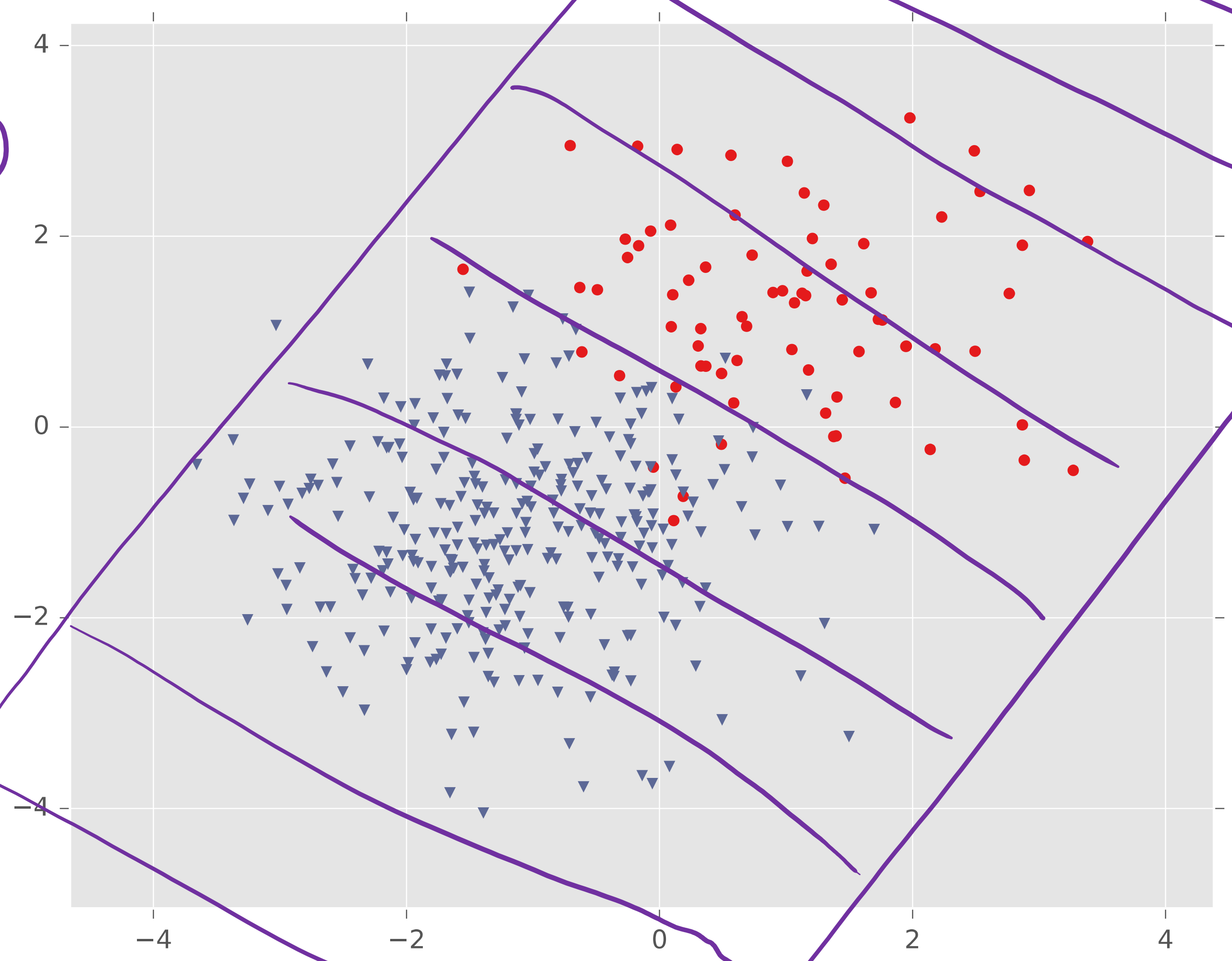
The sigmoid function $\hat{y} = g(z)$ then squashed the plane such that any z values going to $+\infty$ go to 1 and z values going to $-\infty$ go to -1





