10-315
Introduction to ML

Logistic Regression

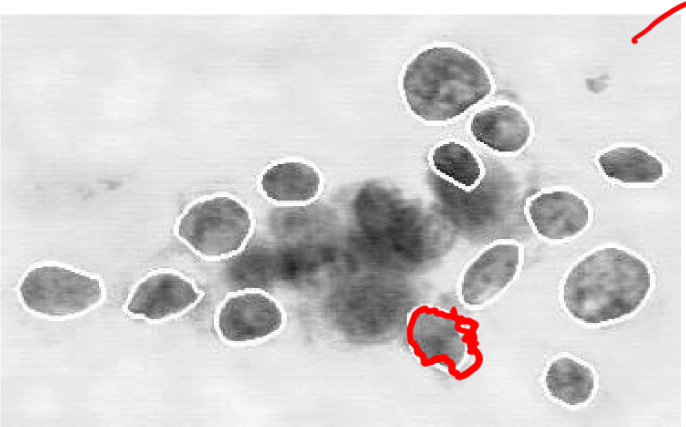
Instructor: Pat Virtue

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.



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

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

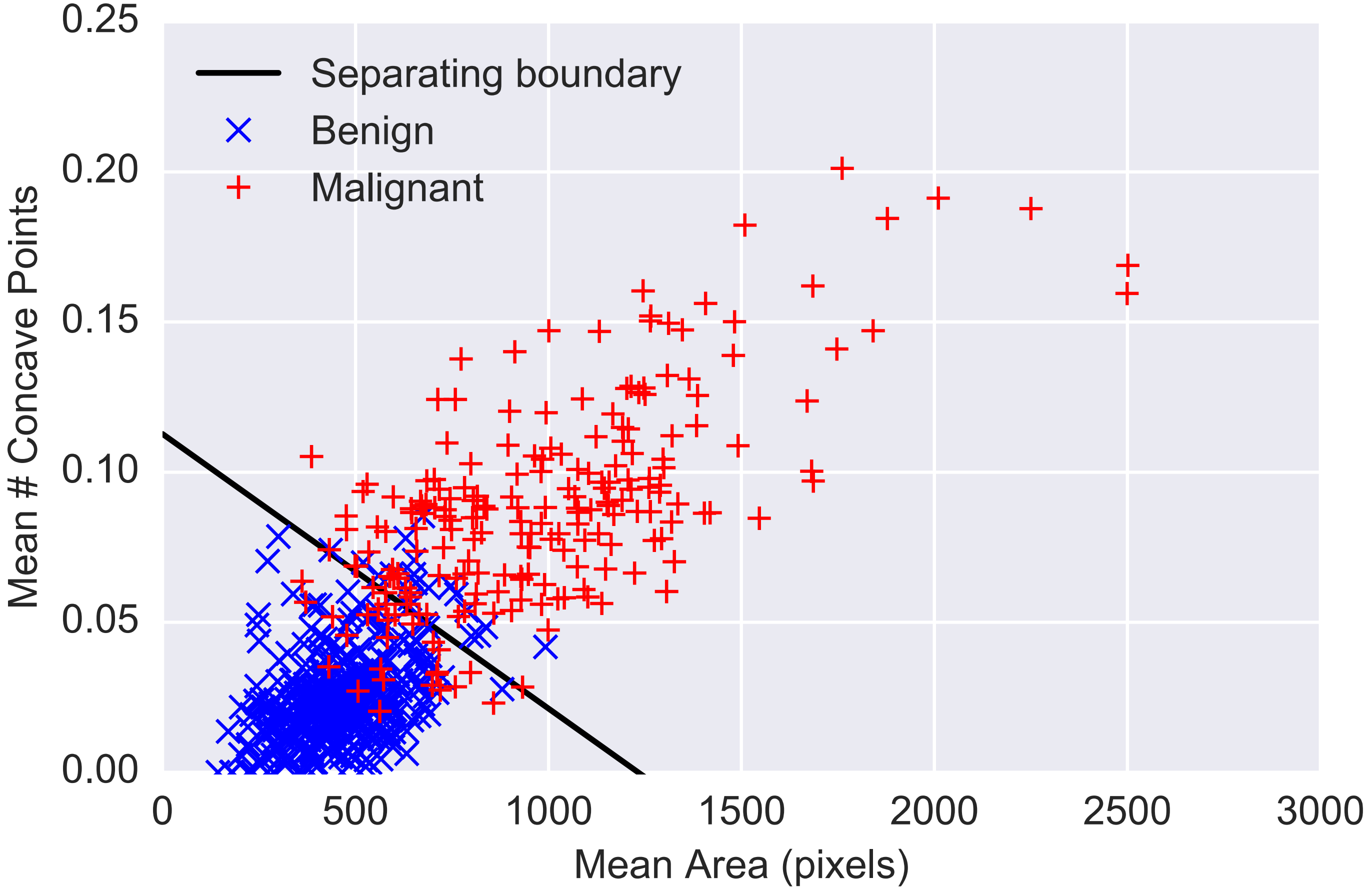
Plot of two features: mean area vs. mean concave points, for two classes

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

Slide credit: CMU AI Zico Kolter

Linear classification example

Linear classification: linear decision boundary

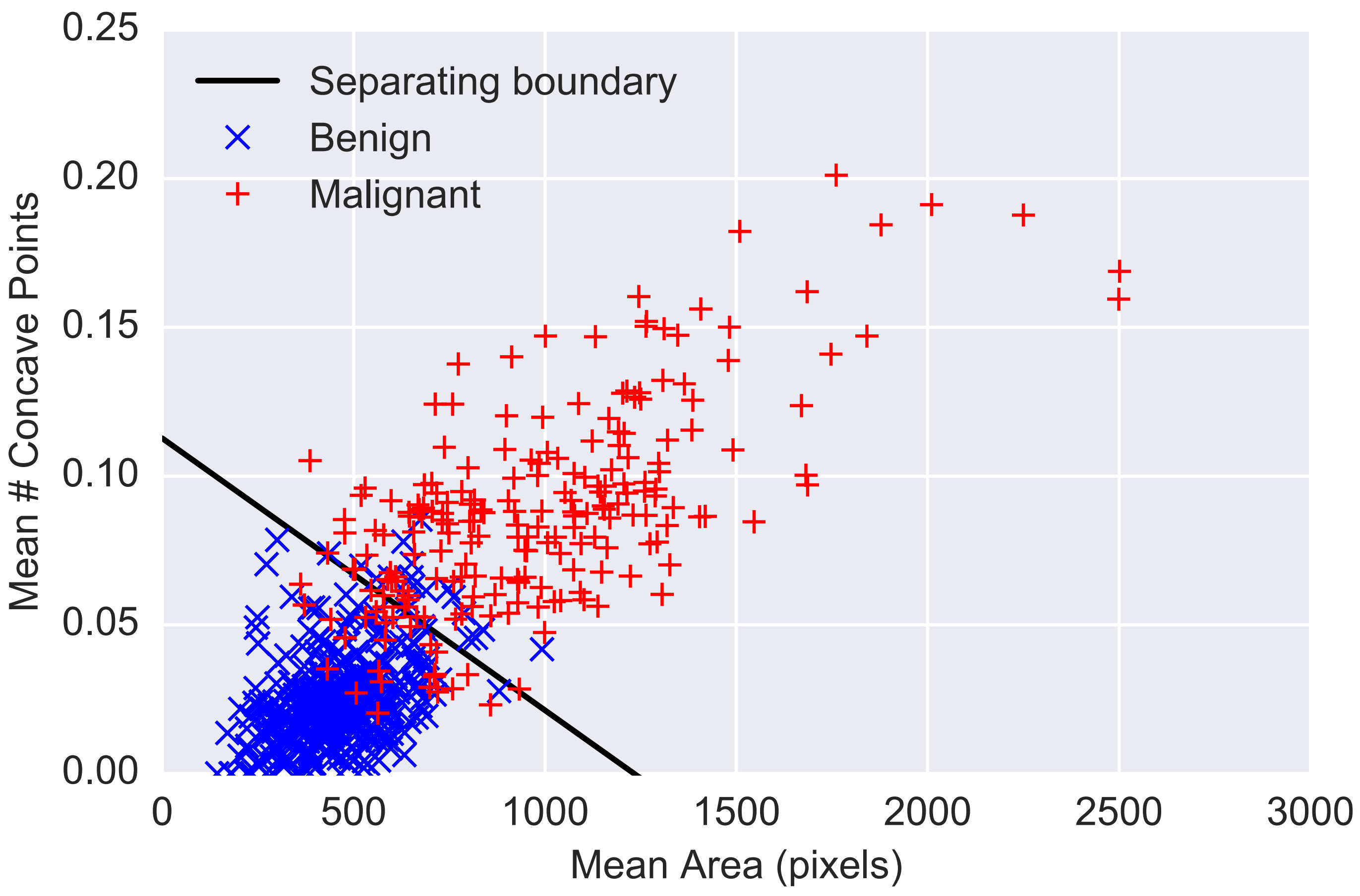


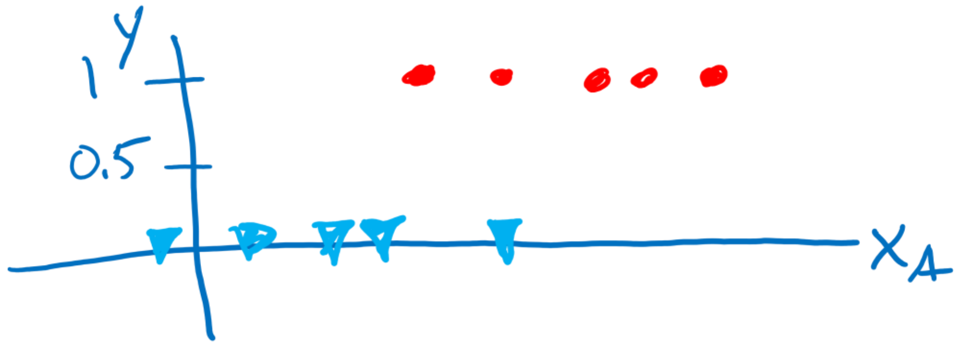
Slide credit: CMU AI Zico Kolter

Logistic regression for classification

Linear classification: linear decision boundary

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



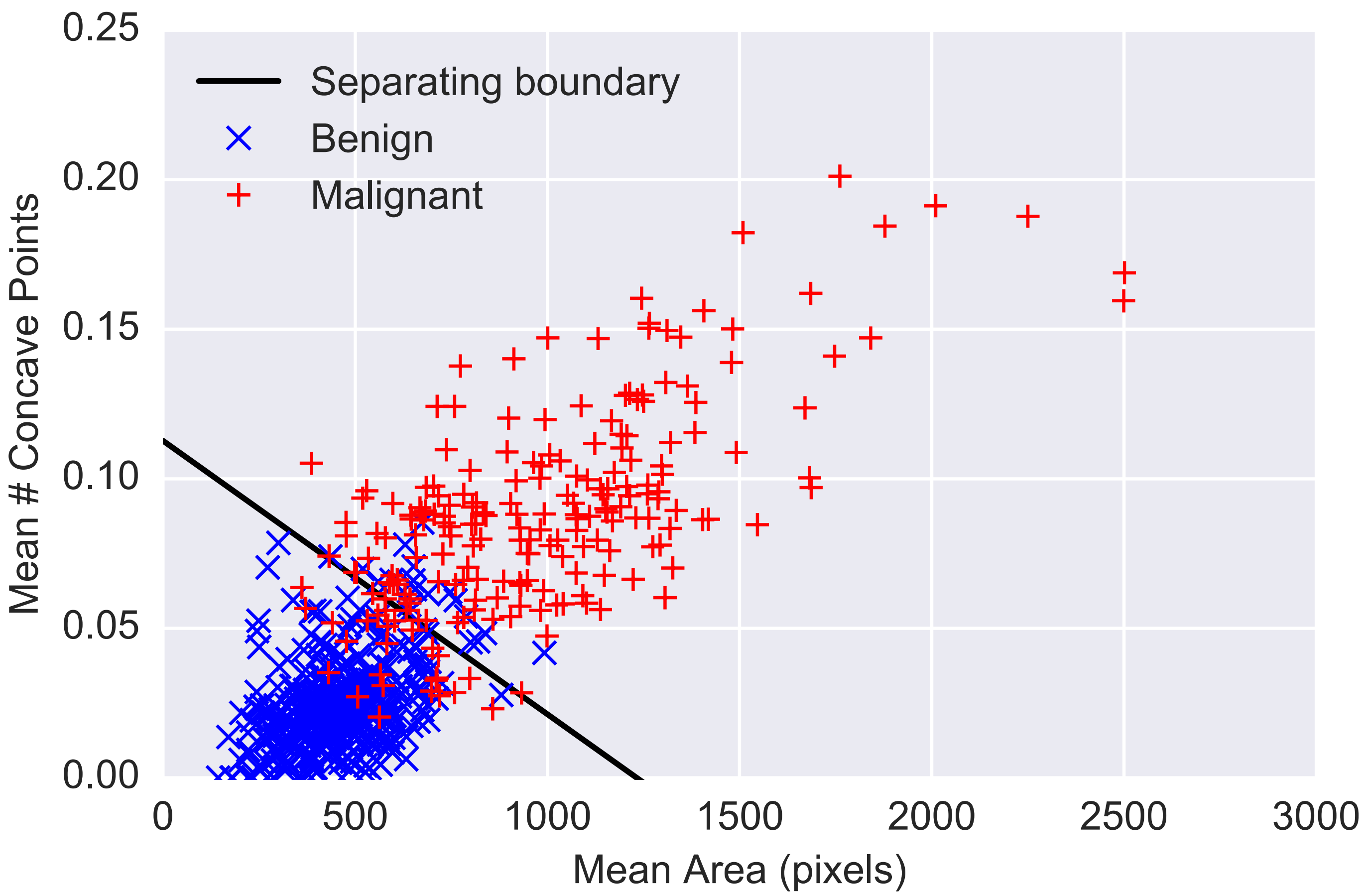


Slide credit: CMU AI Zico Kolter

Logistic regression for classification

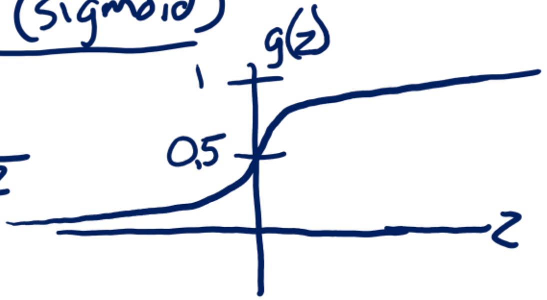
Linear classification: linear decision boundary

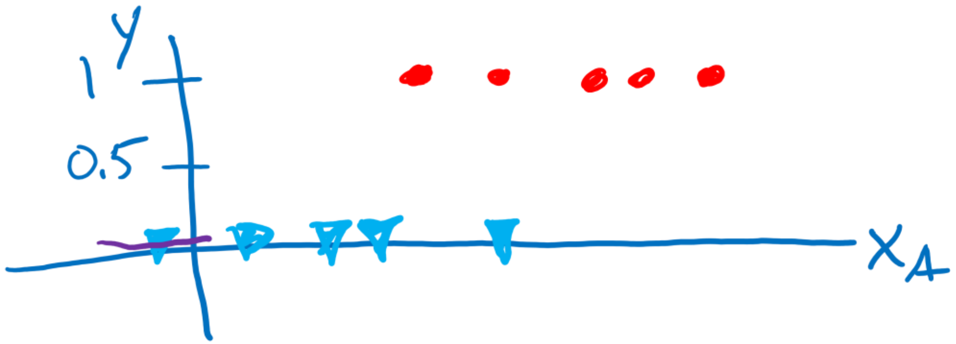
Probabilistic classification: provide $P(Y = 1 \mid x)$ rather than just $\hat{y} \in \{0, 1\}$



logistic function (sigmoid)

