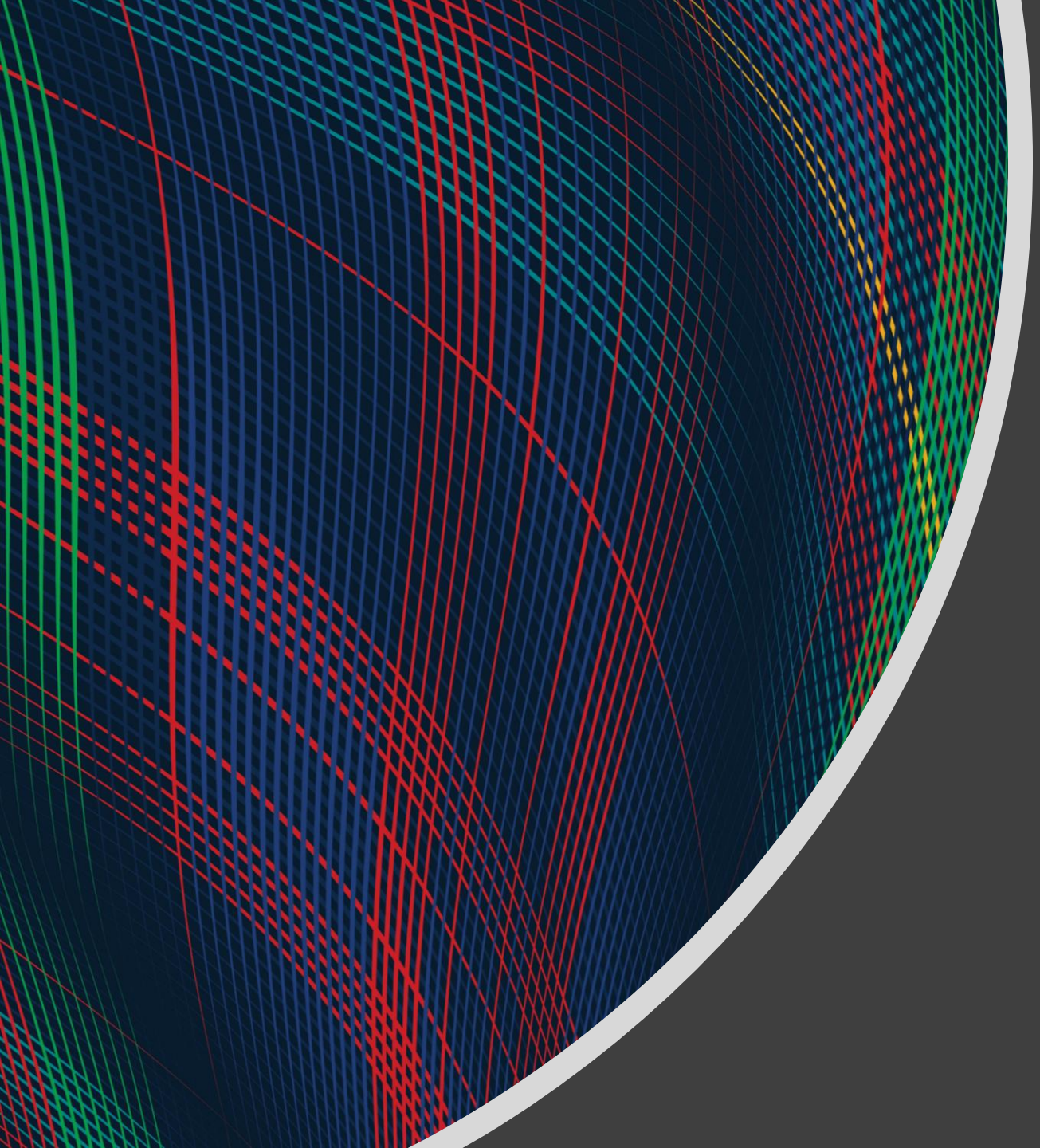


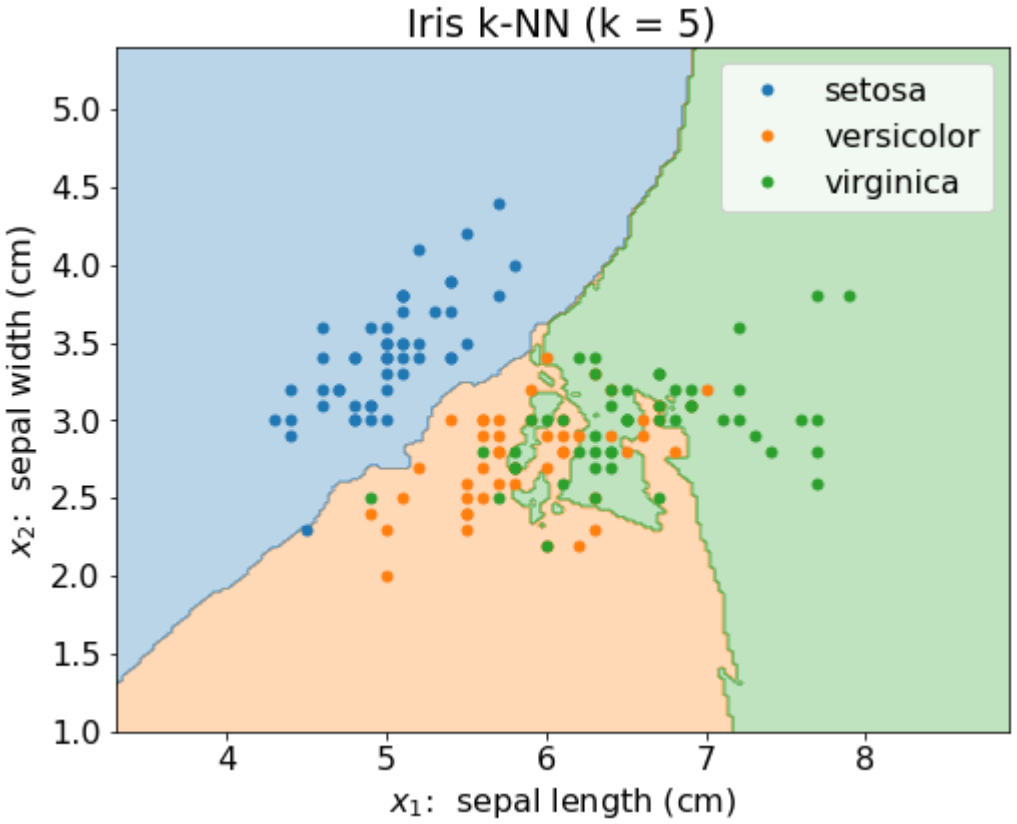
10-315 Introduction to ML