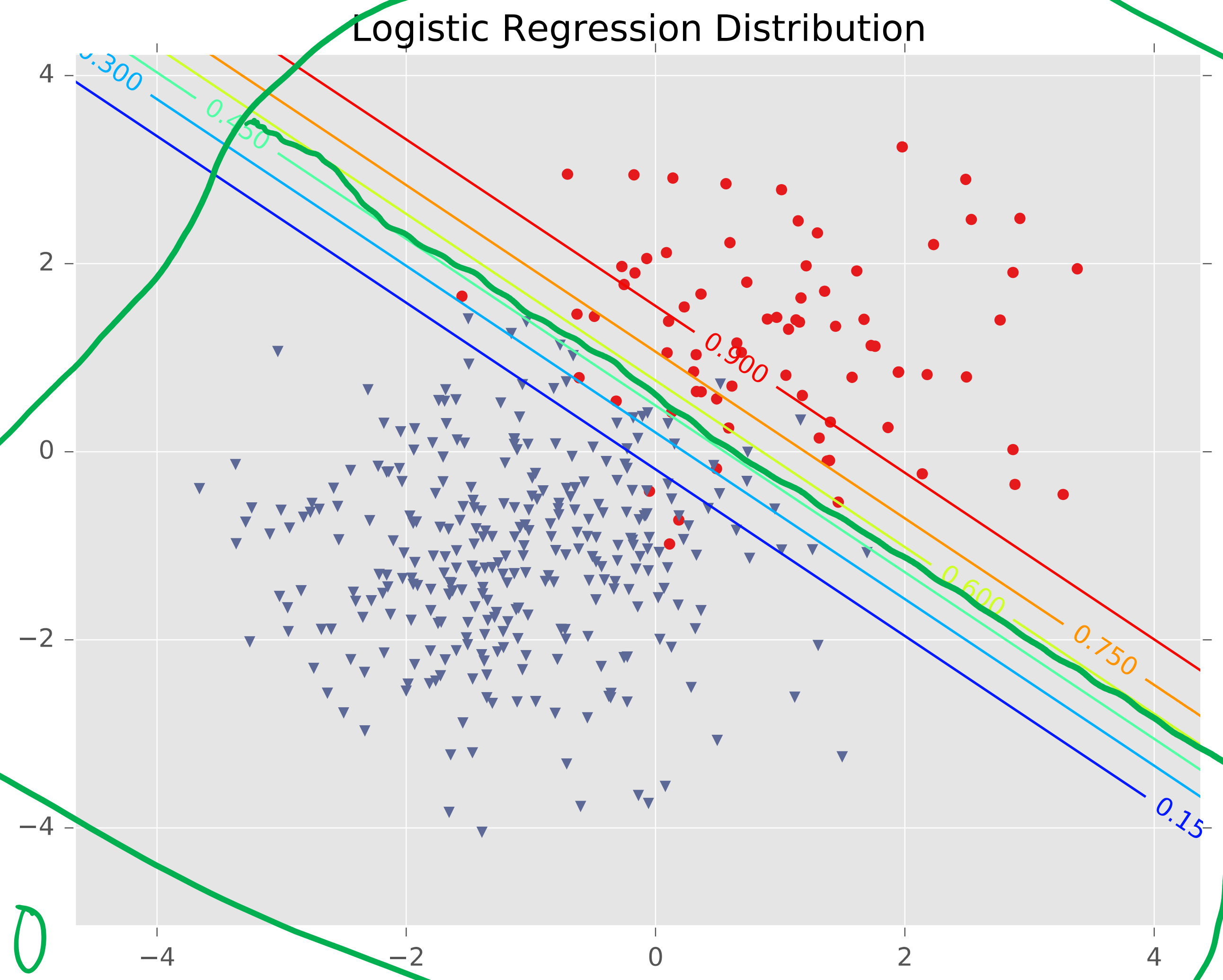
Building on a Linear Model

$z = \vec{w}^T \vec{x} + b$



Building on a Linear Model

$\hat{y} = g(z)$



$P(Y=1|x) \approx 1.0$

$\hat{y} = P(Y=1|\vec{x}) = 0.5$

$P(Y=1|x) = 0.0$

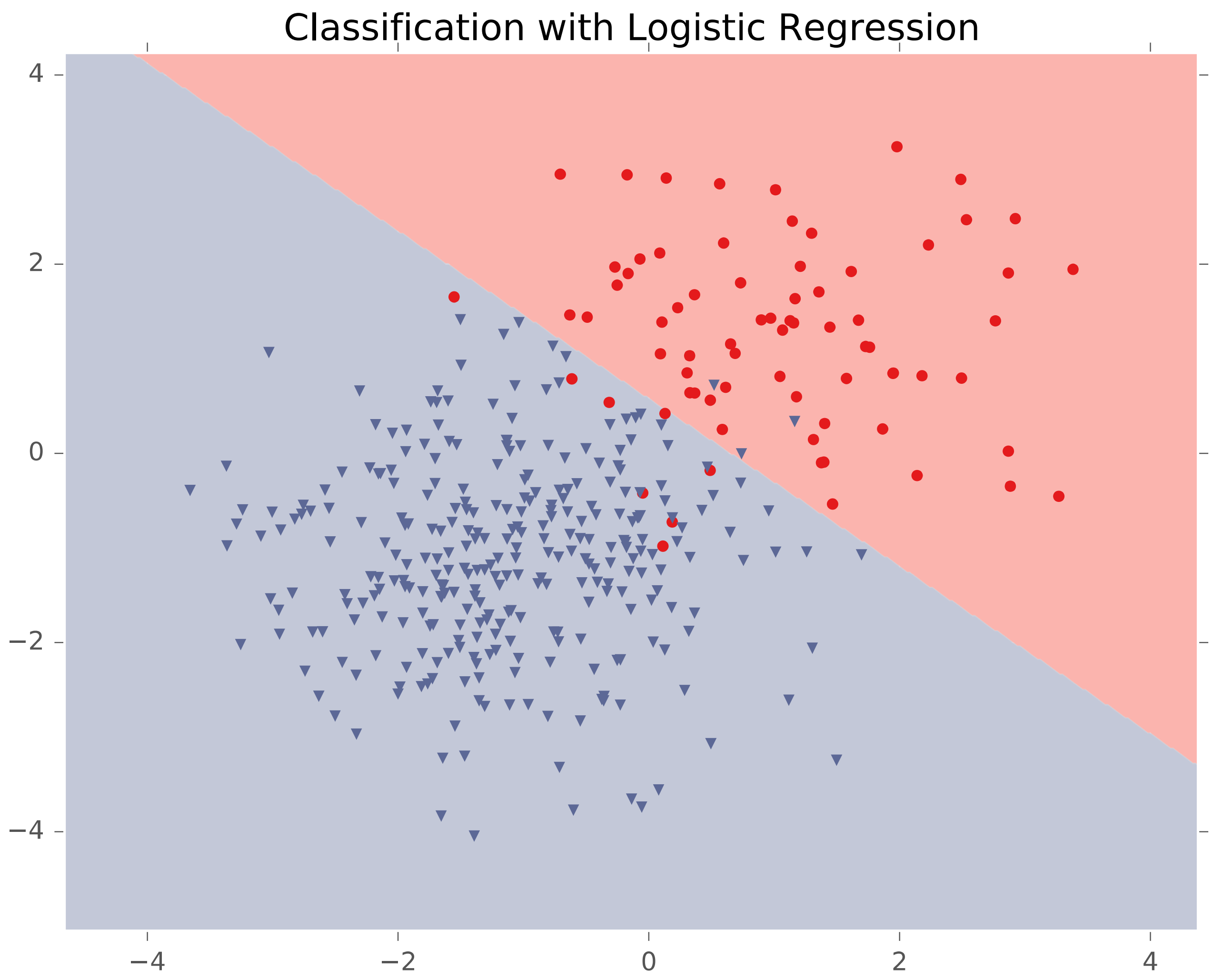
Logistic Regression Decision Boundary

Slide credit: CMU MLD Matt Gormley

Logistic Regression

Linear decision boundary

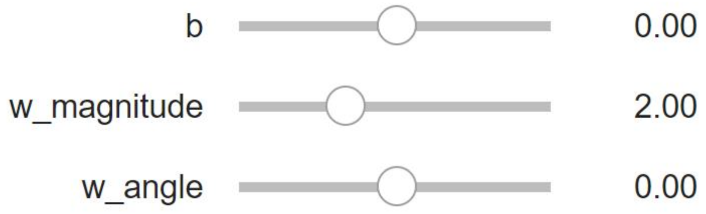
Logistic Regression Decision Boundary

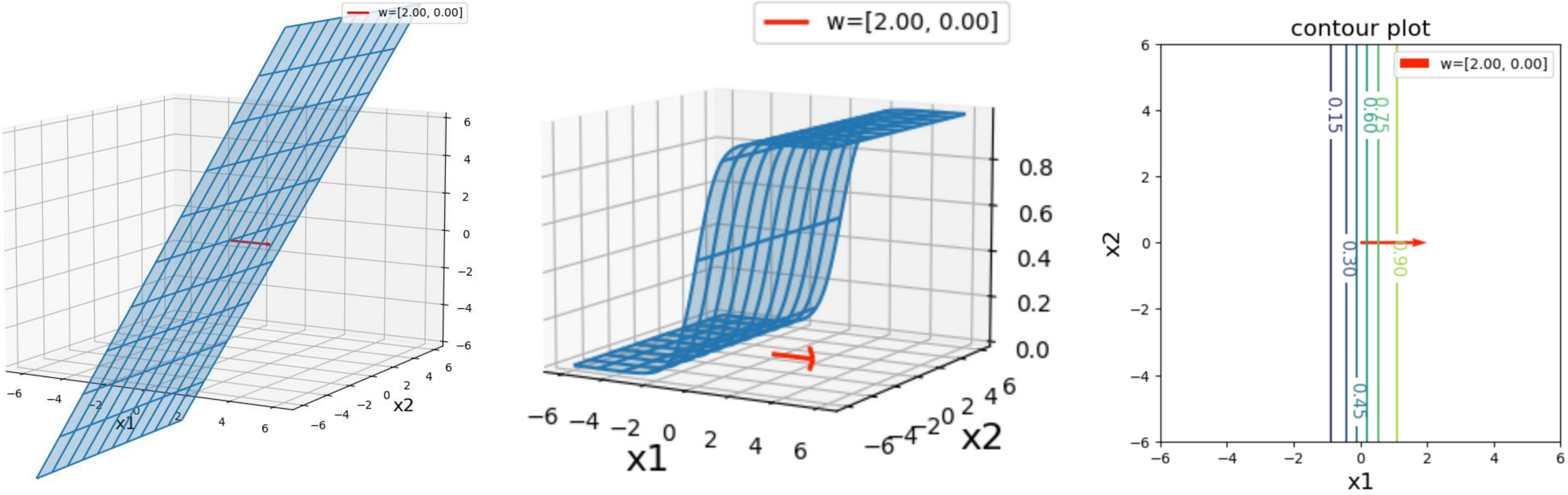


Slide credit: CMU MLD Matt Gormley

Exercise