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



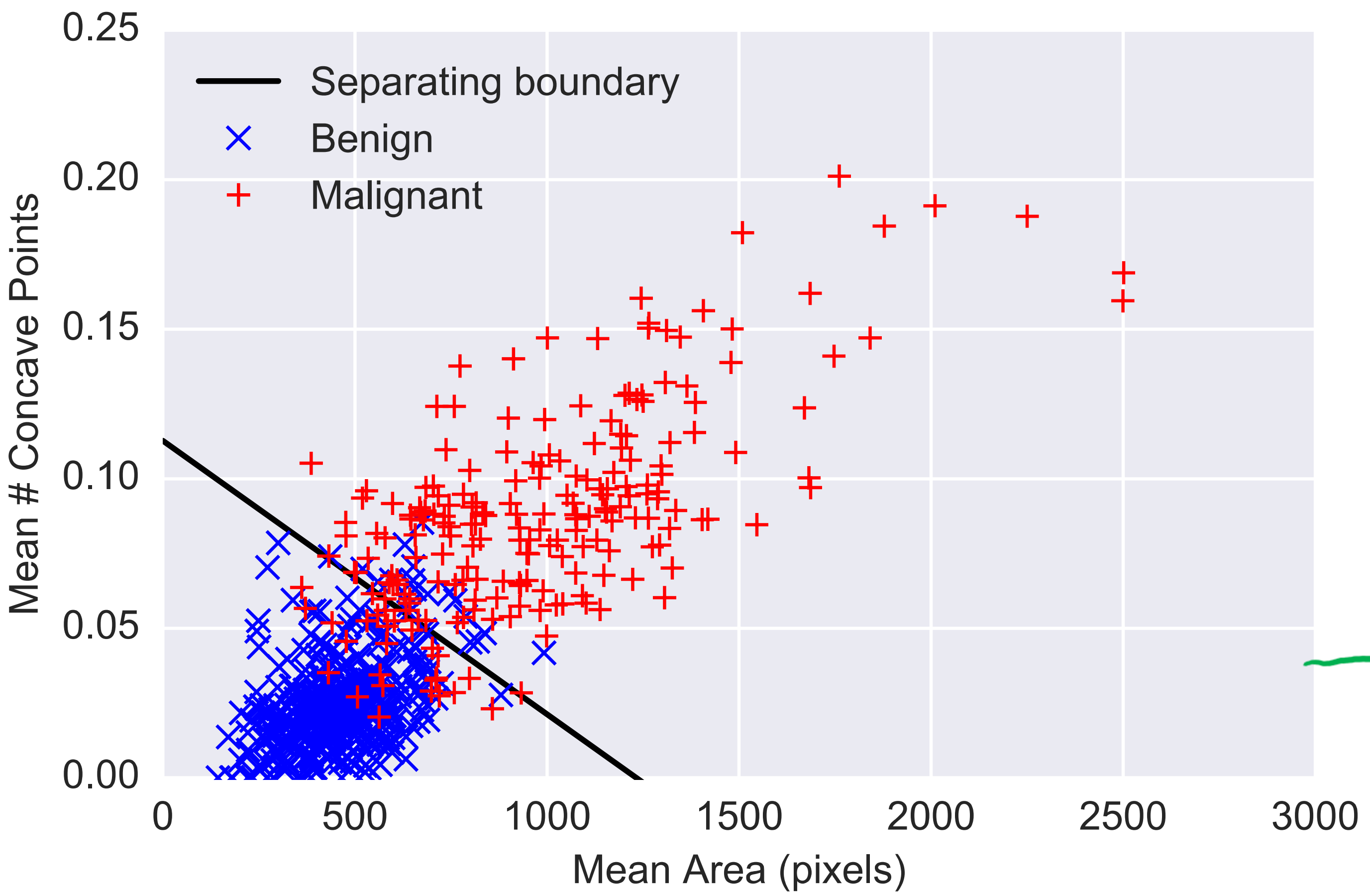


Slide credit: CMU AI Zico Kolter

Logistic regression for classification

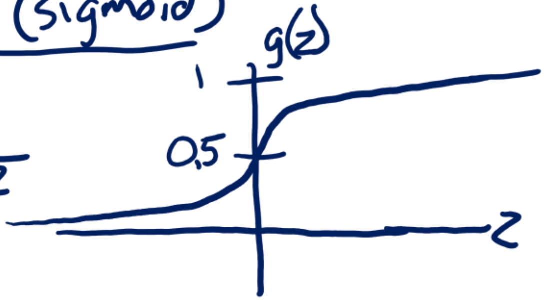
Linear classification: linear decision boundary

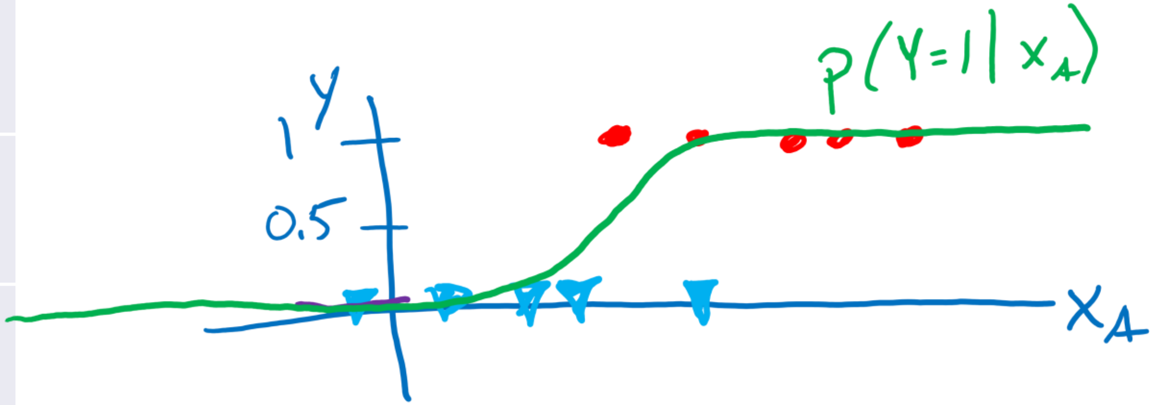
Probabilistic classification: provide $P(Y = 1 \mid x)$ rather than just $\hat{y} \in \{0, 1\}$



logistic function (sigmoid)

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





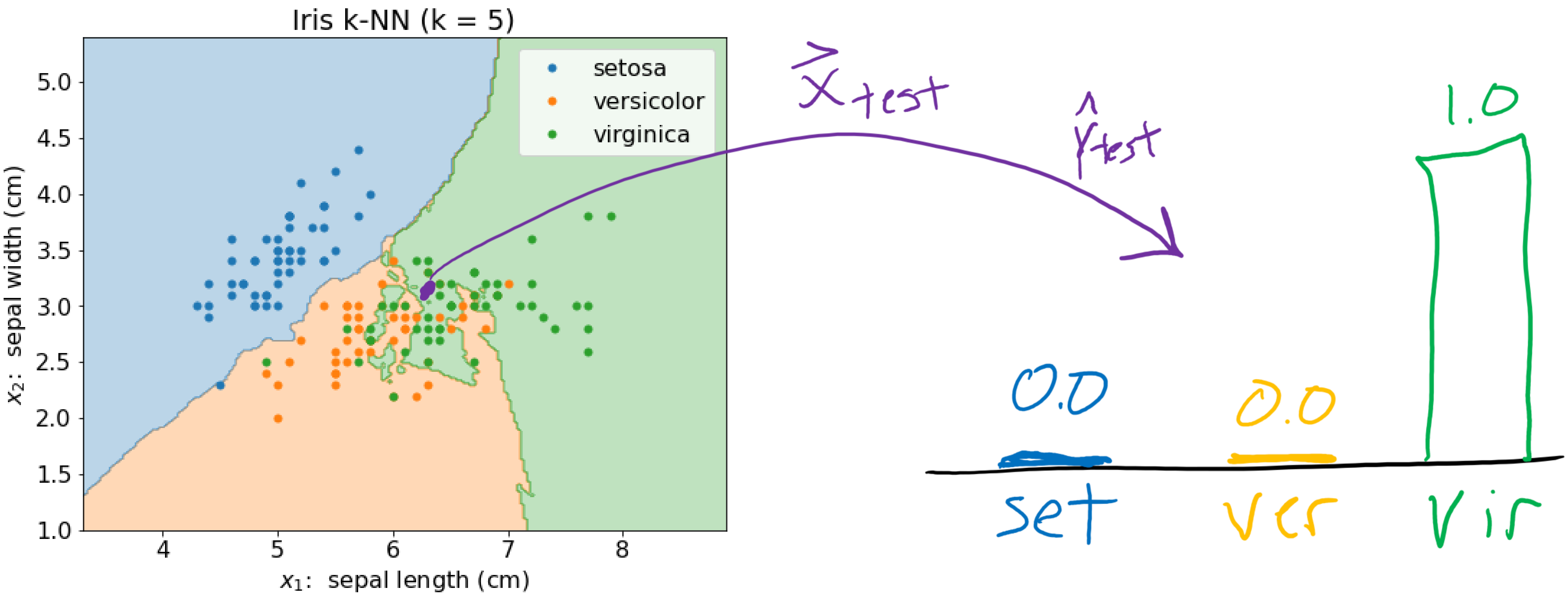
Slide credit: CMU AI Zico Kolter

Pre-reading

Cross-entropy loss

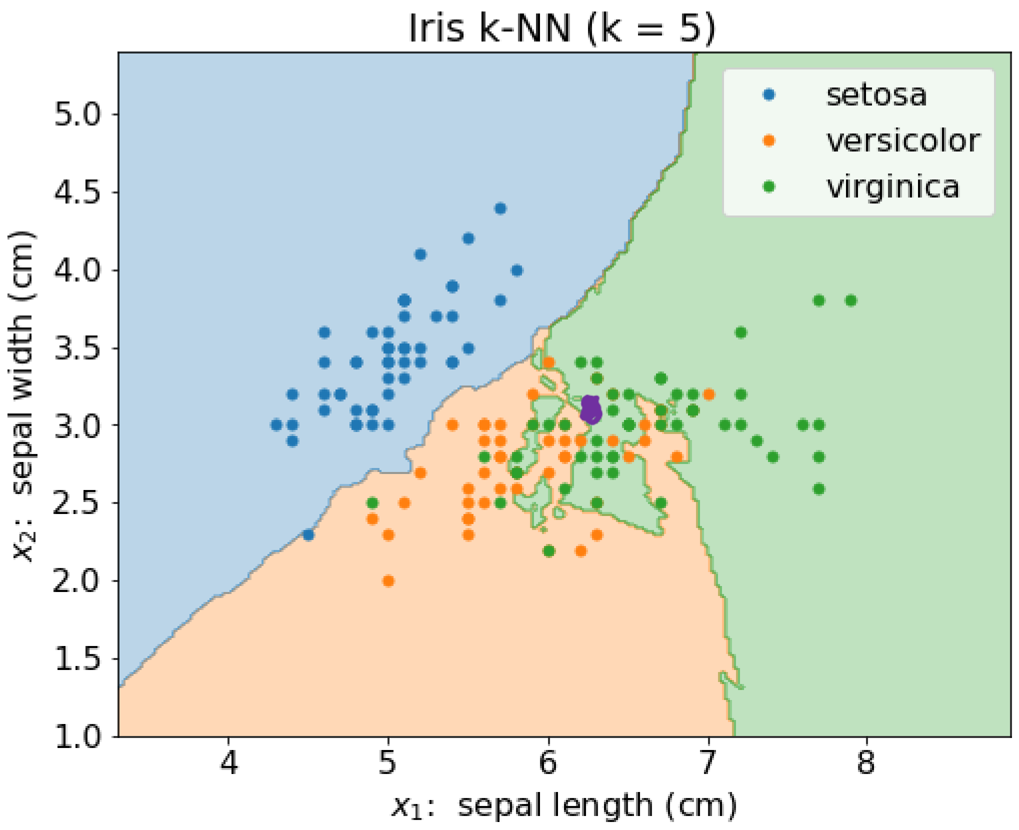
Classification Decisions

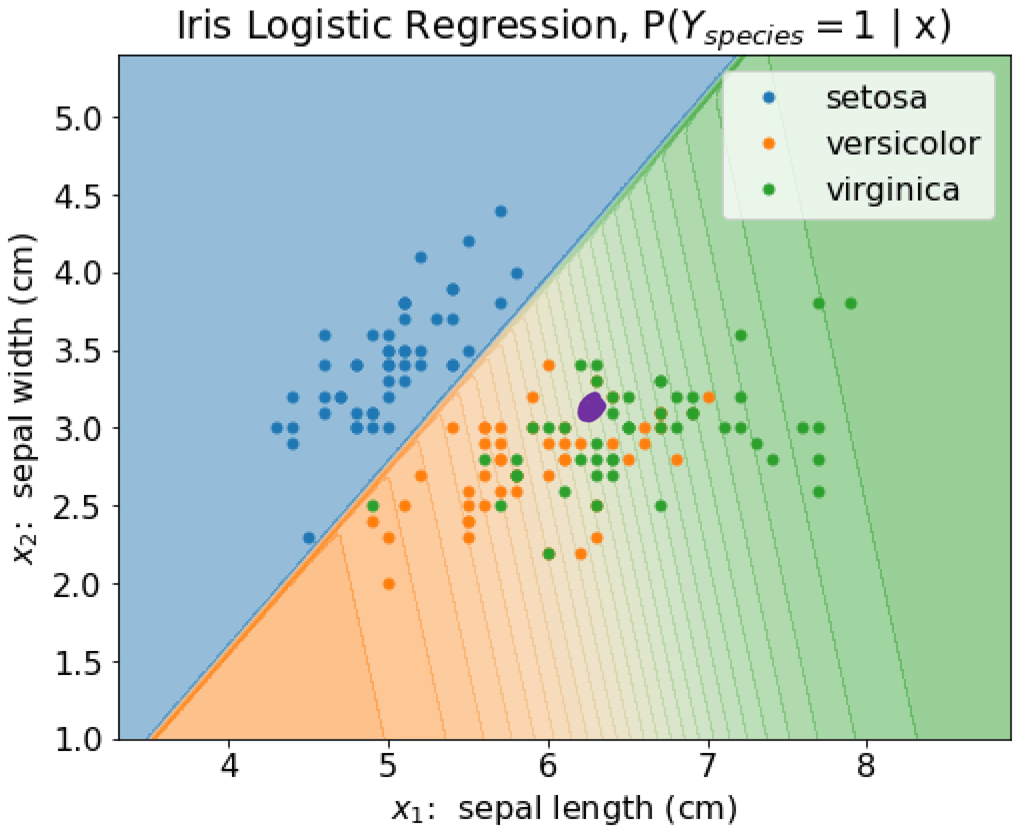
Predicting one specific class is troubling, especially when we know that there is some uncertainty in our prediction



Classification Probability

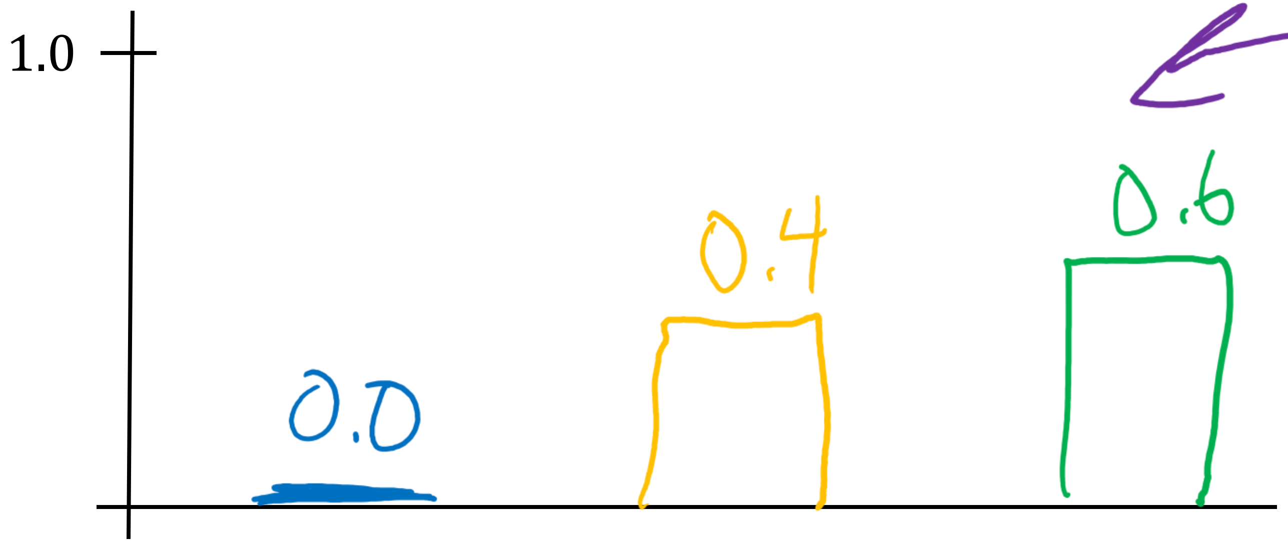
Constructing a model than can return the probability of the output being a specific class could be incredibly useful



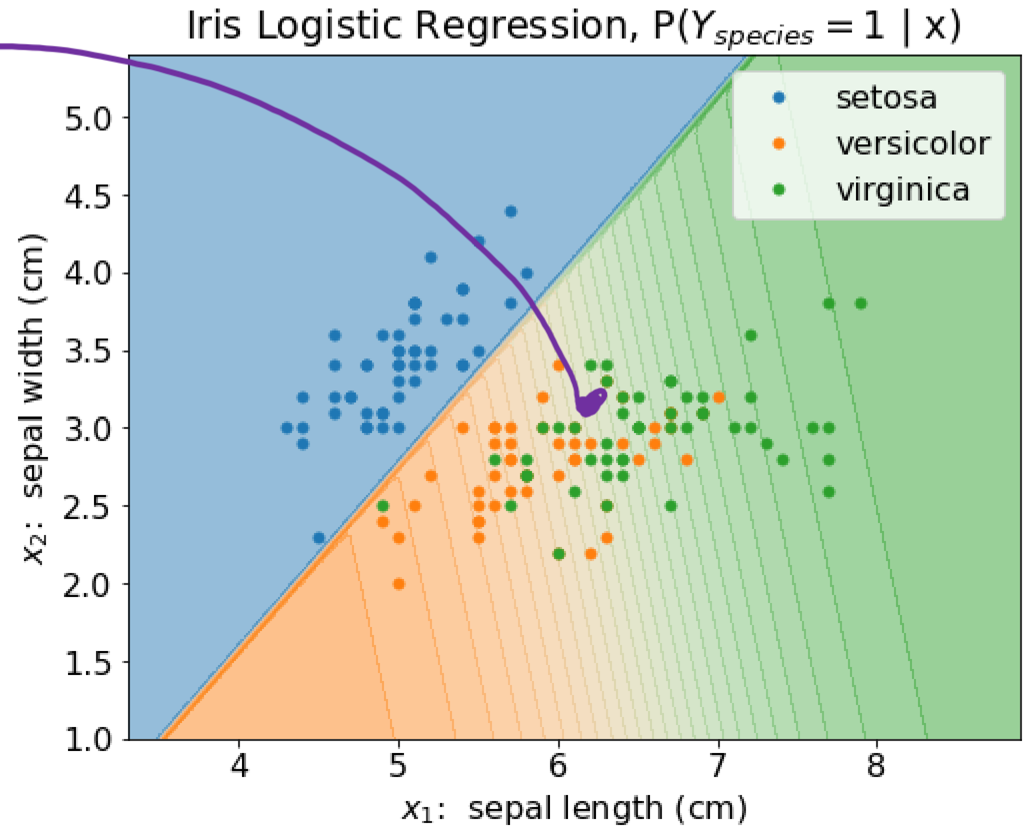


Classification Probability

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



$P(Y_{setosa} = 1 \mid \mathbf{x})$ $P(Y_{vers} = 1 \mid \mathbf{x})$ $P(Y_{virg} = 1 \mid \mathbf{x})$



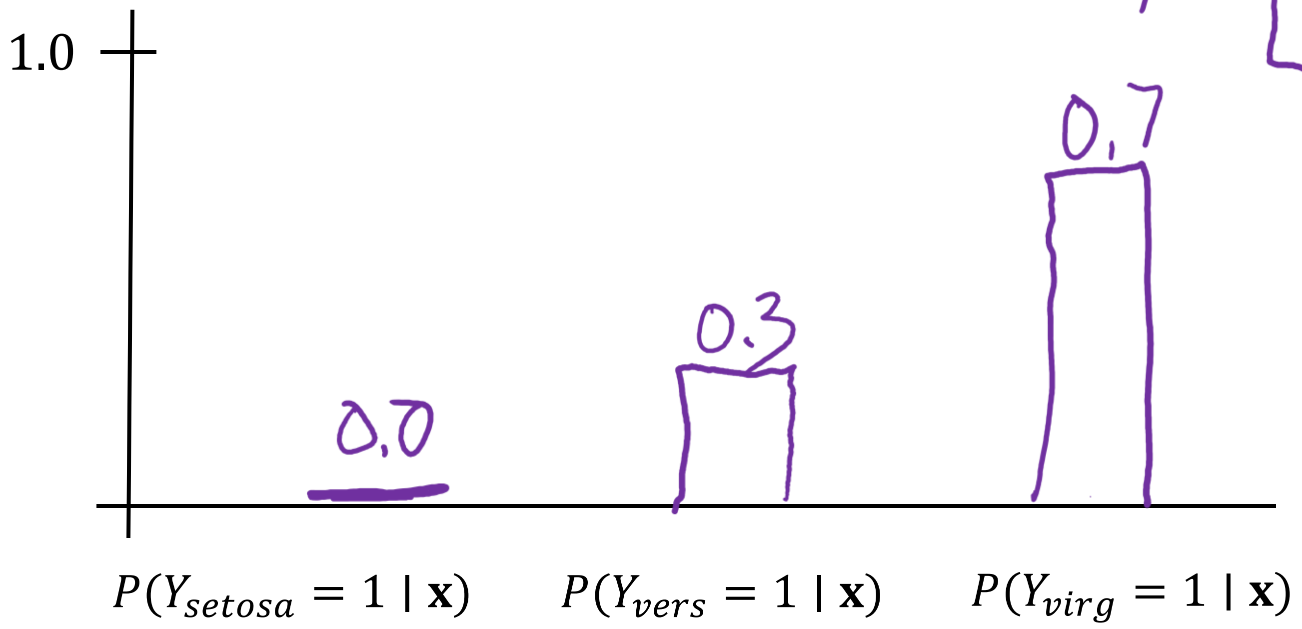
We can still make decisions, .e.g,

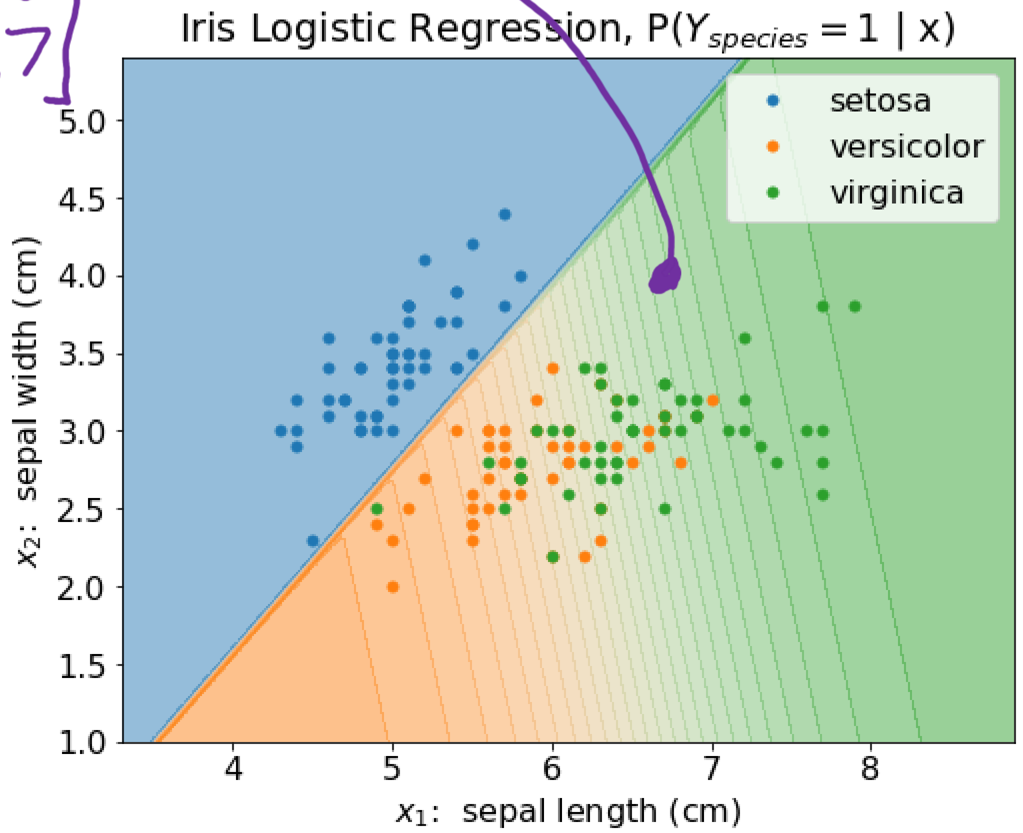
$$\underset{k}{\operatorname{argmax}} P(Y_k = 1 \mid \mathbf{x})$$

Loss for Probabilty Disributions

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

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





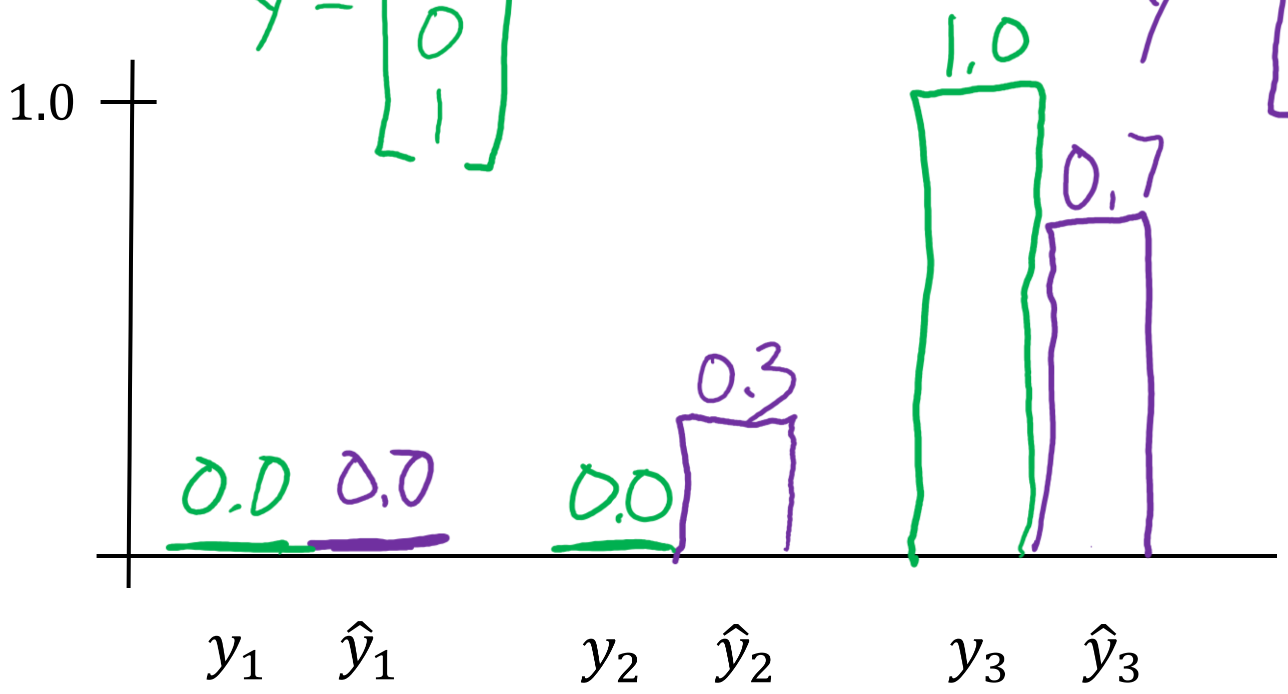
Cross-entropy loss

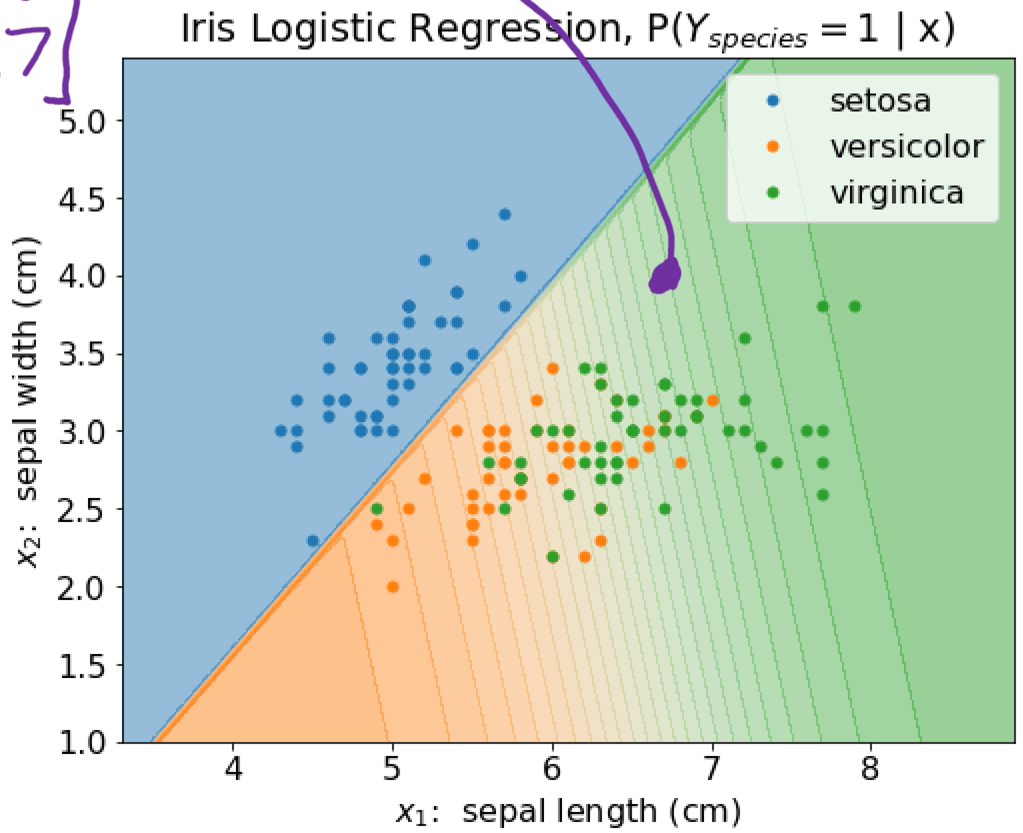
$$\ell(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{k=1}^{K} y_k \log \hat{y}_k$$