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





Linear in Higher Dimensions

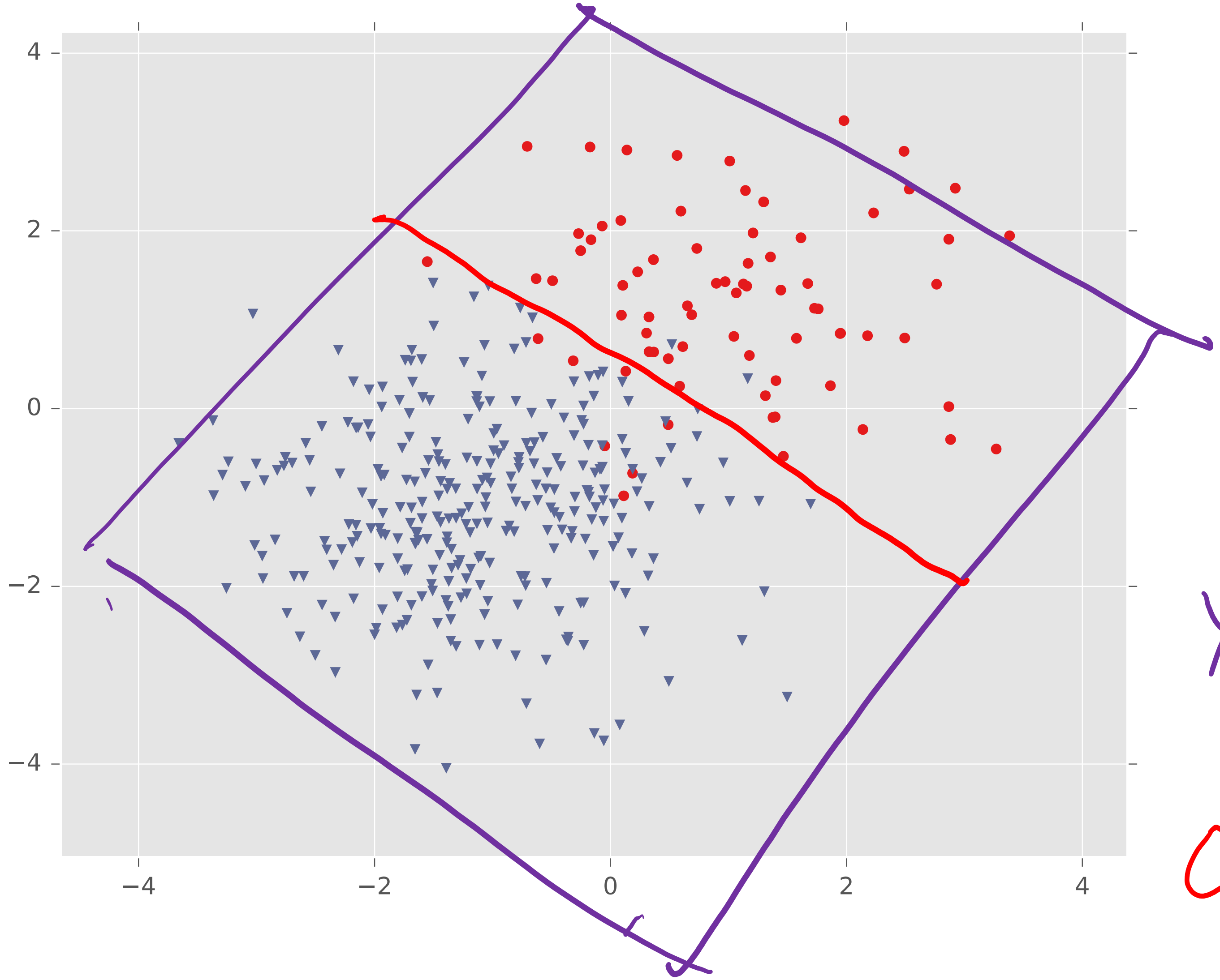
1-D $y = wx + b$

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?

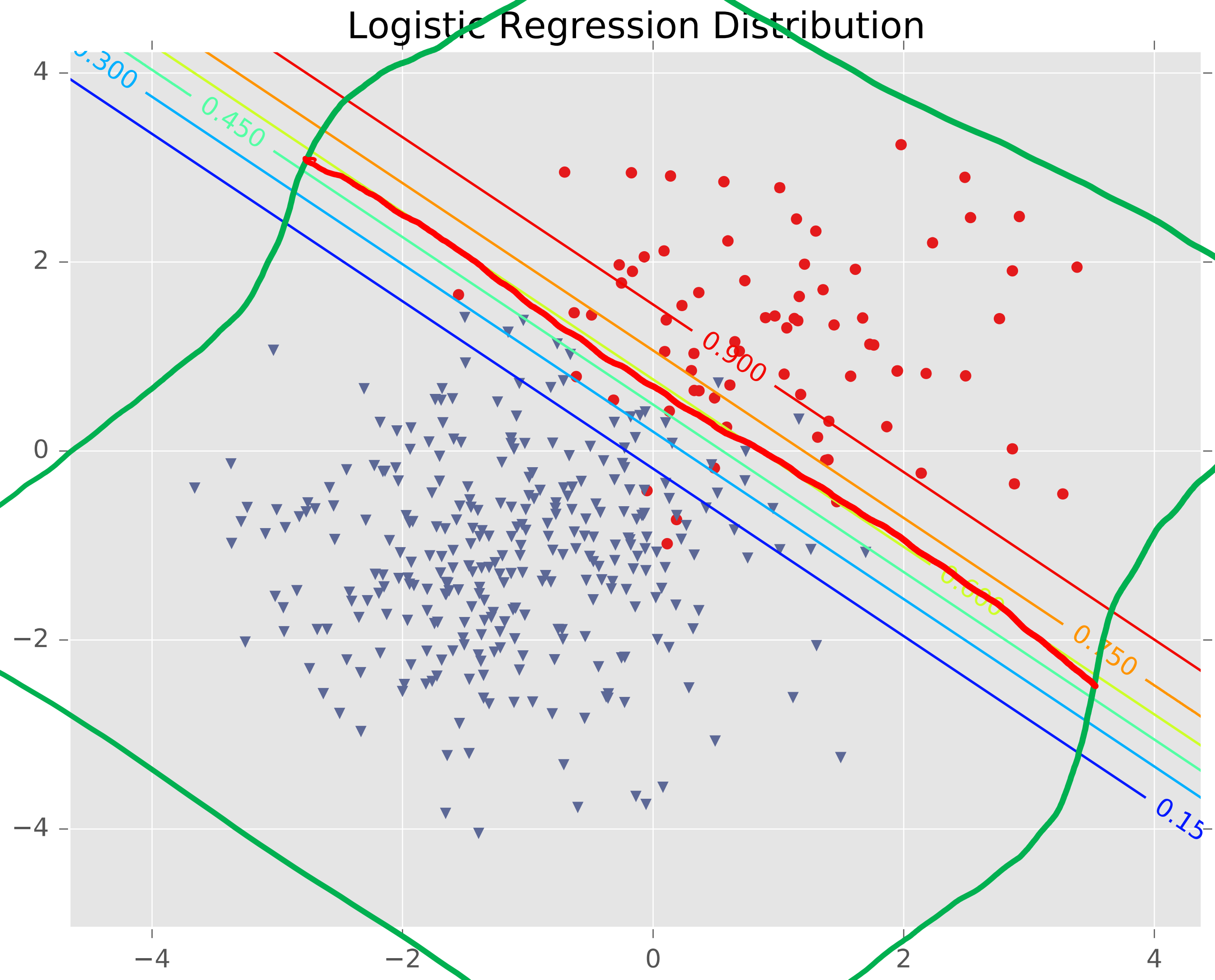
	$\boldsymbol{x} \in \mathbb{R}$	$\boldsymbol{x} \in \mathbb{R}^2$	$\boldsymbol{x} \in \mathbb{R}^3$	$\boldsymbol{x} \in \mathbb{R}^M$
$y = \boldsymbol{w}^T\boldsymbol{x} + b$	line	plane	hyperplane	hyperplane
$\boldsymbol{w}^T\boldsymbol{x} + b = 0$	point	line	plane	hyperplane
$\boldsymbol{w}^T\boldsymbol{x}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