Loss for Probabilty Disributions

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

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





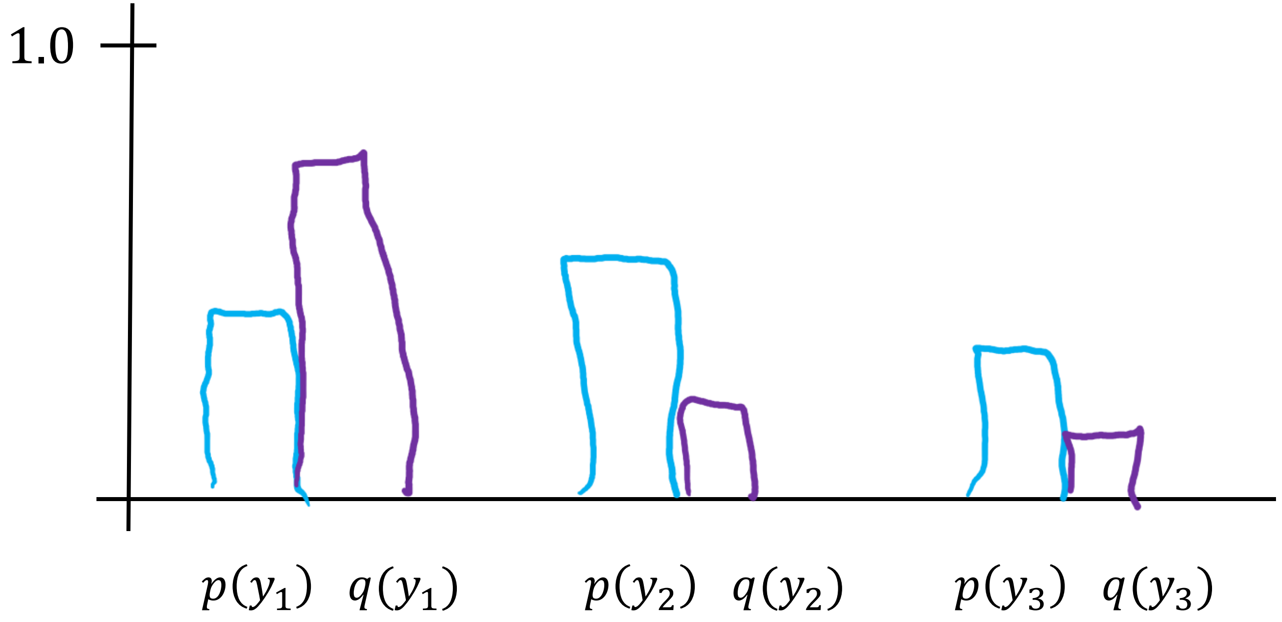
Cross-entropy loss

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

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

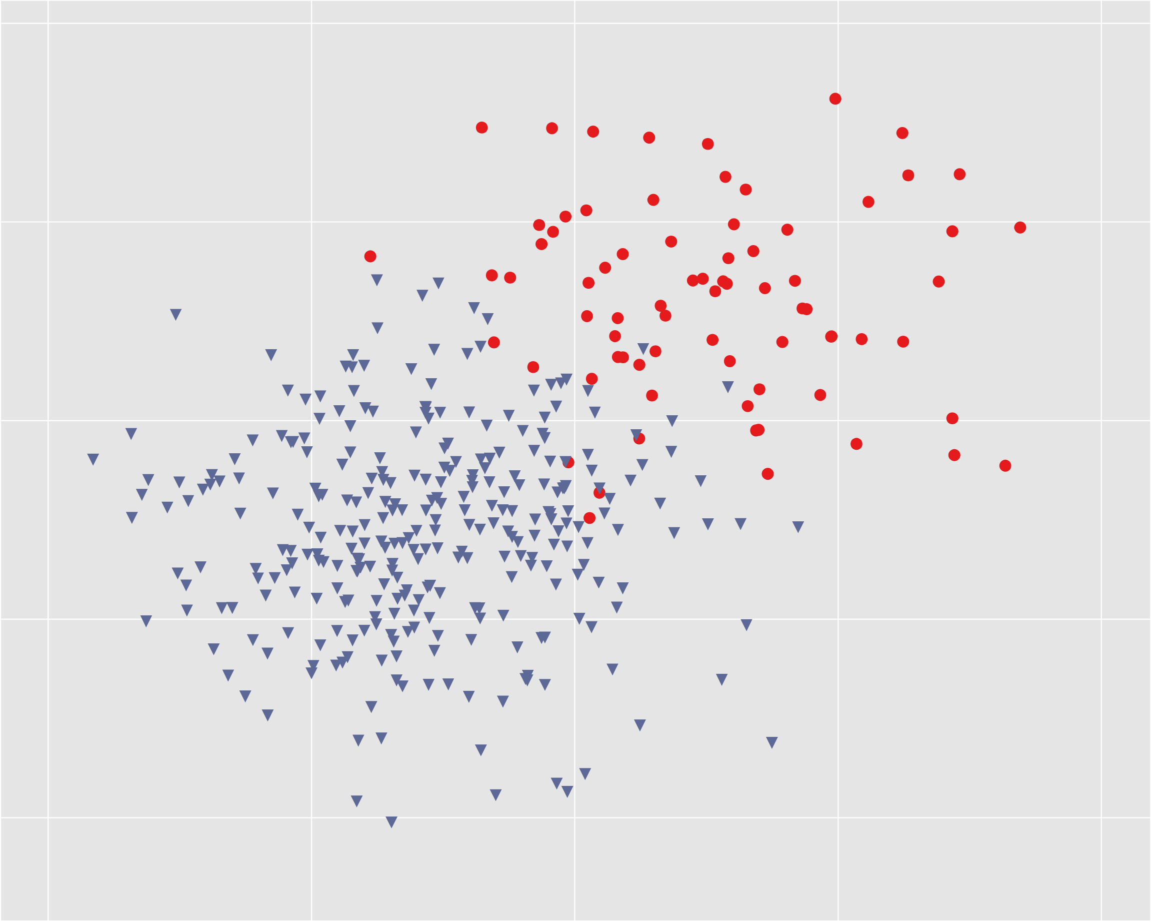
$$H(P, Q) = -\sum_{k=1}^{K} p(y_k) \log q(y_k)$$

Pre-reading

Linear model for classification

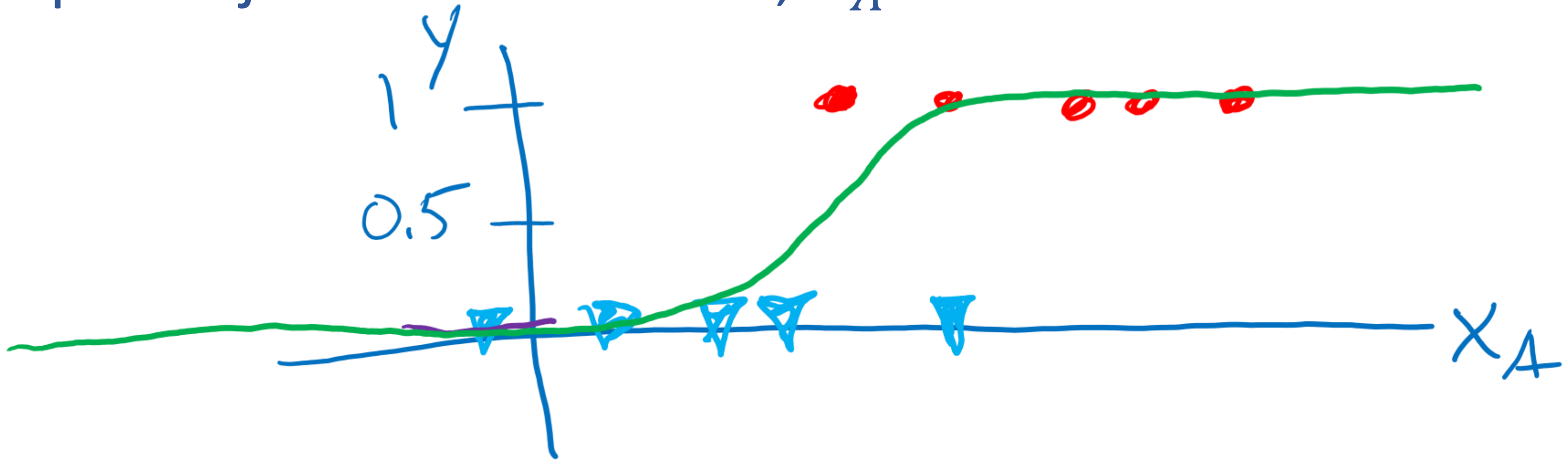
Prediction for Cancer Diagnosis

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

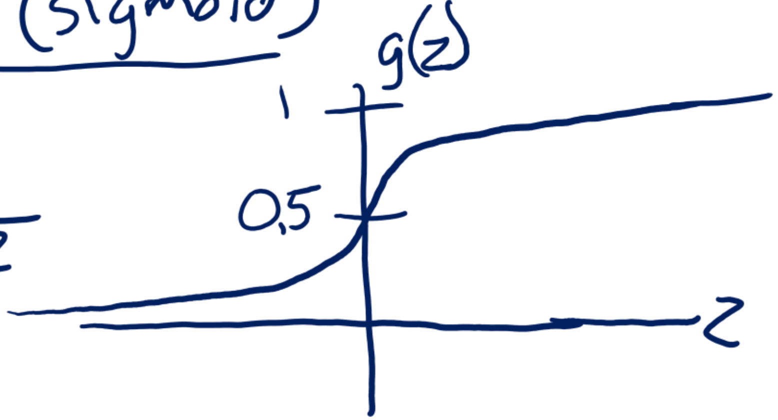
Learn to predict if a patient has cancer ($Y = 1$) or not ($Y = 0$) given the input of just one test result, X_A.



$p(Y=1 \mid x_A)$

logistic function (sigmoid)

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

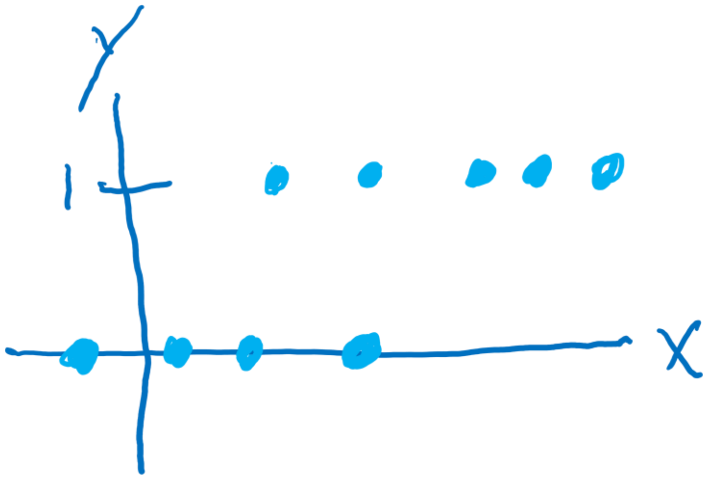


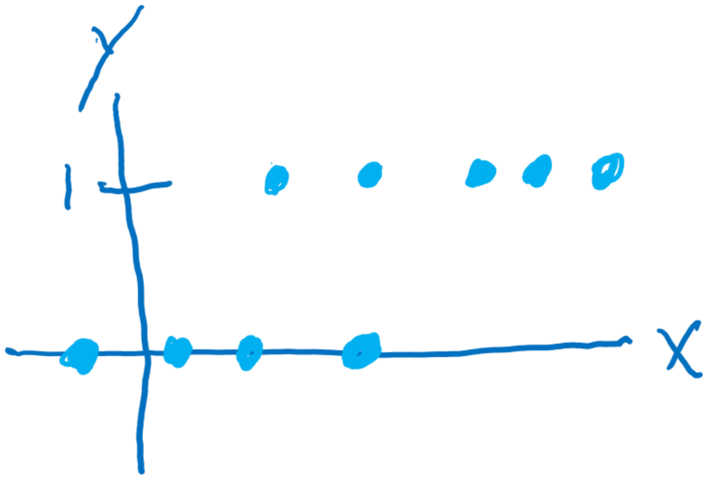
logistic regression

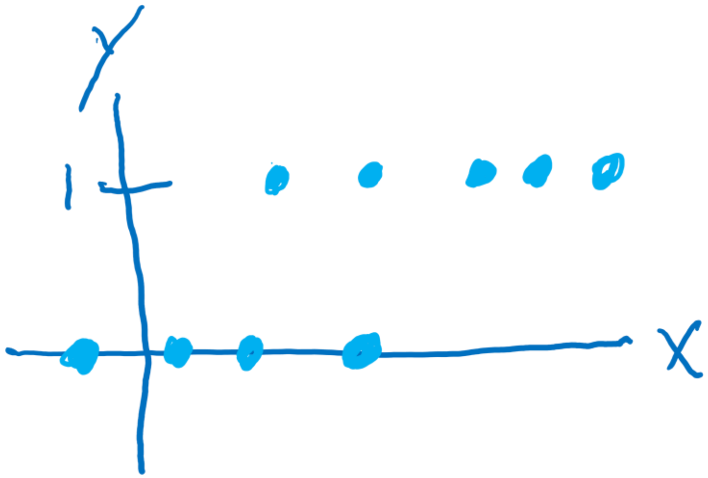
$$p(Y=1 \mid \vec{x}, \vec{\theta}) = g(\theta^T x)$$

Building on a Linear Model

Linear vs Thresholded Linear vs Logistic Linear

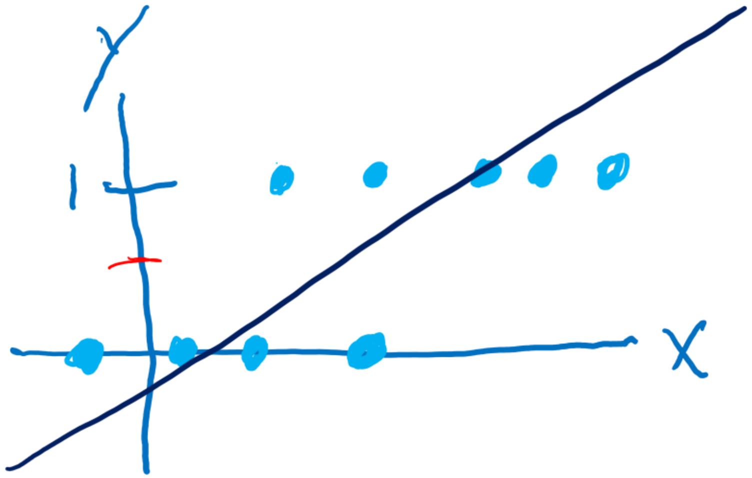


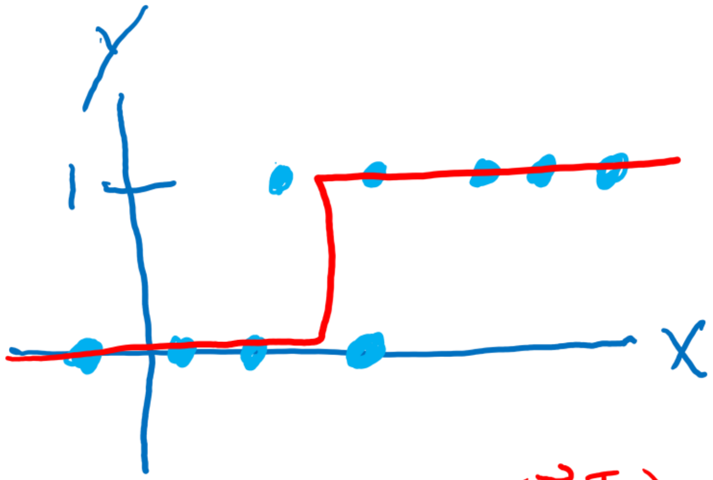


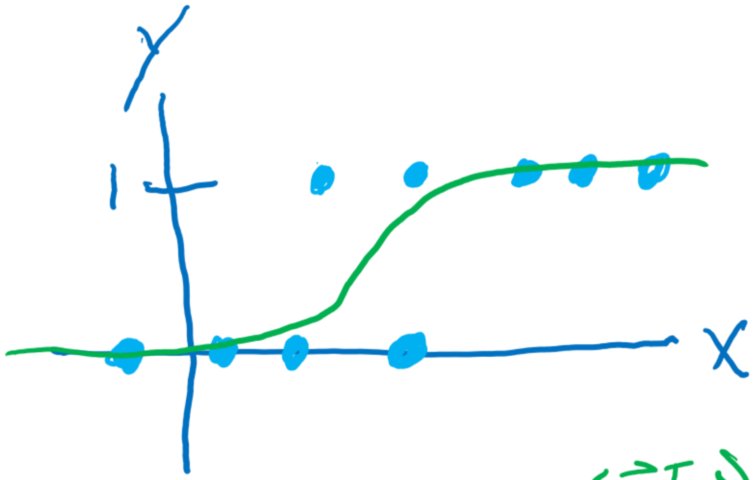


Building on a Linear Model

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

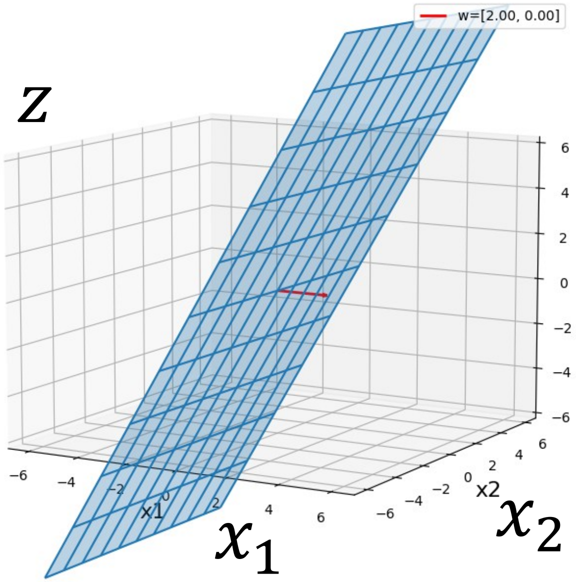
(now with 2+ input features)

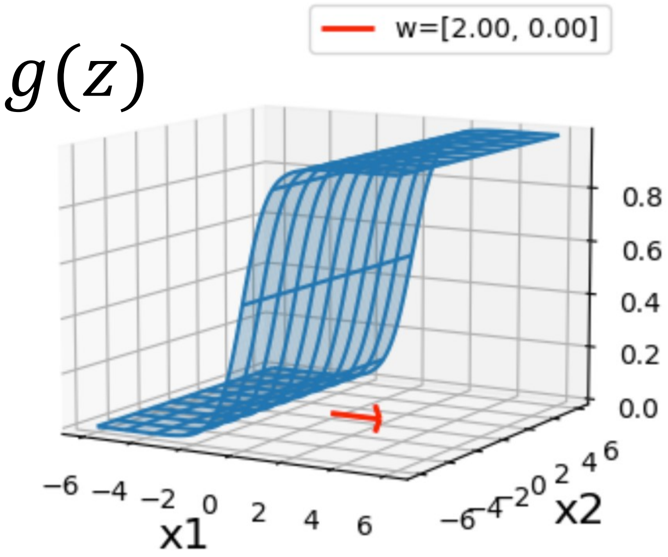
Building on a Linear Model

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:

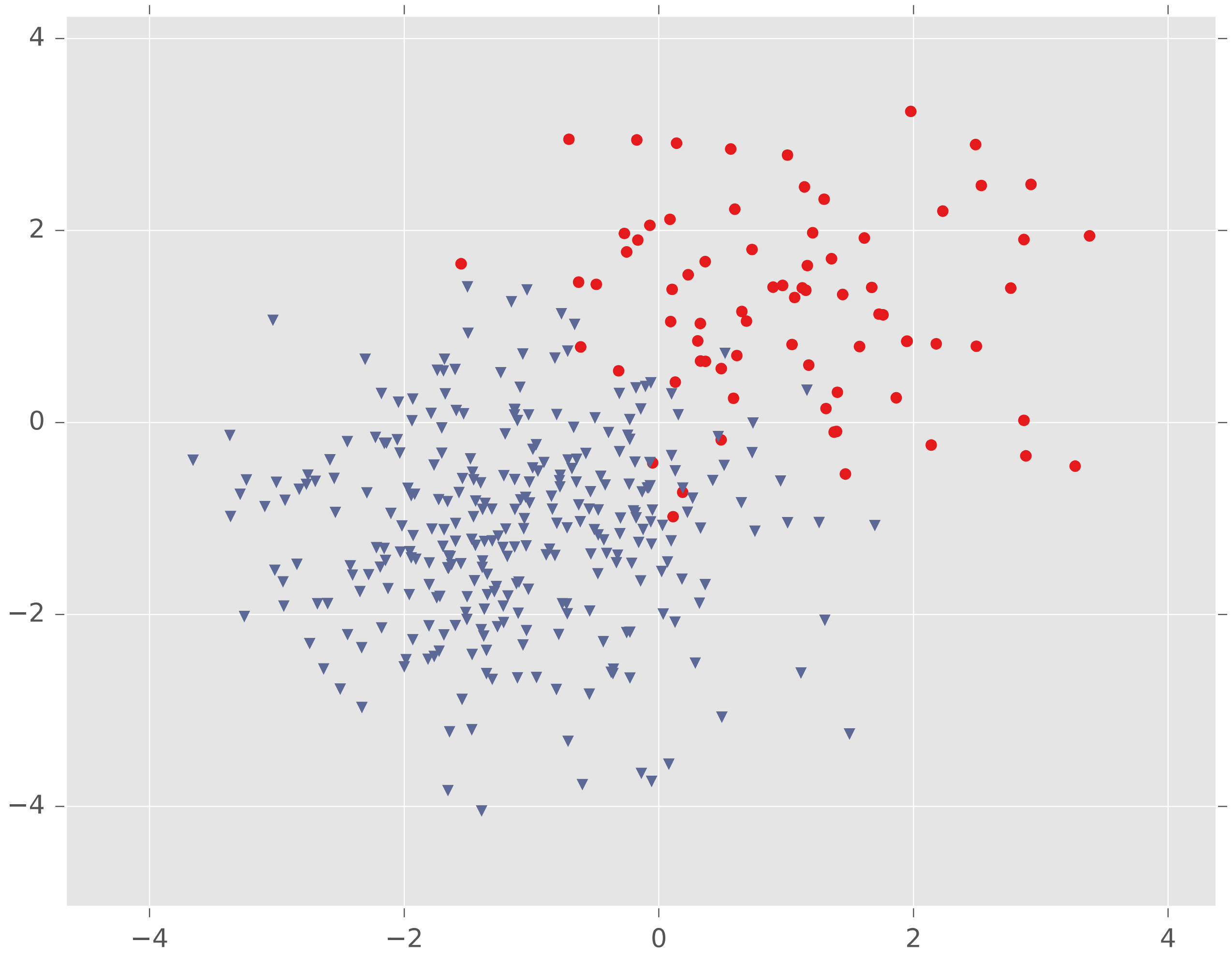
$$z = \mathbf{w}^\top \mathbf{x} + b$$

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



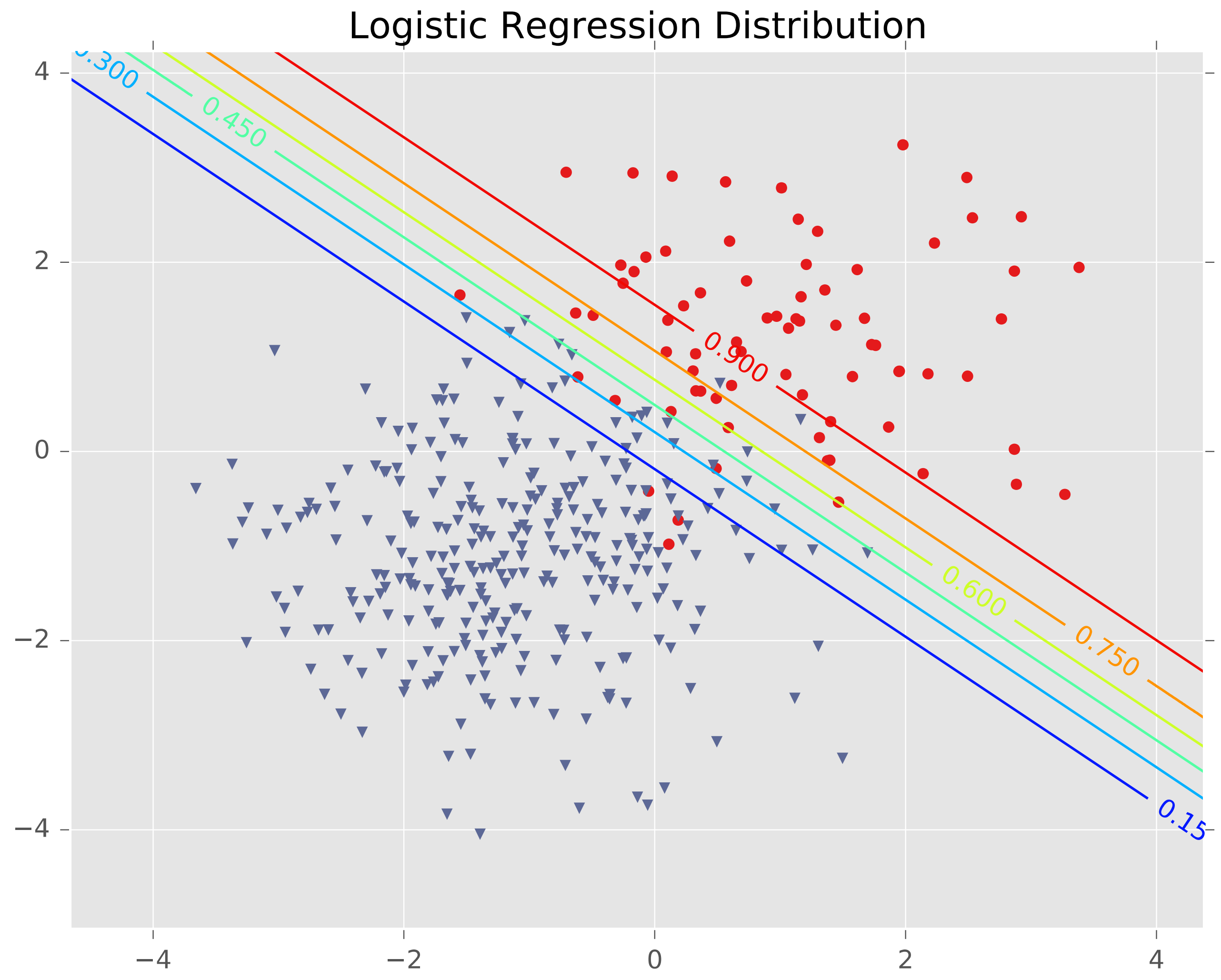


Building on a Linear Model



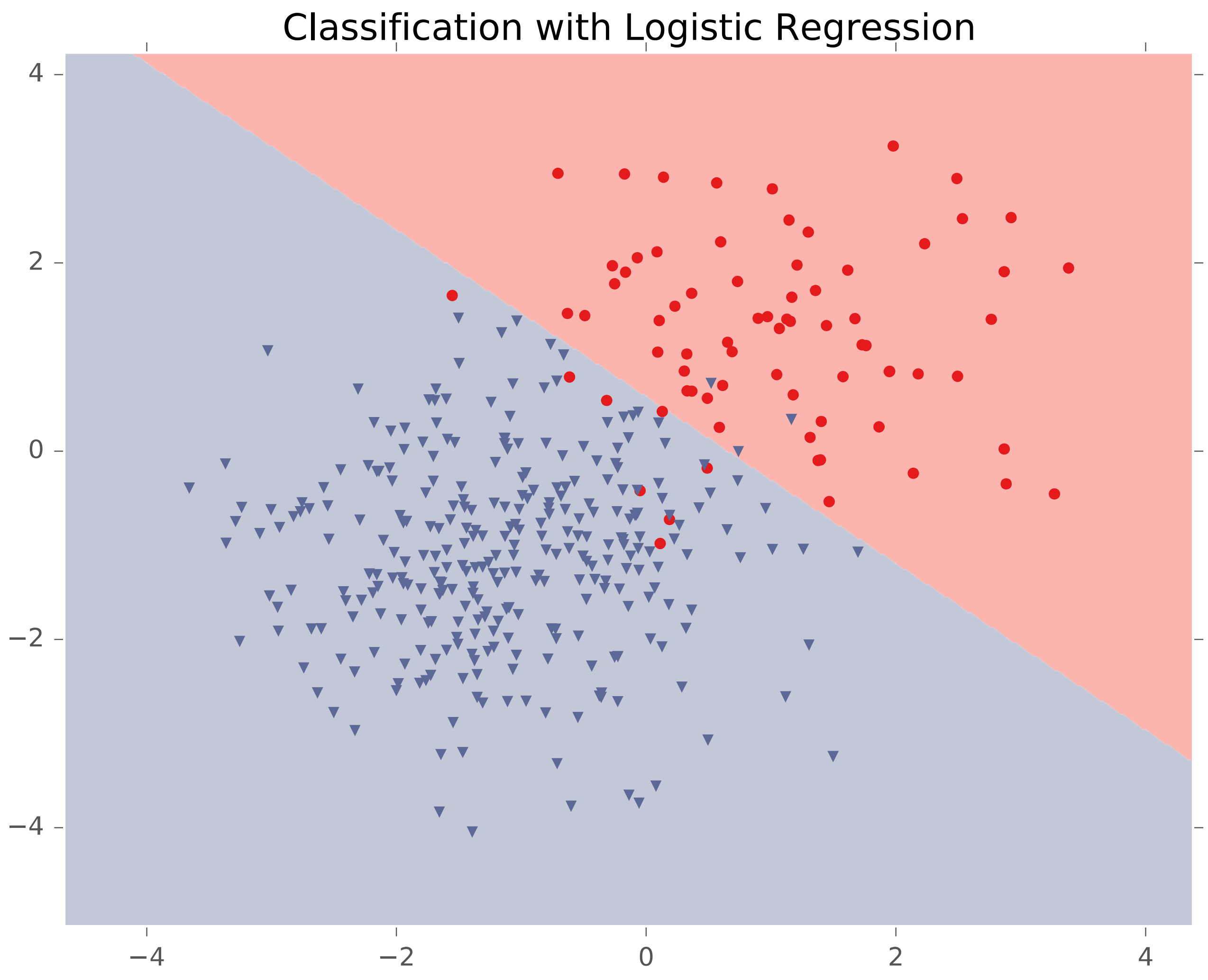
Slide credit: CMU MLD Matt Gormley

Building on a Linear Model



Slide credit: CMU MLD Matt Gormley

Logistic Regression Decision Boundary

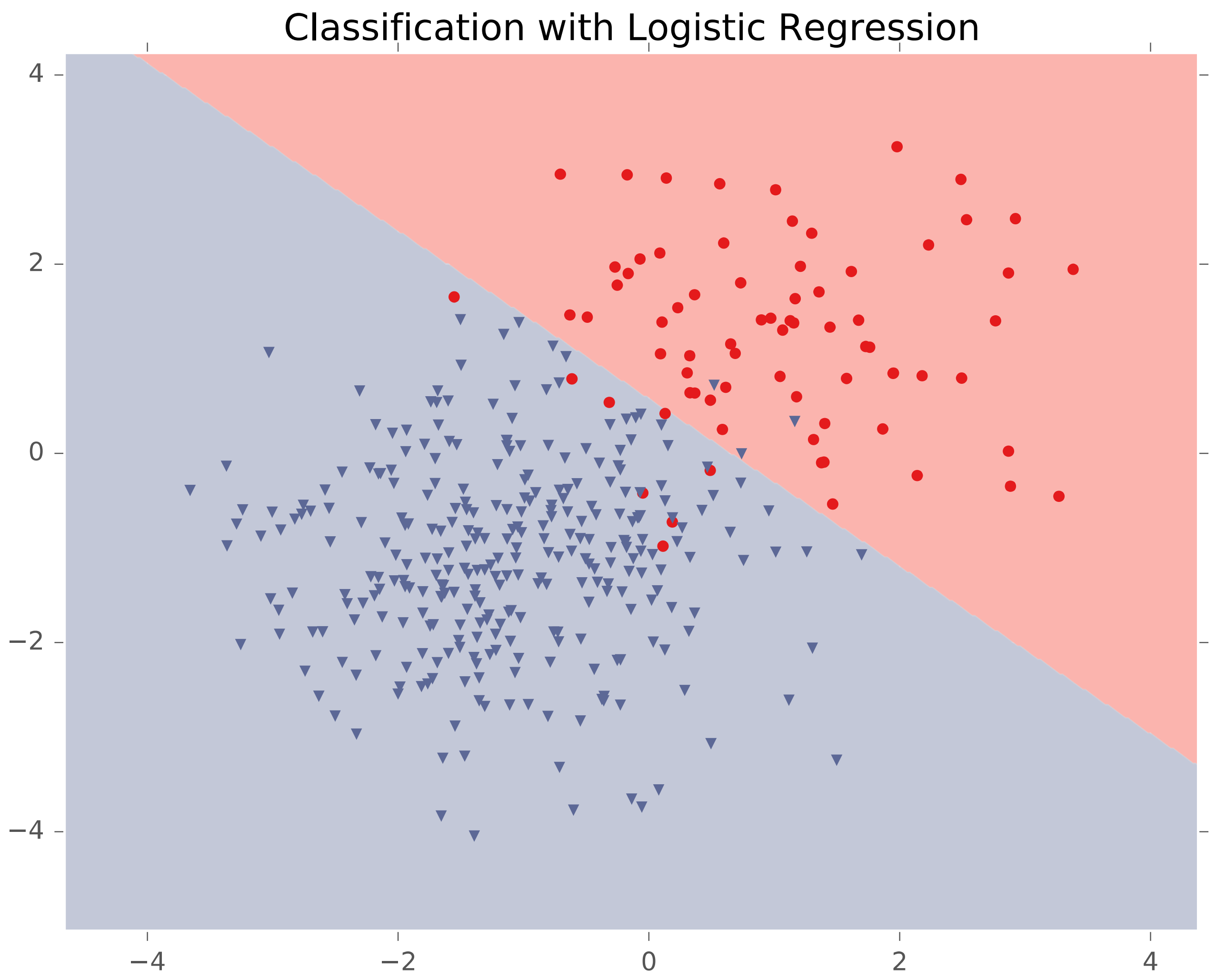


Slide credit: CMU MLD Matt Gormley

Logistic Regression

Linear decision boundary

Logistic Regression Decision Boundary



Exercise

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

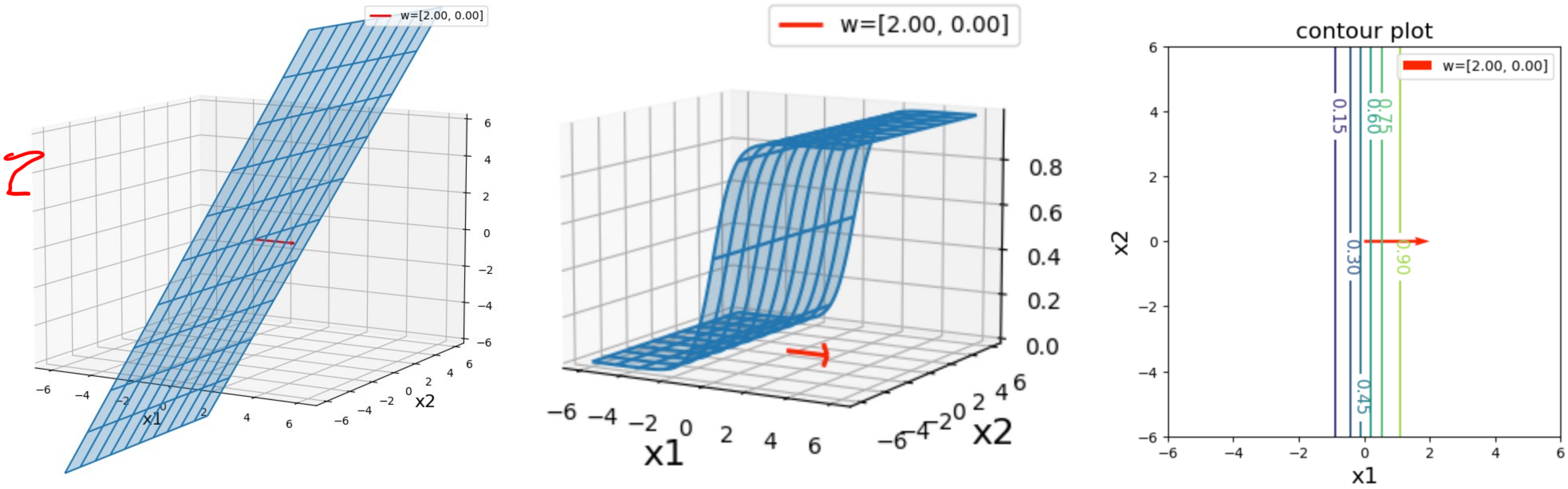
$z = w_1 x_1 + w_2 x_2 + b$

b 0.00

w_magnitude 2.00

w_angle 0.00

0,5



Linear in Higher Dimensions

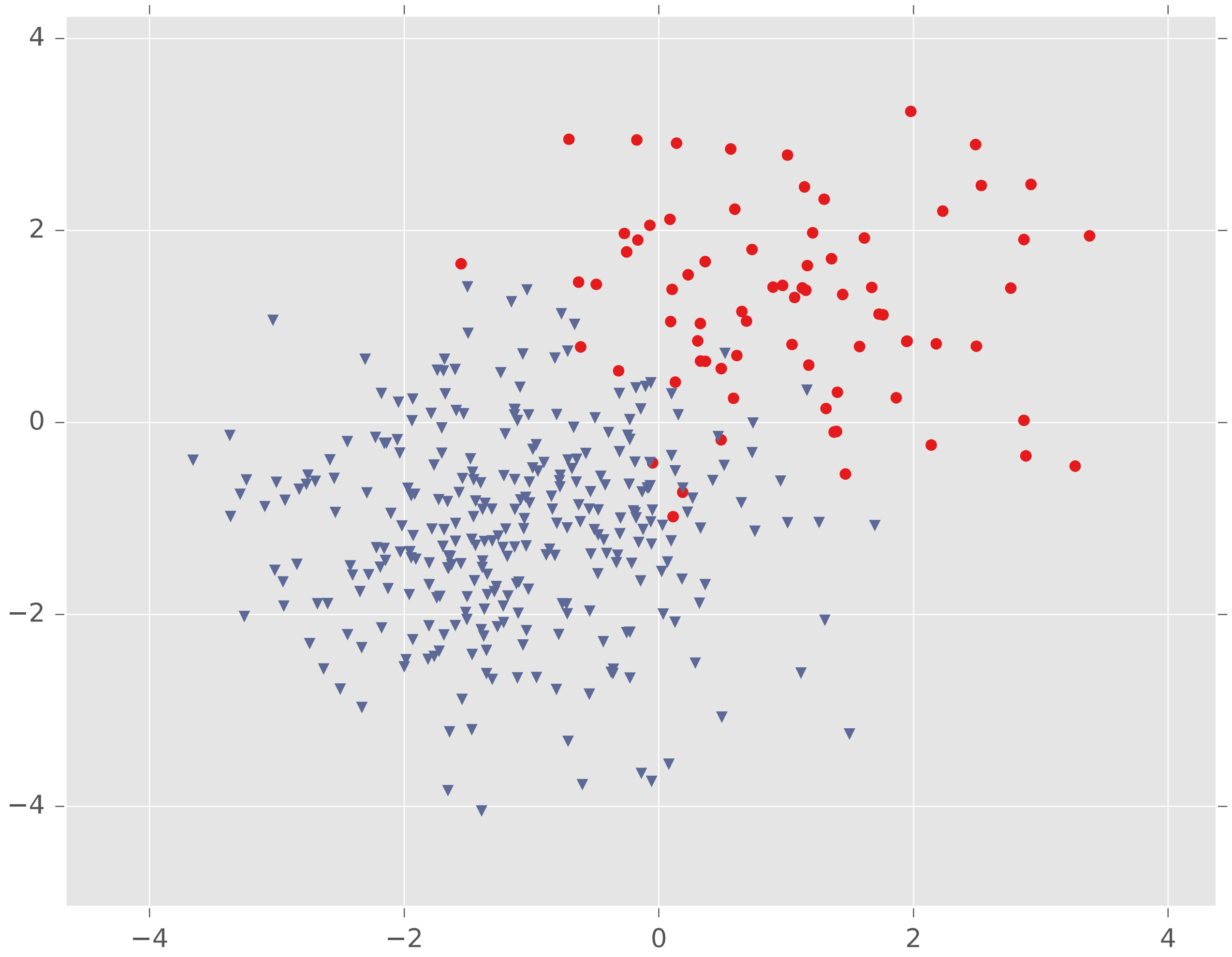