Logistic Regression



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

Logistic Regression



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

$$= g(w^T x + b)$$

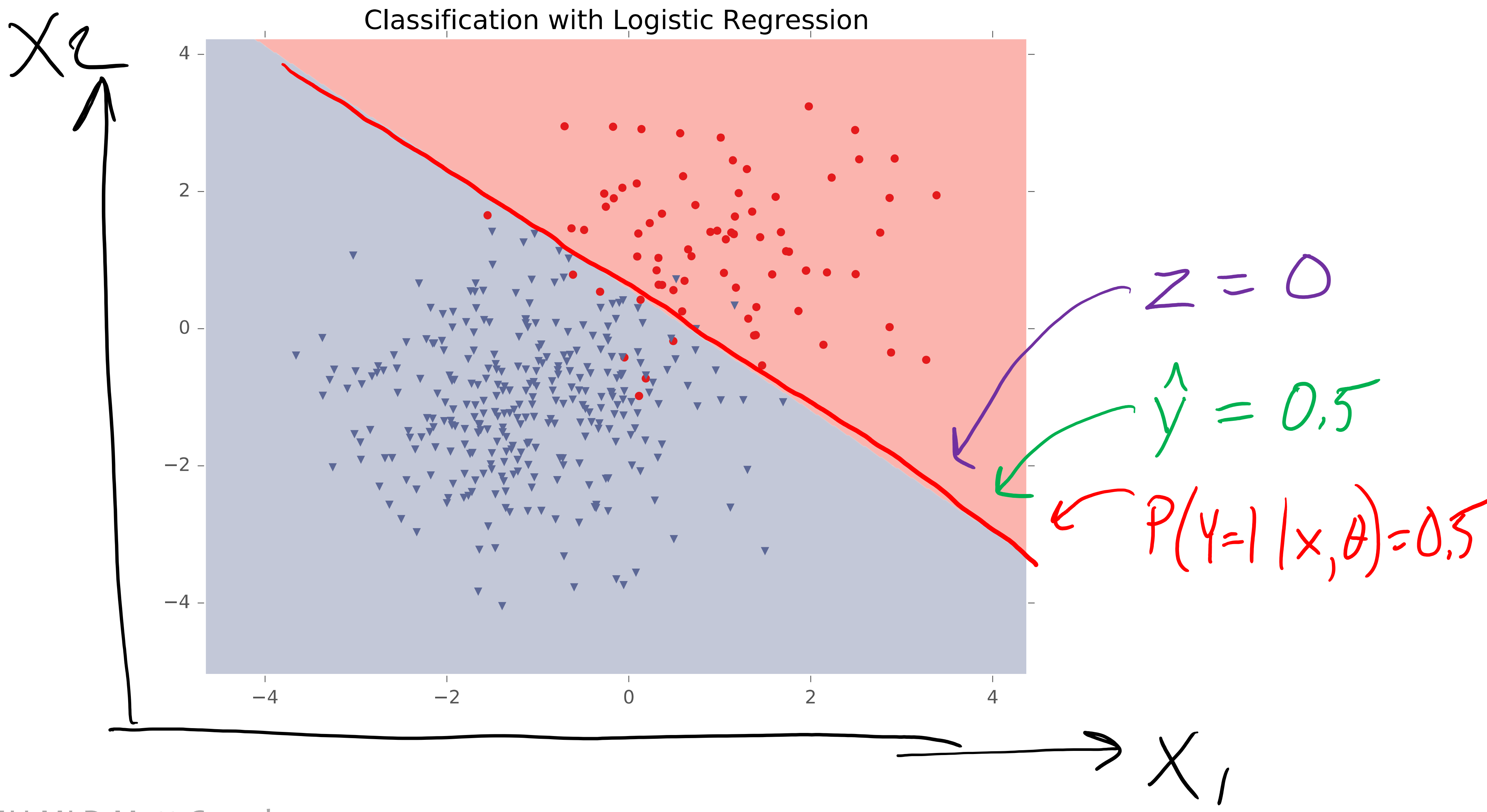
$$= \frac{1}{1+e^{-\theta^T x}}$$

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

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

$$= 0.5$$

Logistic Regression



Slide credit: CMU MLD Matt Gormley

Poll 3

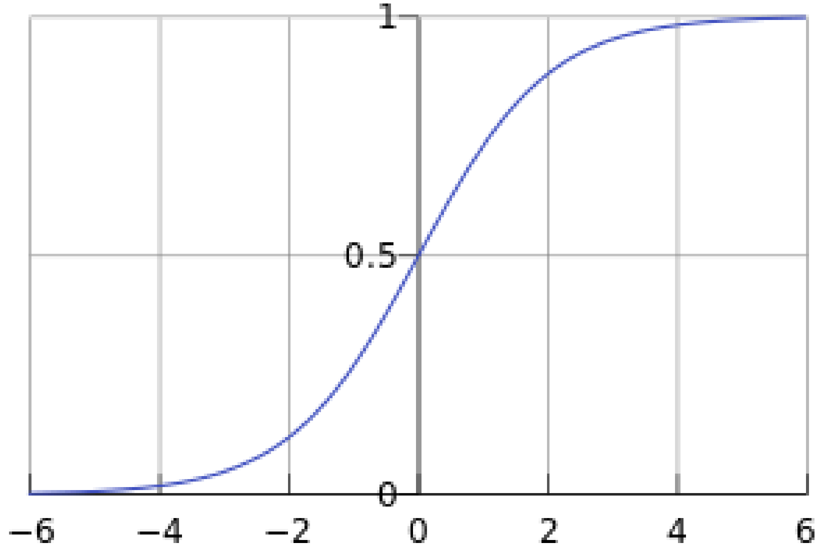
For a point $\mathbf{x}$ on the decision boundary of logistic regression, does $g(\mathbf{w}^T\mathbf{x} + b) = \mathbf{w}^T\mathbf{x} + b$?

A) Yes

B) No

C) I have no idea

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



Logistic Regression

Optimization

Optimizing a Model for Cancer Diagnosis

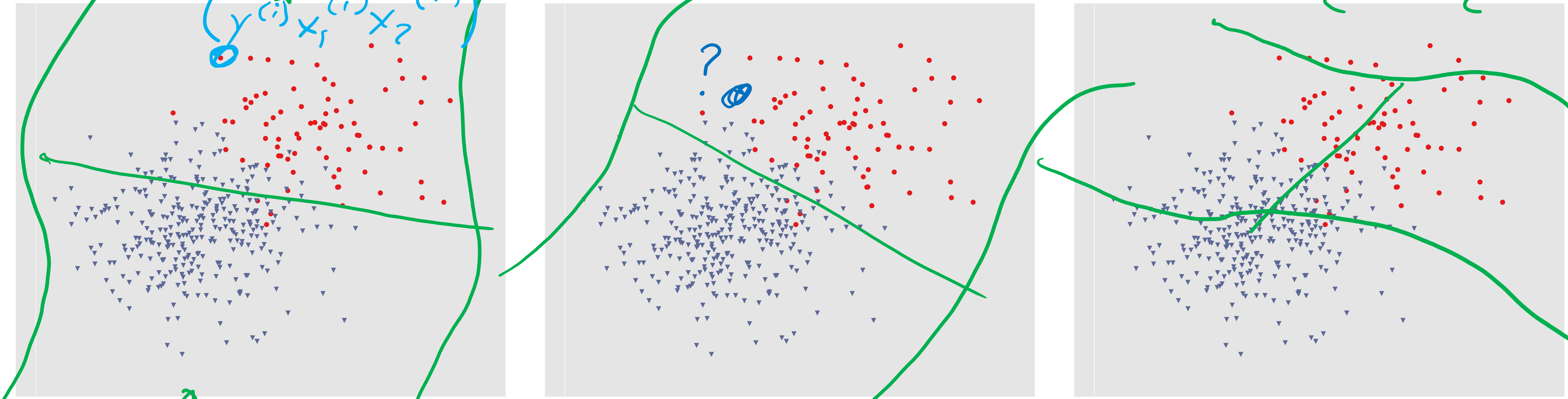
Learn to predict if a patient has cancer ($Y = 1$) or not ($Y = 0$) given the input of two test results, X_A, X_B. Note: bias term included in $\mathbf{x}$.

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

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

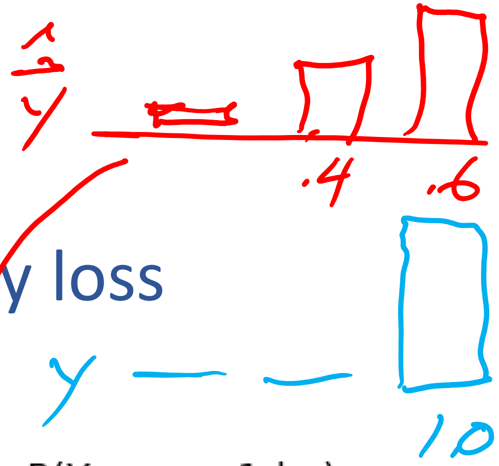
$x_1 w_2 + x_2 w_2 + b$

$(y^{(i)}, x_1^{(i)}, x_2^{(i)})$



θ_A θ_B θ_C

Empirical Risk Minimization



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

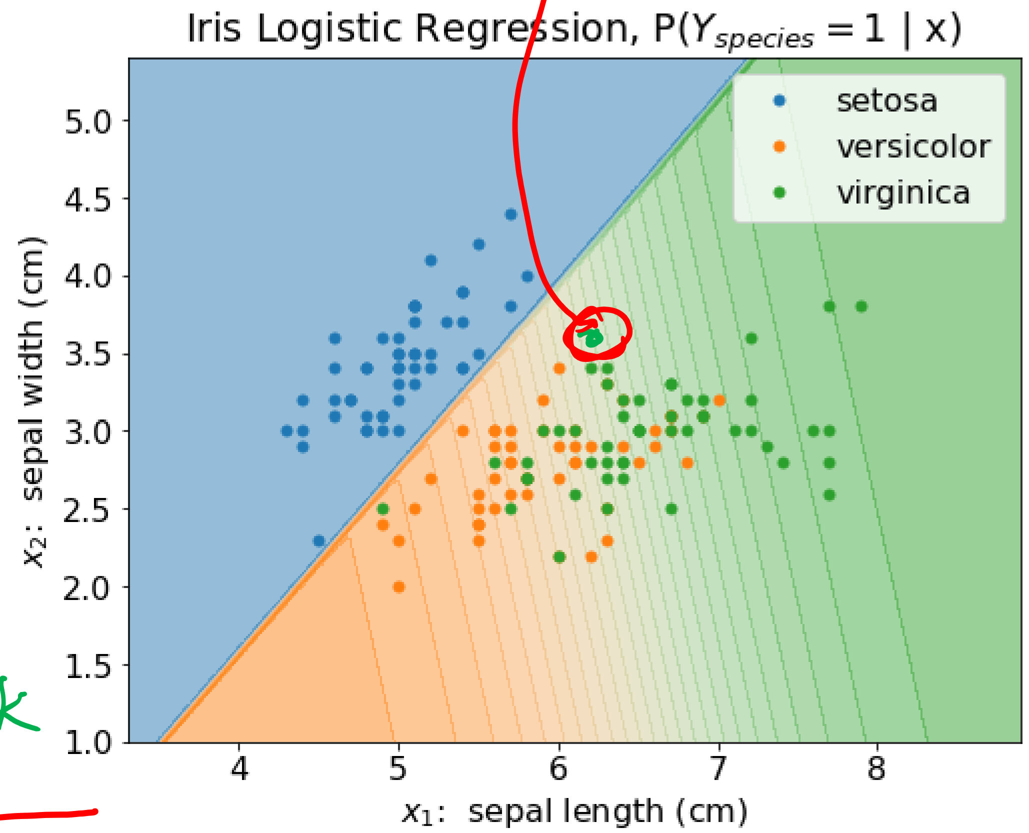
$$h^* = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{R}(h)$$

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

true one hot

$= g(\theta^T x)$

$$\ell(\vec{y}, \hat{\vec{y}}) = -\sum_k y_k \log \hat{y}_k$$



Empirical Risk Minimization

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

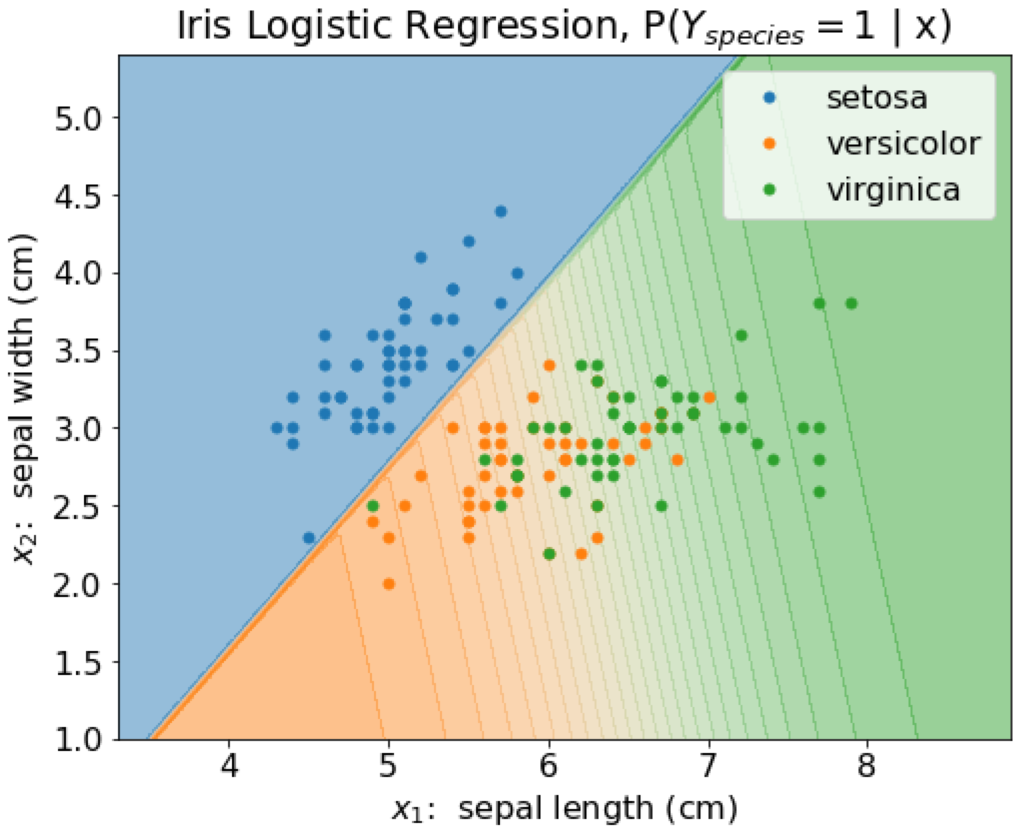
$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \; \hat{R}(h)$$

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

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



Binary Logistic Regression

1) Model

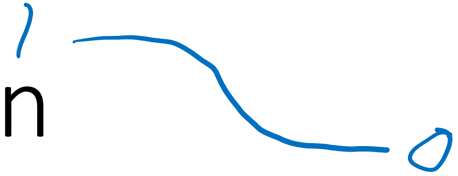
$$\hat{y} = g_{logistic}(\theta^T x)$$

2) Objective function

$$J(\theta) = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_k^{(i)} \log \underline{\hat{y}_k^{(i)}}$$

3) Solve for $\widehat{\boldsymbol{\theta}}$ ← S.G.D.

Binary Logistic Regression



$\hat{y} = g(z) = \frac{1}{1+e^{-z}}$ $\quad P(Y=1|x)$

$= P(Y_1=1|x)$

Objective: Special case for binary logistic regression

$\hat{y} = g(\boldsymbol{\theta}^T \mathbf{x})$

$J^{(i)}(\theta)$

$$J(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i}\sum_{k=1}^{2} y_k^{(i)} \log \hat{y}_k^{(i)}$$

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

$$= -\frac{1}{N}\sum_{i}\left(y^{(i)} \log \hat{y}^{(i)} + \left(1-y^{(i)}\right) \log\left(1-\hat{y}^{(i)}\right)\right)$$

$J^{(i)}(\theta)$

Multiclass

$y^{(i)} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} y \\ (1-y) \end{bmatrix}$

$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} = \begin{bmatrix} \hat{y} \\ (1-\hat{y}) \end{bmatrix}$

Binary

$y^{(i)} \in \{0, 1\}$

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

Solve Logistic Regression

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

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

1 0

0

0 0

1

$y^{(i)} = 1$

$y^{(i)} = 0$

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

Solve Logistic Regression

$z = \theta^T x$

$\hat{y} = g(z)$

$\frac{d}{du} u^T v = v$ or v^T

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

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

$$J^{(i)}(\boldsymbol{\theta}) = -\underbrace{\left[y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})\right]}_{\ell(y^{(i)}, \hat{y}^{(i)})}$$

$$\frac{\partial J}{\partial \theta} = \frac{\partial \ell}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \theta}$$

$$= -\left[\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right] \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial \theta}$$

$$= -\left[\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right] \hat{y}(1-\hat{y}) \vec{x}$$

$$\frac{\partial J^{(i)}}{\partial \boldsymbol{\theta}} = -(y^{(i)} - \hat{y}^{(i)}) \mathbf{x}^{(i)}$$

Solve Logistic Regression

$z = \theta^T x$

$\hat{y} = g(z)$

$\frac{d}{du} u^T v = v$ or v^T

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

$$J^{(i)}(\boldsymbol{\theta}) = \underbrace{-\left[y^{(i)} \log \hat{y}^{(i)} + \left(1-y^{(i)}\right)\log\left(1-\hat{y}^{(i)}\right)\right]}_{\ell(y^{(i)},\, \hat{y}^{(i)})}$$

$$\frac{\partial J}{\partial \theta} = \frac{\partial \ell}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \theta}$$

$$= -\left[\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right] \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial \theta}$$

$$= -\left[\cancel{\frac{y}{\hat{y}}} - \frac{1}{\cancel{1-\hat{y}}}\right] \hat{y}\cancel{(1-\hat{y})}\,\vec{x}$$