1-D $y = wx + b$

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?

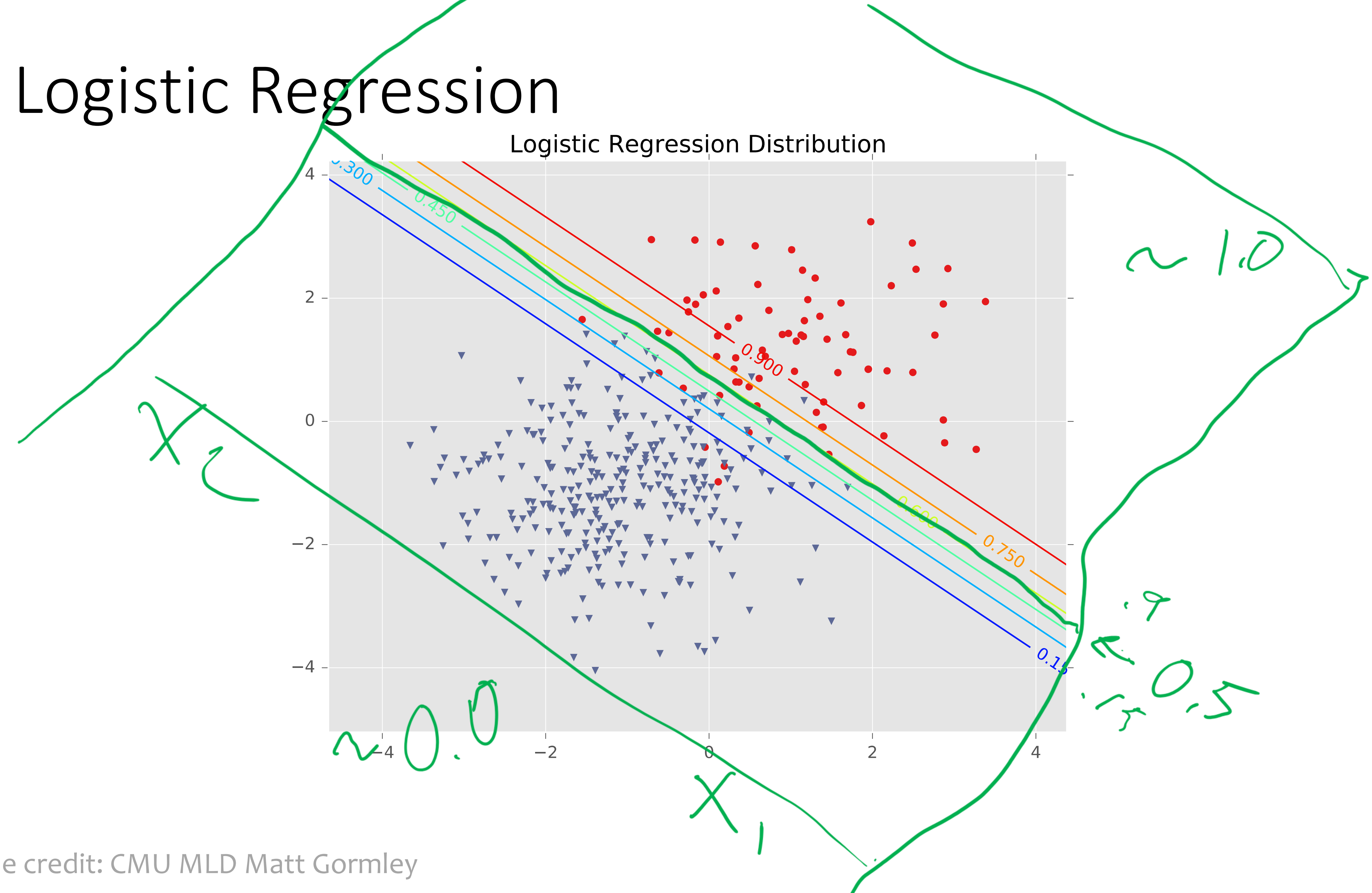
	$x \in \mathbb{R}$	$\mathbf{x} \in \mathbb{R}^2$	$\mathbf{x} \in \mathbb{R}^3$	$\mathbf{x} \in \mathbb{R}^M$
$y = \mathbf{w}^\top \mathbf{x} + b$	line	plane	hyperplane	hyperplane
$\mathbf{w}^\top \mathbf{x} + b = 0$	point	line	plane	hyperplane
$\mathbf{w}^\top \mathbf{x} + b \geq 0$	halfline	halfplane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

Logistic Regression



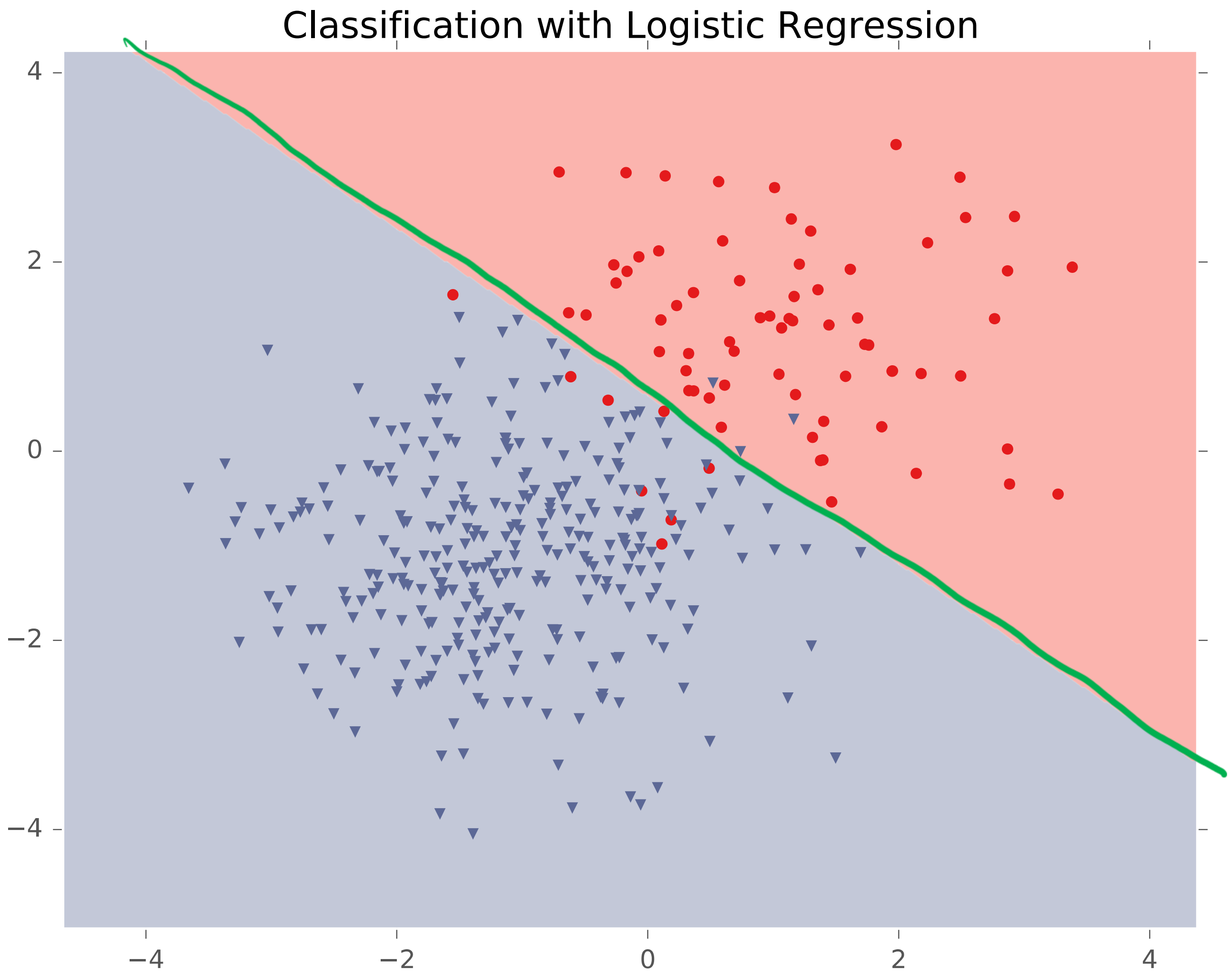
Slide credit: CMU MLD Matt Gormley

Logistic Regression



Slide credit: CMU MLD Matt Gormley

Logistic Regression



Slide credit: CMU MLD Matt Gormley

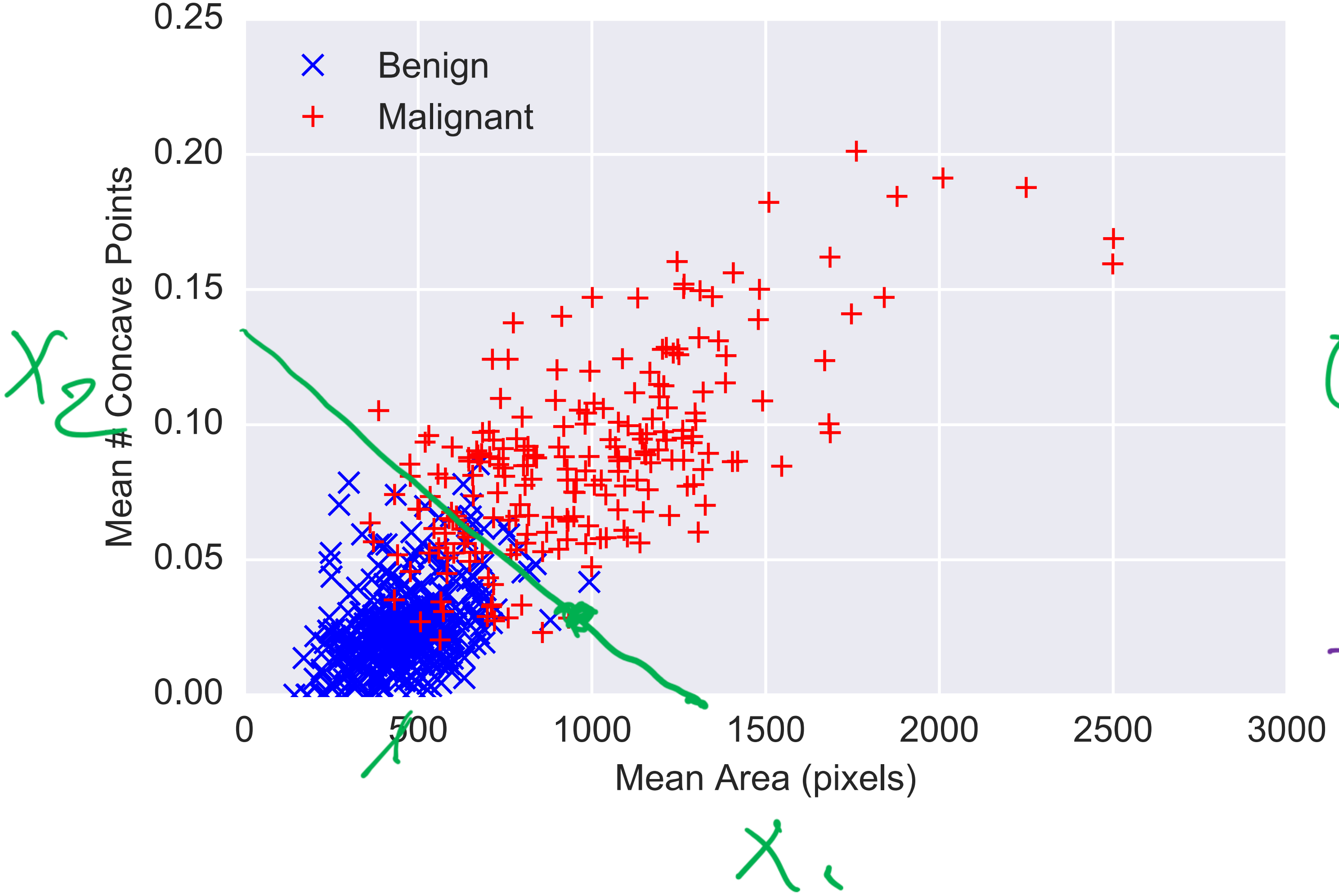
Exercise

Write the equation for this is decision boundary line (in terms of θ).