Case: $y^{(i)} = 0$

$= -(0 - \hat{y})\vec{x}$

$= -(y - \hat{y})\vec{x}$

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

Solve Logistic Regression

$z = \theta^T x$

$\hat{y} = g(z)$

$\frac{d}{du} u^T v = v$ or v^T

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

$$J^{(i)}(\boldsymbol{\theta}) = \underbrace{-\left[y^{(i)} \log \hat{y}^{(i)} + \left(1 - y^{(i)}\right) \log\left(1 - \hat{y}^{(i)}\right)\right]}_{} \quad \ell(y^{(i)}, \hat{y}^{(i)})$$

$$\frac{\partial J}{\partial \theta} = \frac{\partial \ell}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \theta}$$

$$= -\left[\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right] \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial \theta}$$

$$= -\left[\frac{\cancel{y}}{\cancel{\hat{y}}} - \cancel{\frac{1-y}{1-\hat{y}}}\right] \cancel{\hat{y}}(1 - \hat{y}) \vec{x}$$

Case: $y^{(i)} = 1$

$$= -(1 - \hat{y}) \vec{x}$$

$$= -(y - \hat{y}) \vec{x}$$

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

Solve Logistic Regression

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

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

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

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

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

No closed form solution ☹

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

Solve Logistic Regression

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

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

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

No closed form solution ☹

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

Good news: The logistic regression optimization function is convex!

Logistic Regression

Convexity

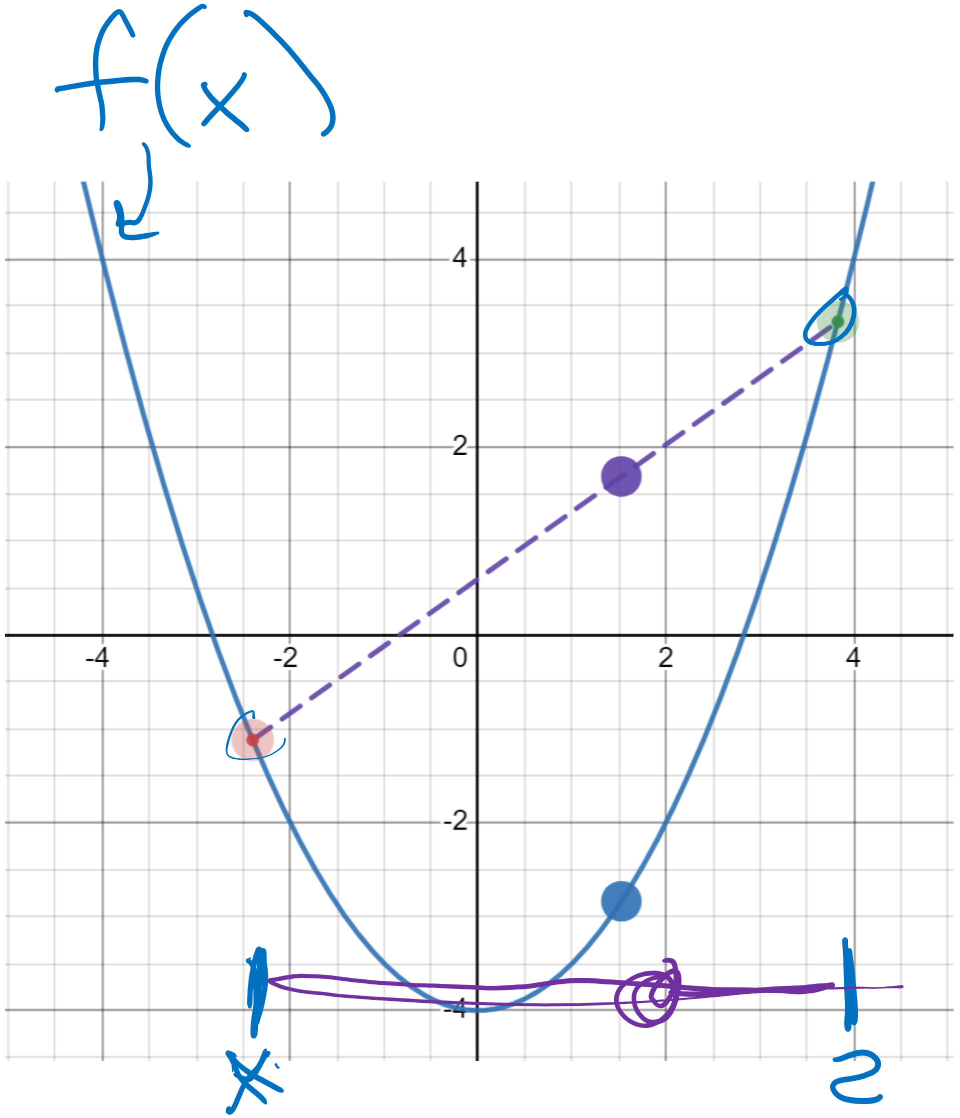
Optimization

Convex function

If $f(\boldsymbol{x})$ is convex, then:

- $f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{z}) \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{z})$
  $\forall\ 0 \leq \alpha \leq 1$

[Demo on Desmos]()



Optimization

Convex function

If $f(\boldsymbol{x})$ is convex, then:

- $f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{z}) \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{z}) \quad \forall\, 0 \leq \alpha \leq 1$

Convex optimization

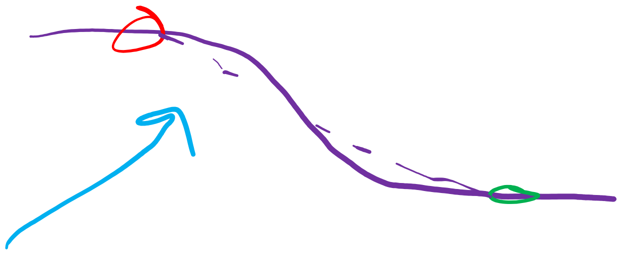
If second derivative is ≥ 0 everywhere then function is convex

If $f(\boldsymbol{x})$ is convex, then:

- Every local minimum is also a global minimum ☺

Optimization

$h(\mathbf{x}) = g(\mathbf{w}^\top \mathbf{x} + b)$ is definitely not convex



But…what are we optimizing over in logistic regression?

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

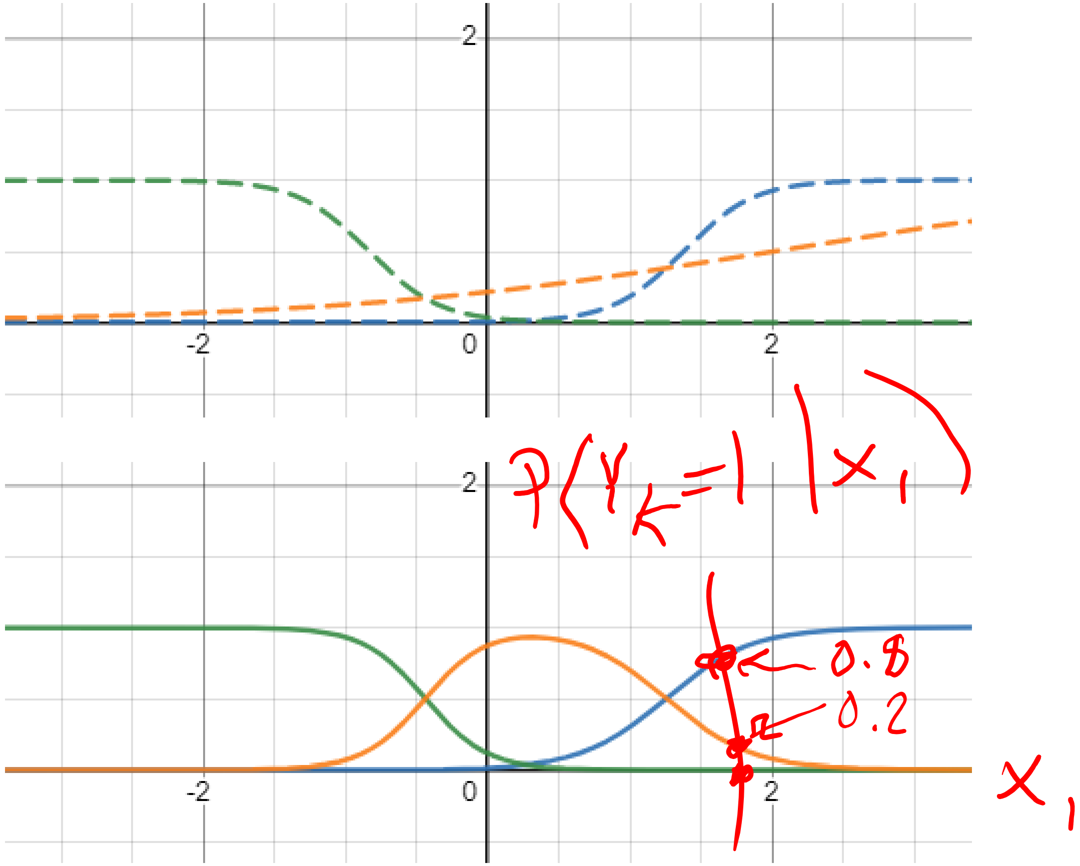
Multi-class Logistic Regression

Multi-class Logistic Regression

$$\Gamma = e^{\theta_g^T x} + e^{\theta_o^T x} + e^{\theta_b^T x}$$

Desmos Demo:

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

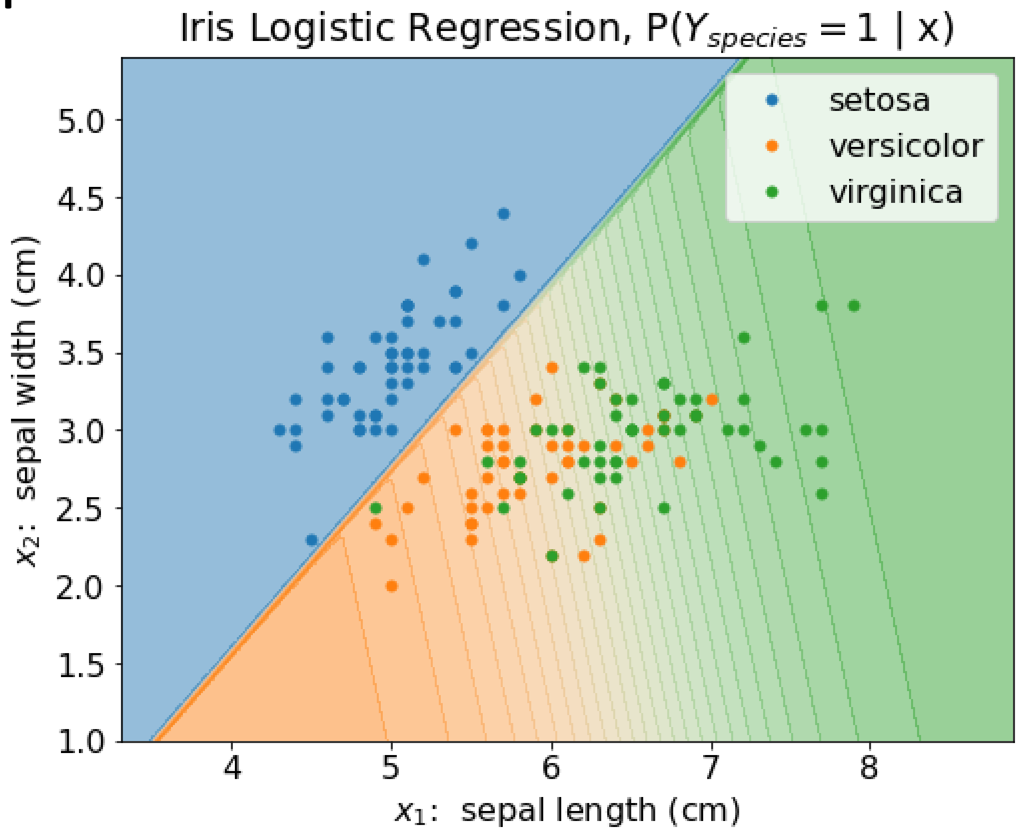


$$P(y_{green} = 1 \mid x_1) = \frac{e^{\theta_g^T x}}{\Gamma}$$

$$P(y_{orange} = 1 \mid x_1) = \frac{e^{\theta_o^T x}}{\Gamma}$$

$$P(y_{blue} = 1 \mid x_1) = \frac{e^{\theta_b^T x}}{\Gamma}$$

Multi-class Logistic Regression



Cross-entropy loss

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

Model

$$\hat{\mathbf{y}} = h(\mathbf{x}) = g_{softmax}(\mathbf{z})$$

$$\mathbf{z} = \Theta\mathbf{x}$$

$$z_k = \boldsymbol{\theta}_k \mathbf{x}$$

One vector of parameters for each class

Stacked into a matrix of $K \times M$ parameters

$$\mathbf{x} = \begin{bmatrix} 1 \\ \mathrm{x}_1 \\ \mathrm{x}_2 \end{bmatrix} \qquad \boldsymbol{\theta}_k = \begin{bmatrix} b_k \\ w_{k,1} \\ w_{k,2} \end{bmatrix} \qquad \Theta = \begin{bmatrix} - & \boldsymbol{\theta}_1^\top & - \\ - & \boldsymbol{\theta}_2^\top & - \\ - & \boldsymbol{\theta}_3^\top & - \end{bmatrix} = \begin{bmatrix} b_1 & w_{1,1} & w_{1,2} \\ b_2 & w_{2,1} & w_{2,2} \\ b_3 & w_{3,1} & w_{3,2} \end{bmatrix}$$

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.

$z_k = \theta_k^T x$

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

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

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

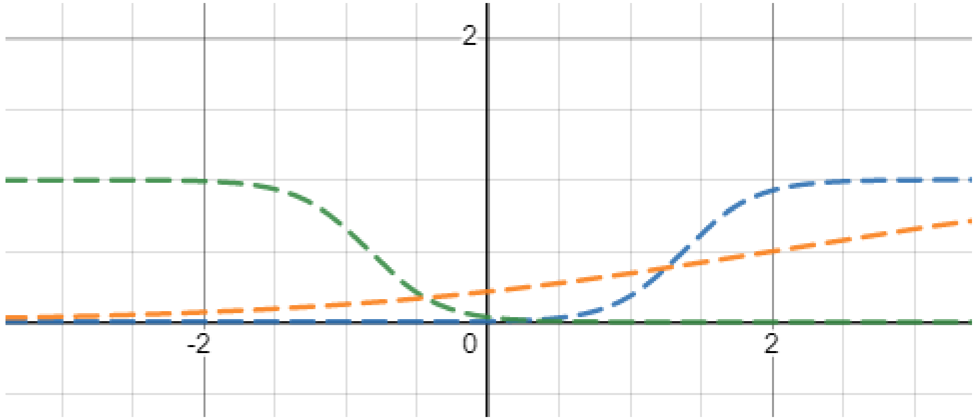
Multiclass Predicted Probability

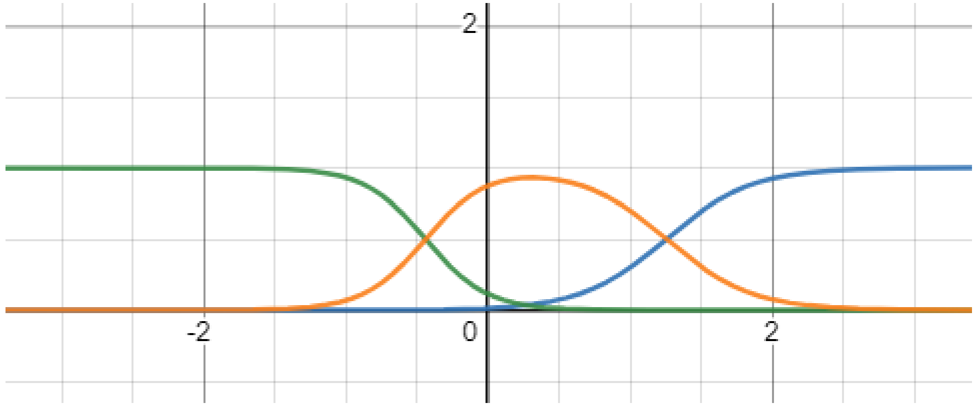
Multiclass logistic regression uses the parameters learned across all K classes to predict the discrete conditional probability distribution of the output Y given a specific input vector $\mathbf{x}$