$w \quad b_1 \quad b_2$

$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix}$

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



$0.5 = g(\vec{\theta}^T \vec{x})$

$0 = \theta^T x$

Poll 1

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

$g(\theta^T x) = \theta^T x$?

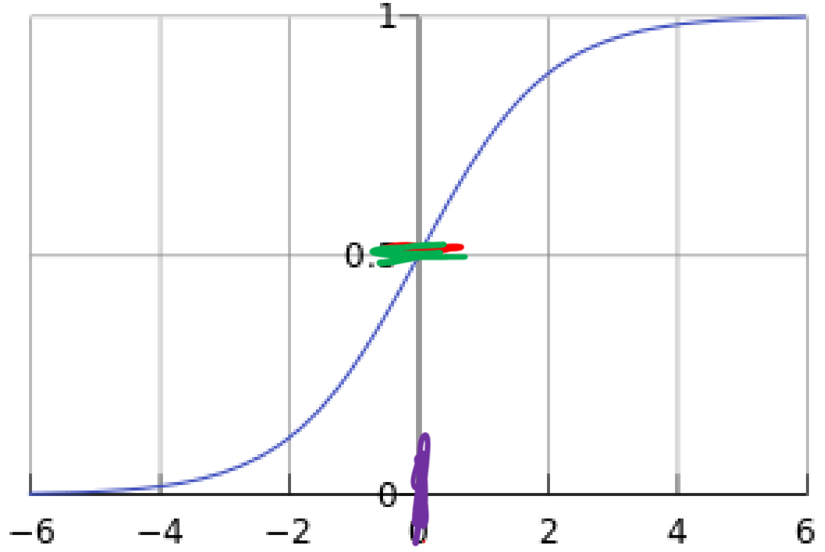
A) Yes

B) No

C) I have no idea

D) I'm not paying attention

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

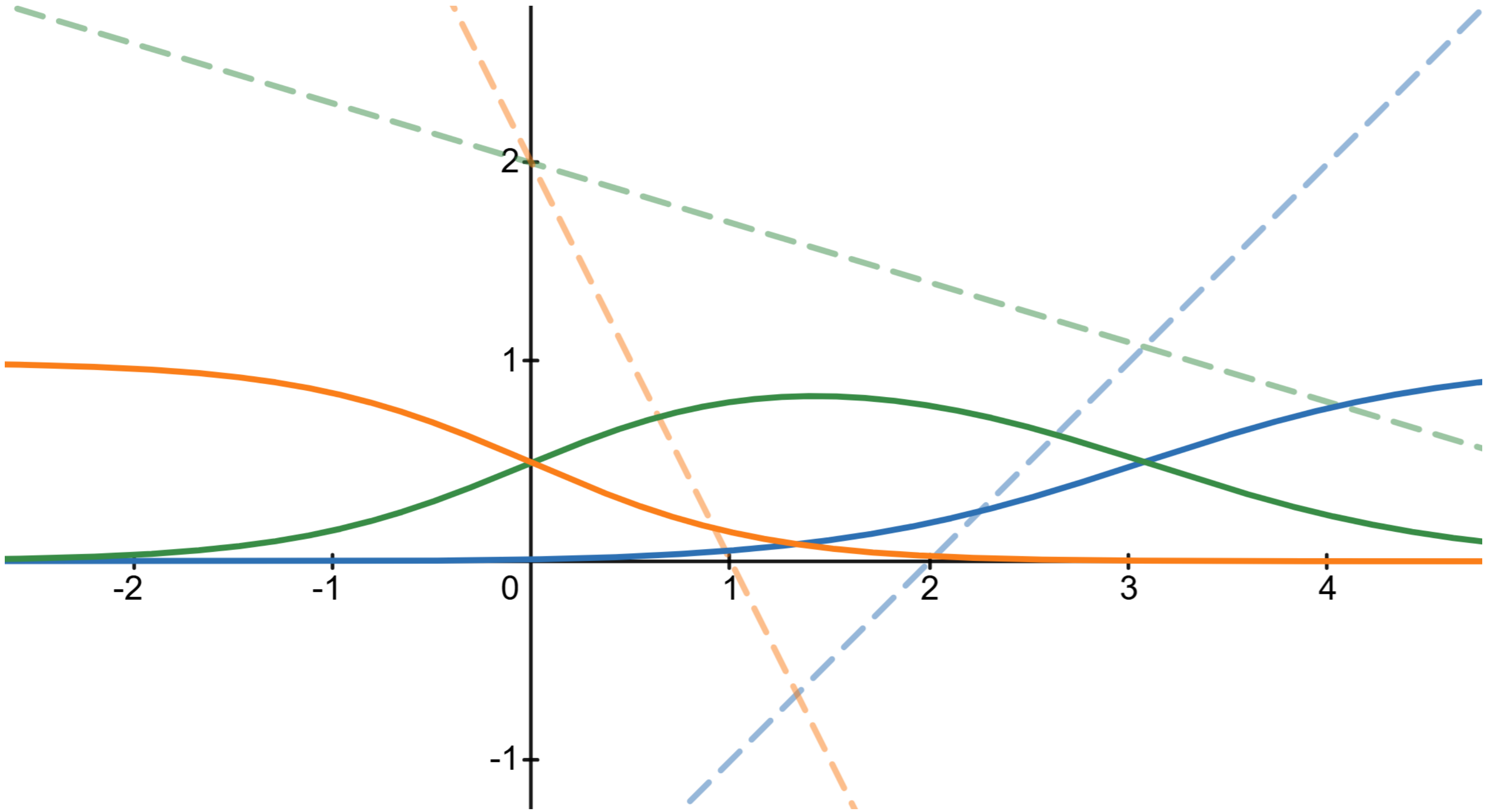


Multi-class Logistic Regression

Multi-class Logistic Regression

Desmos Demo:

https://www.desmos.com/calculator/tu2wg7bhnn



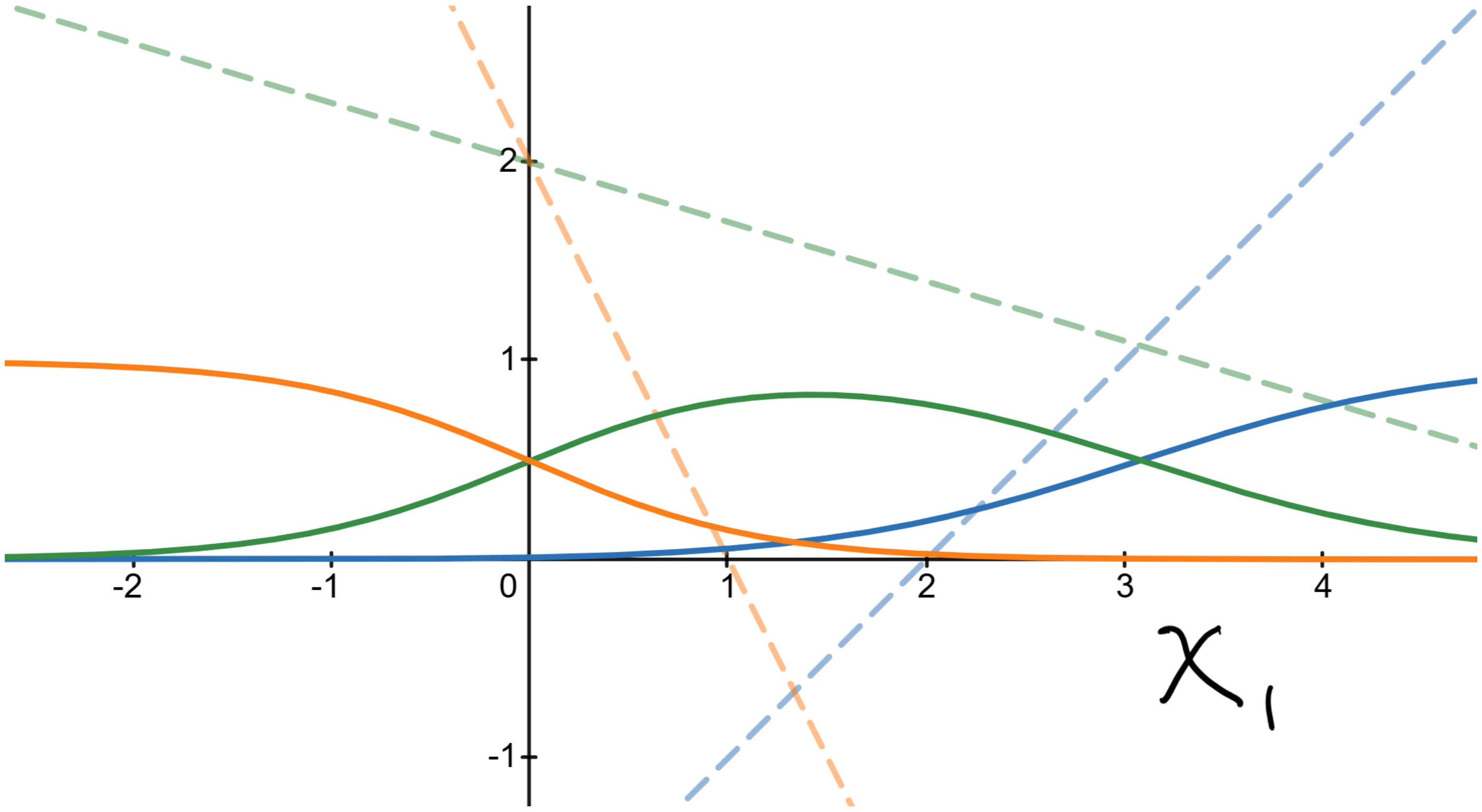
Poll 2

How many parameters are in this model?

3

6

$$\vec{\theta}_1 = \begin{bmatrix} b_1 \\ w_{11} \end{bmatrix}$$

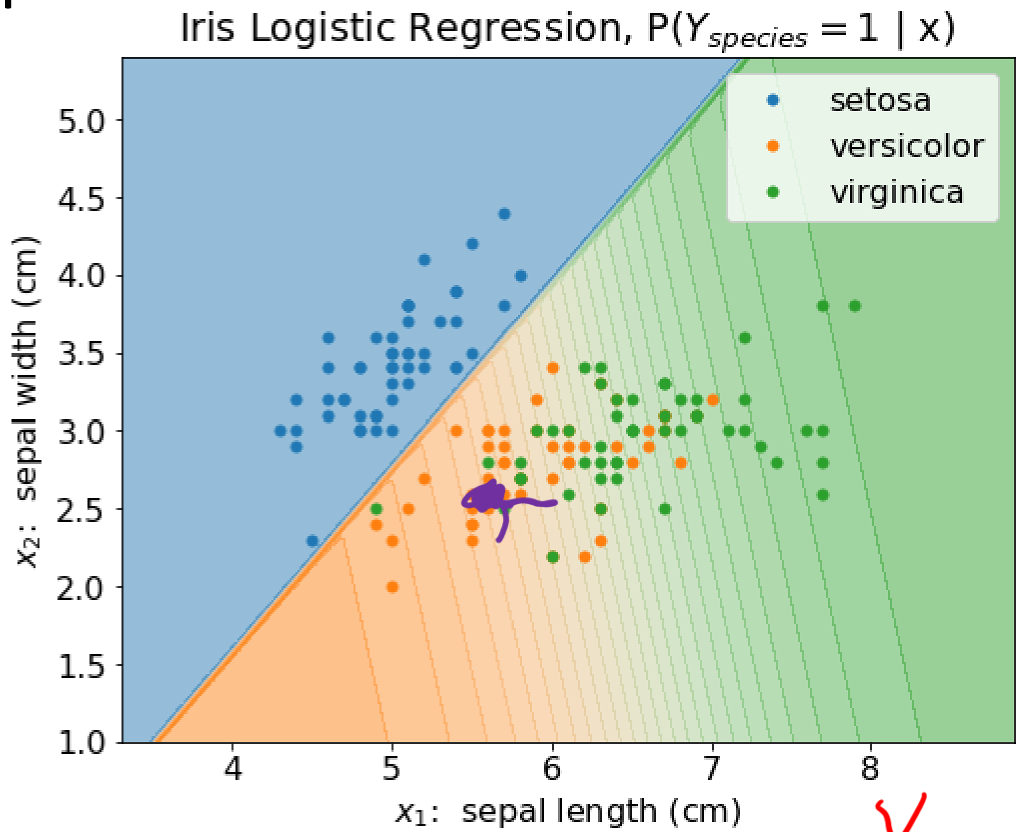


$$\vec{\theta}_2 = \begin{bmatrix} b_2 \\ w_{21} \end{bmatrix}$$

$$\vec{\theta}_3 = \begin{bmatrix} b_3 \\ w_{31} \end{bmatrix}$$

$\vec{y} = h(x)$

Multi-class Logistic Regression



Cross-entropy loss

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

Model

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

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

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

One vector of parameters for each class

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

$y \quad \hat{y}$

$$\mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \qquad \boldsymbol{\theta}_k = \begin{bmatrix} b_k \\ w_{k,1} \\ w_{k,2} \end{bmatrix}$$

$$\Theta = \begin{bmatrix} - & \boldsymbol{\theta}_1^{\top} & - \\ - & \boldsymbol{\theta}_2^{\top} & - \\ - & \boldsymbol{\theta}_3^{\top} & - \end{bmatrix} = \begin{bmatrix} b_1 & w_{1,1} & w_{1,2} \\ b_2 & w_{2,1} & w_{2,2} \\ b_3 & w_{3,1} & w_{3,2} \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0.01 \\ 0.6 \\ 0.39 \end{bmatrix}$$

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

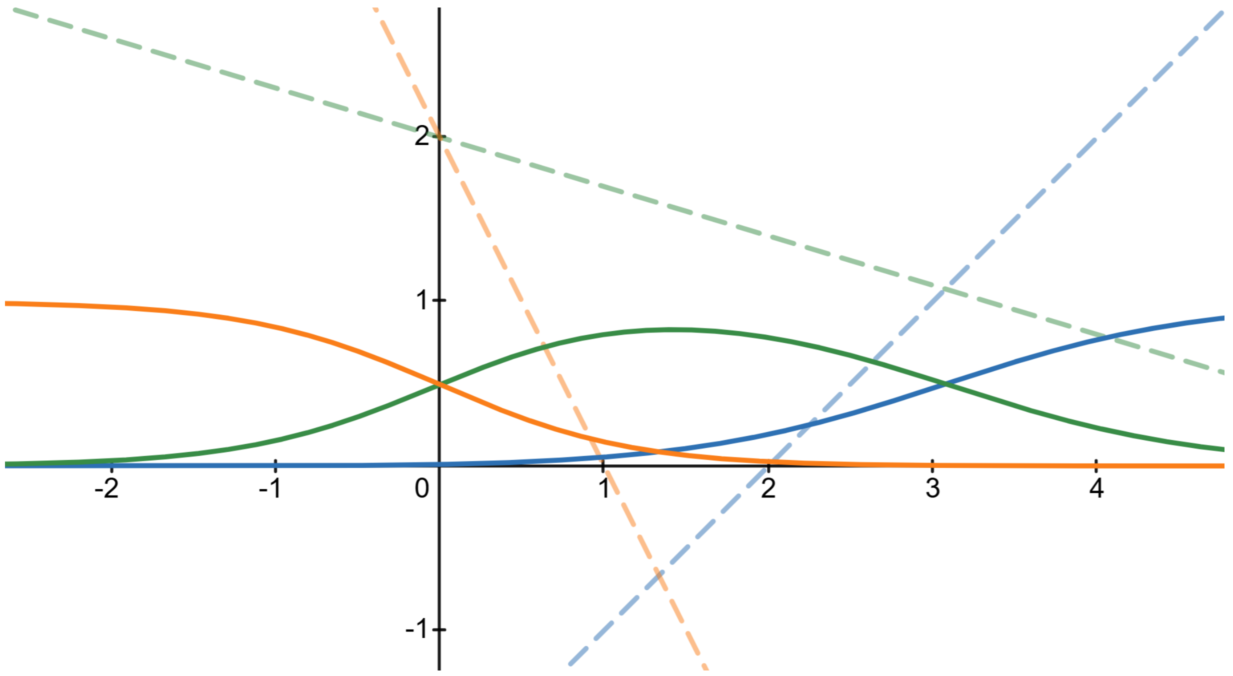
Multiclass Predicted Probability

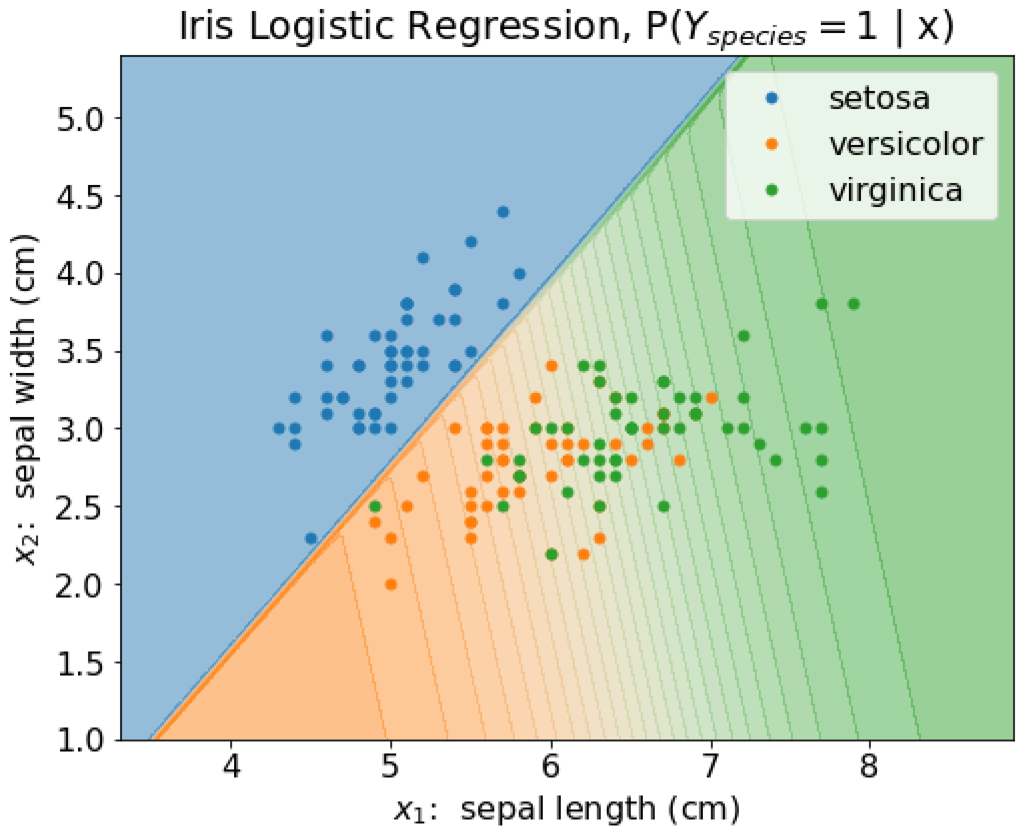
Multiclass logistic regression uses the parameters learned across all K classes to predict the discrete conditional probability distribution of the output Y given a specific input vector $\mathbf{x}$