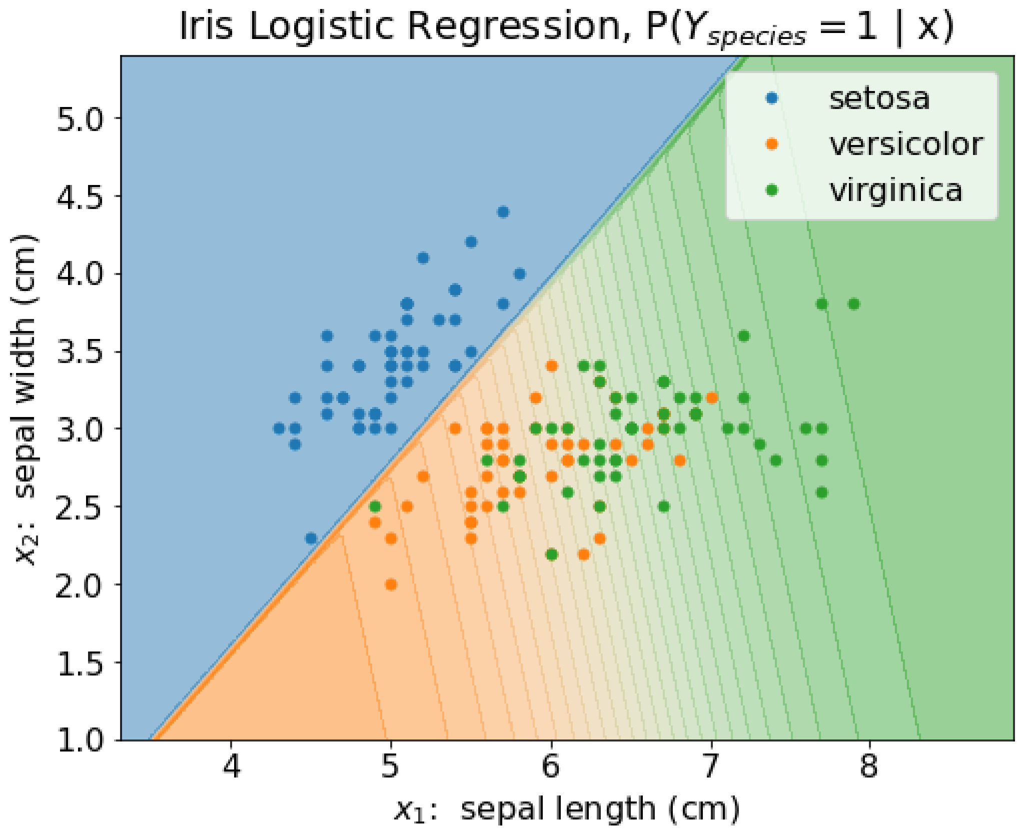
$$\begin{bmatrix} p(Y = 1 \mid \mathbf{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) \\ p(Y = 2 \mid \mathbf{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) \\ p(Y = 3 \mid \mathbf{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)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

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 $\mathbf{x}$

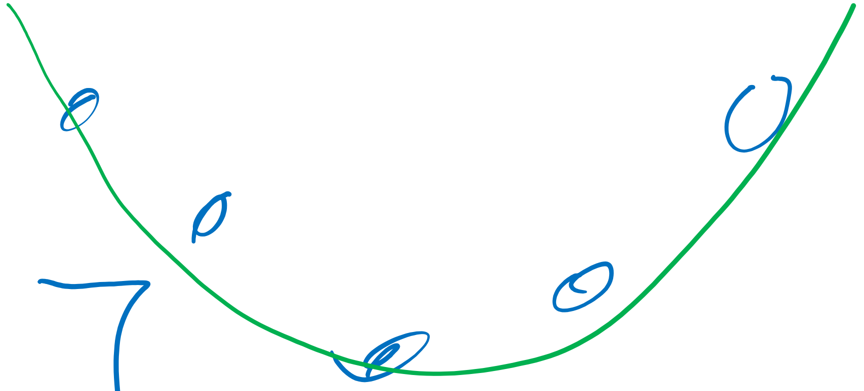






$\vec{x}$

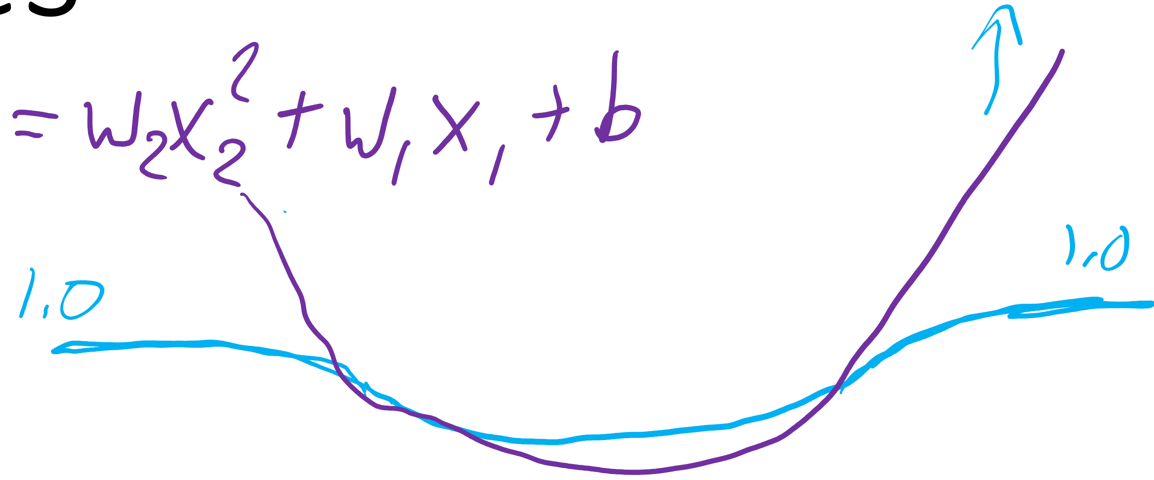
$$x_1 \rightarrow \begin{bmatrix} 1 \\ x_1 \\ x^2 \end{bmatrix}$$



Logistic Regression with Polynomial Features

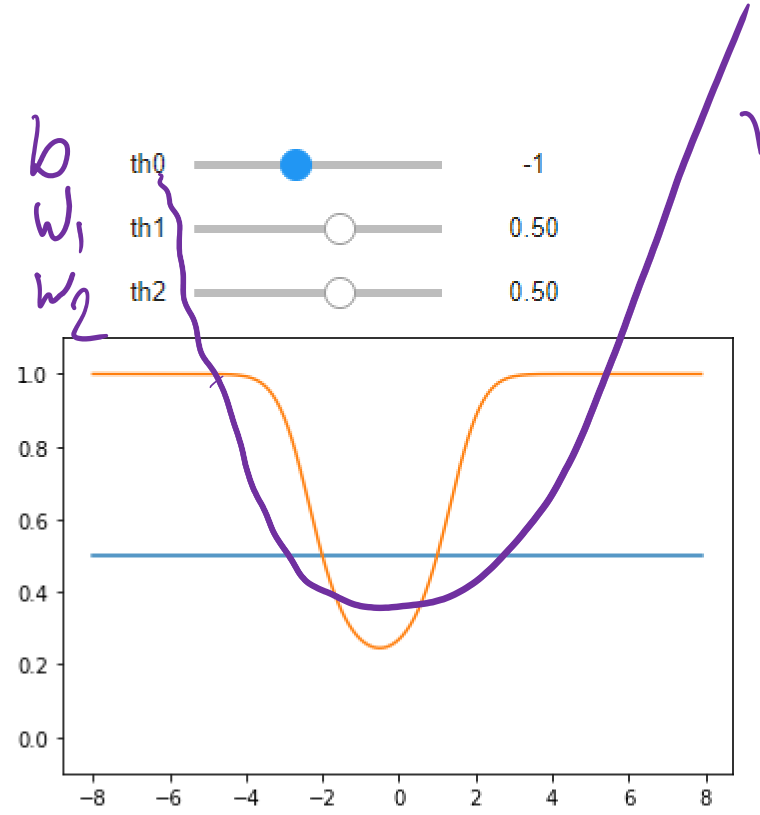
$g(z)$

$$z = w_2 x_2^2 + w_1 x_1 + b$$



Exercise

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



$$w_2 x^2 + w_1 x + b$$

Exercise

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