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

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 $\mathbf{x}$





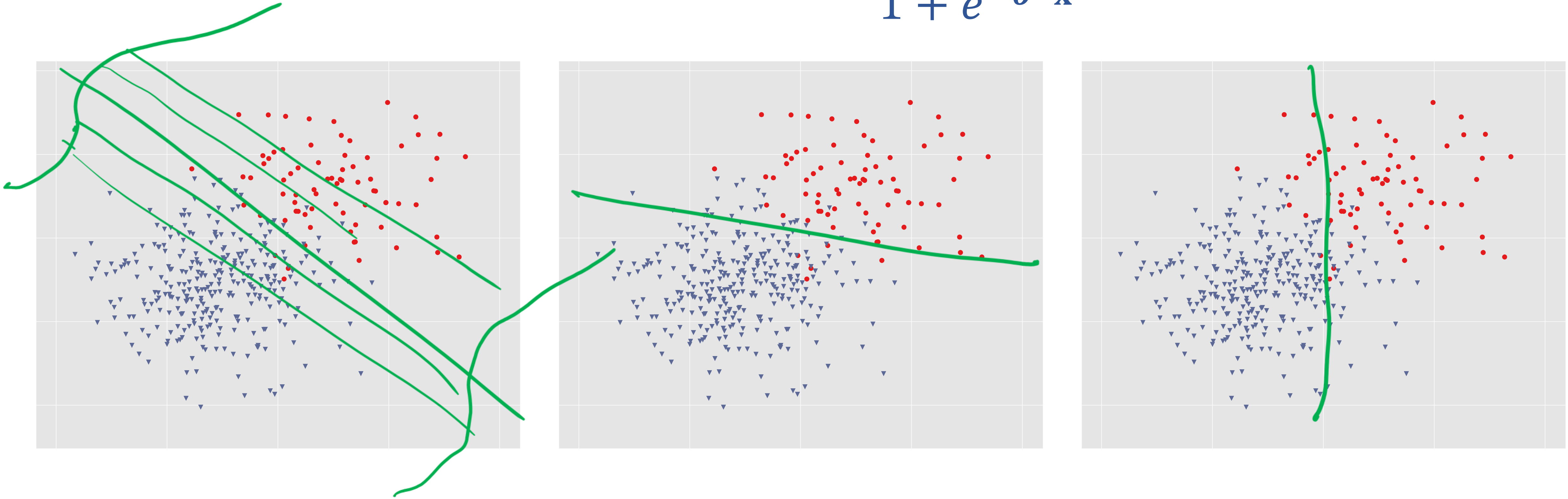
Logistic Regression

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 $\mathbf{x}$.

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

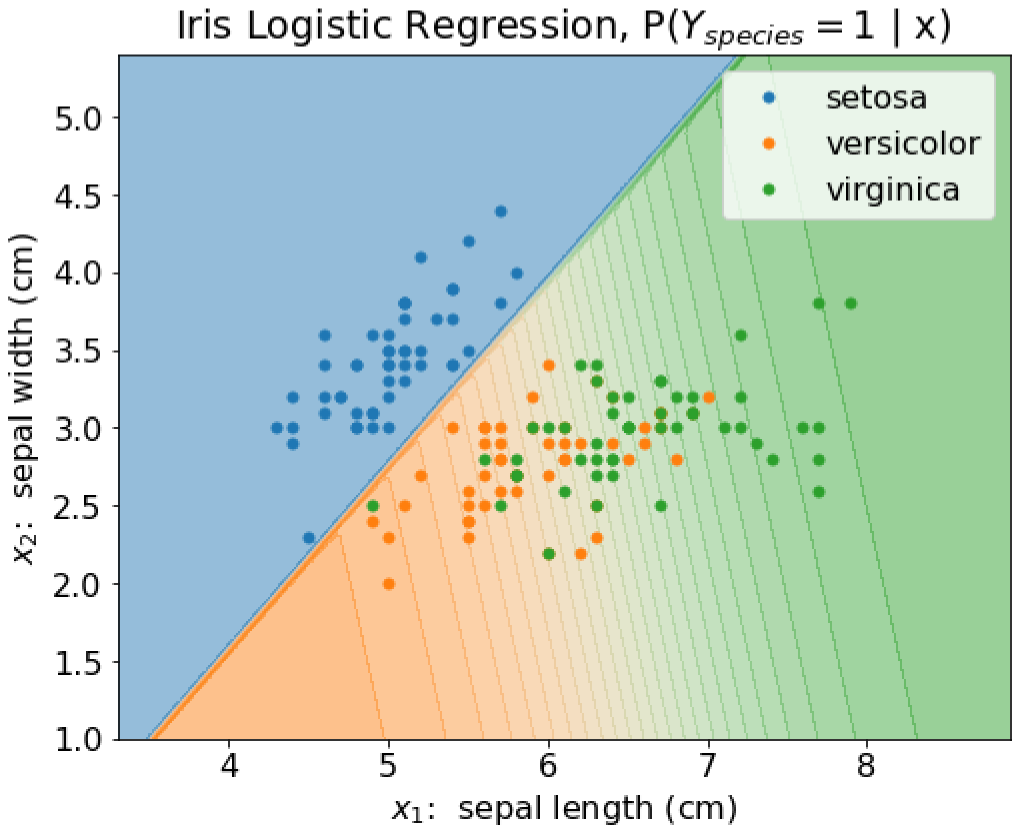


Empirical Risk Minimization

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

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

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



Empirical Risk Minimization

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

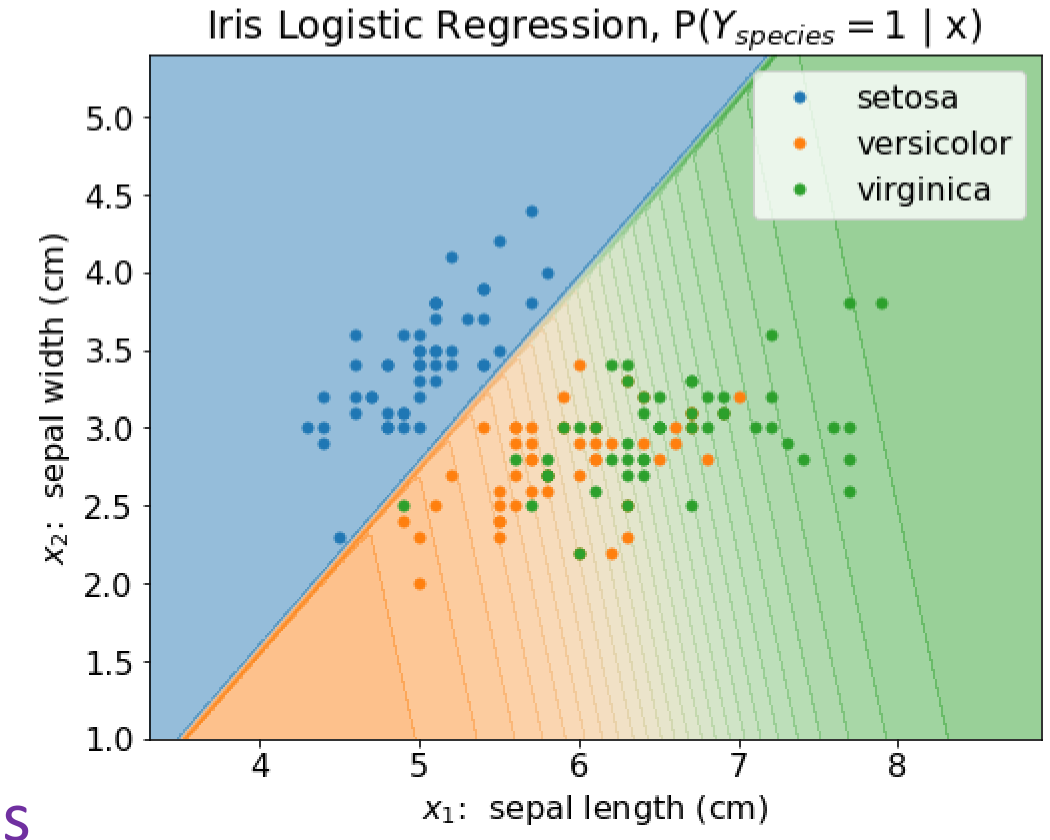
$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\ \hat{R}(h)$$

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

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

$$h(x) = g_{logistic}(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

2) Objective function

$$\ell(\vec{y}^{(i)}, \hat{\vec{y}}^{(i)})$$

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} \bigcirc$$

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

Binary Logistic Regression

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

Objective: Special case for binary logistic regression

$$y_1 = P(y=1|x)$$
$$y_0 = 1 - y_1$$

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

$$J(\boldsymbol{\theta}) = -\frac{1}{N}\sum_i \boxed{\sum_k y_k^{(i)} \log \hat{y}_k^{(i)}} \leftarrow \frac{1}{N}\sum_{i=1}^{N}\left(y_1^{(i)} \log \hat{y}_1^{(i)} + y_0^{(i)} \log \hat{y}_0^{(i)}\right)$$

$$\left(y_1 \log \hat{y}_1 + (1-y_1)\log(1-\hat{y}_1)\right)$$

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

Solve Logistic Regression

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

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

case $y^{(i)} = 1$ $\quad -\left[1 \log \hat{y}^{(i)} + 0 \cdot \log\left(1-\hat{y}^{(i)}\right)\right] = -\log \hat{y}^{(i)}$

case $y^{(i)} = 0$ $\quad -\left[0 \log \hat{y}^{(i)} + 1 \cdot \log\left(1-\hat{y}^{(i)}\right)\right] = -\log\left(1-\hat{y}^{(i)}\right)$

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

Solve Logistic Regression

$z = \theta^T x$

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

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

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

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

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

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

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

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

Solve Logistic Regression

$$z = \theta^T x$$

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

$$\frac{\partial}{\partial u} u^T v = v \text{ or } v^T$$

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

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

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

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

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

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

$$= -\left[y - \cancel{y\hat{y}} - (\hat{y} - \cancel{y\hat{y}})\right]$$

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

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

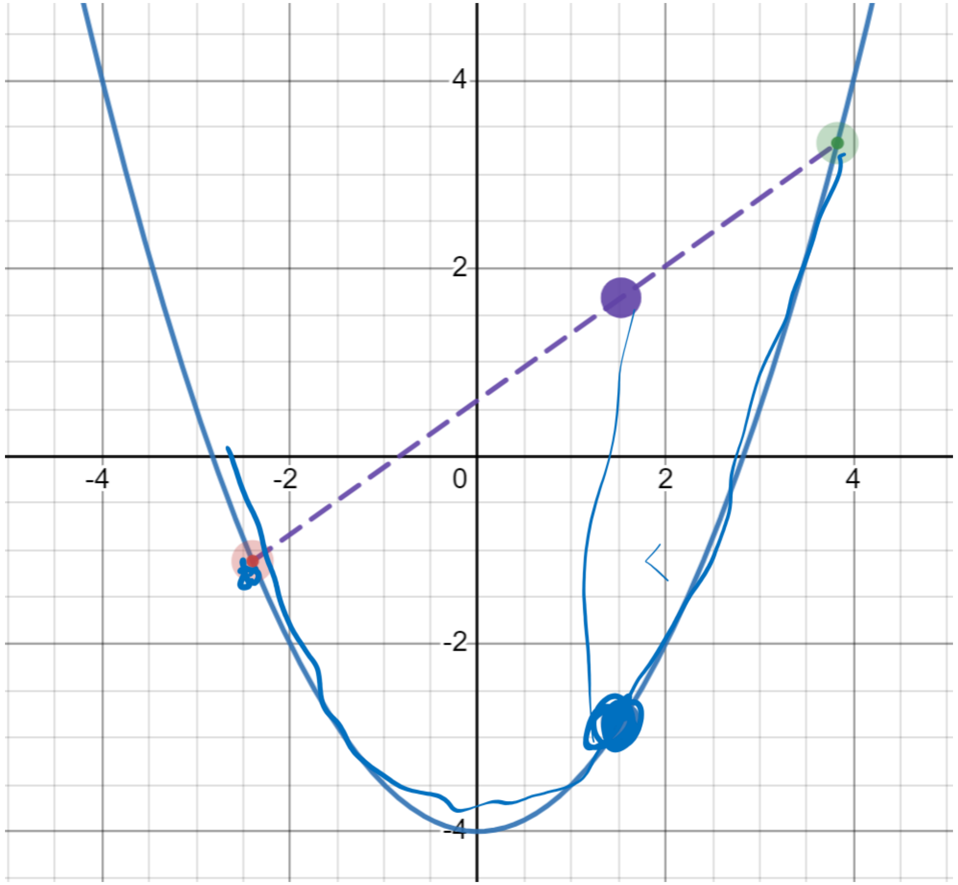
Optimization

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:

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

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

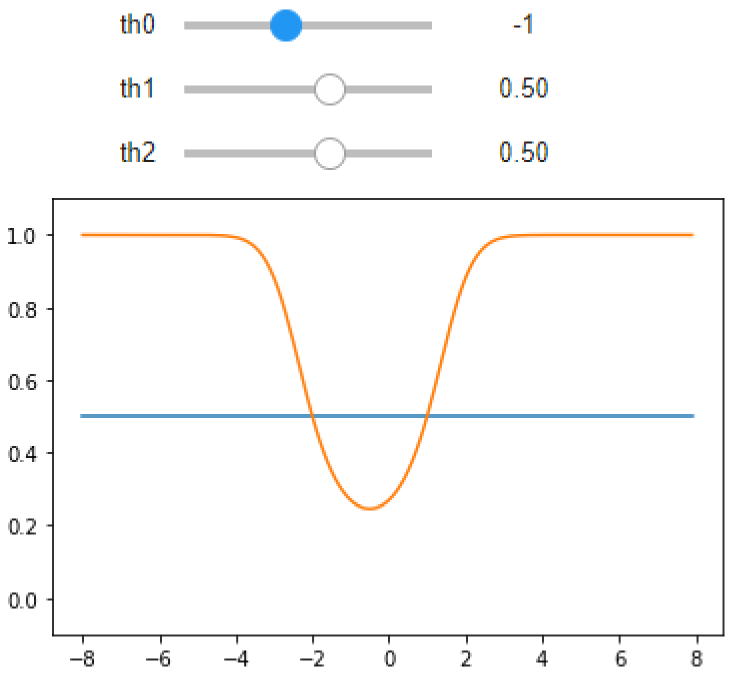
But...what are we optimizing over in logistic regression?

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

Logistic Regression with Polynomial Features

Exercise

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



Exercise

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