Logistic Regression

Instructor: Pat Virtue

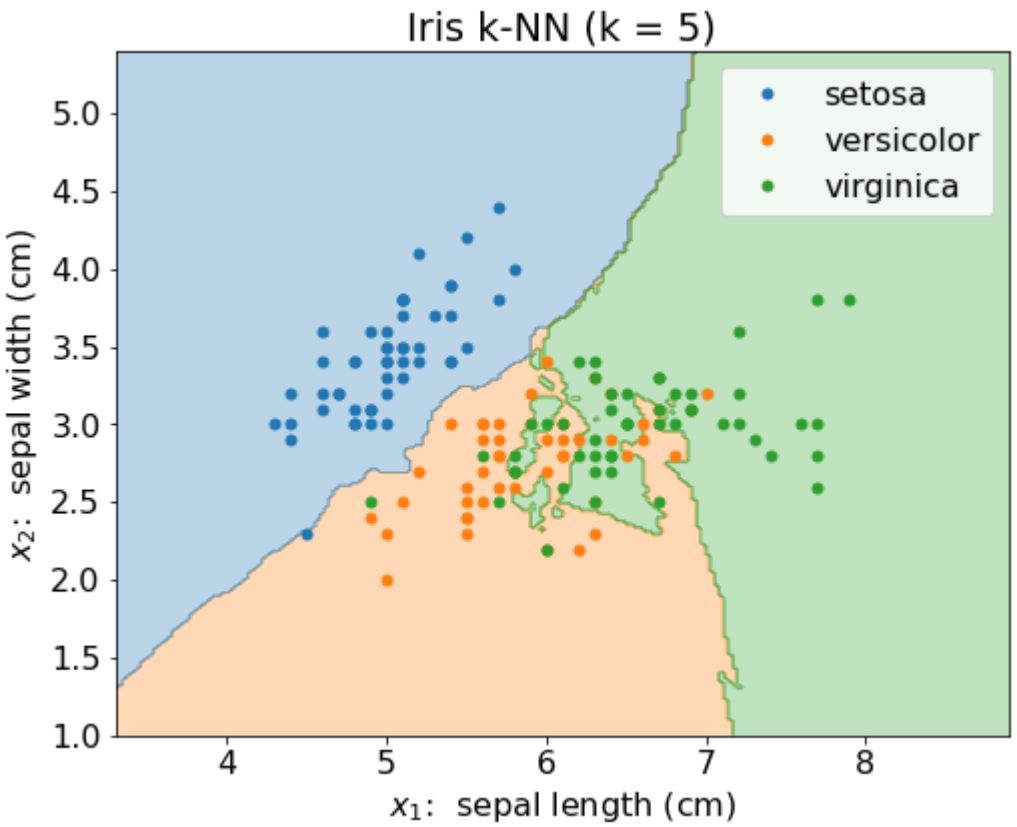
Classification Decisions

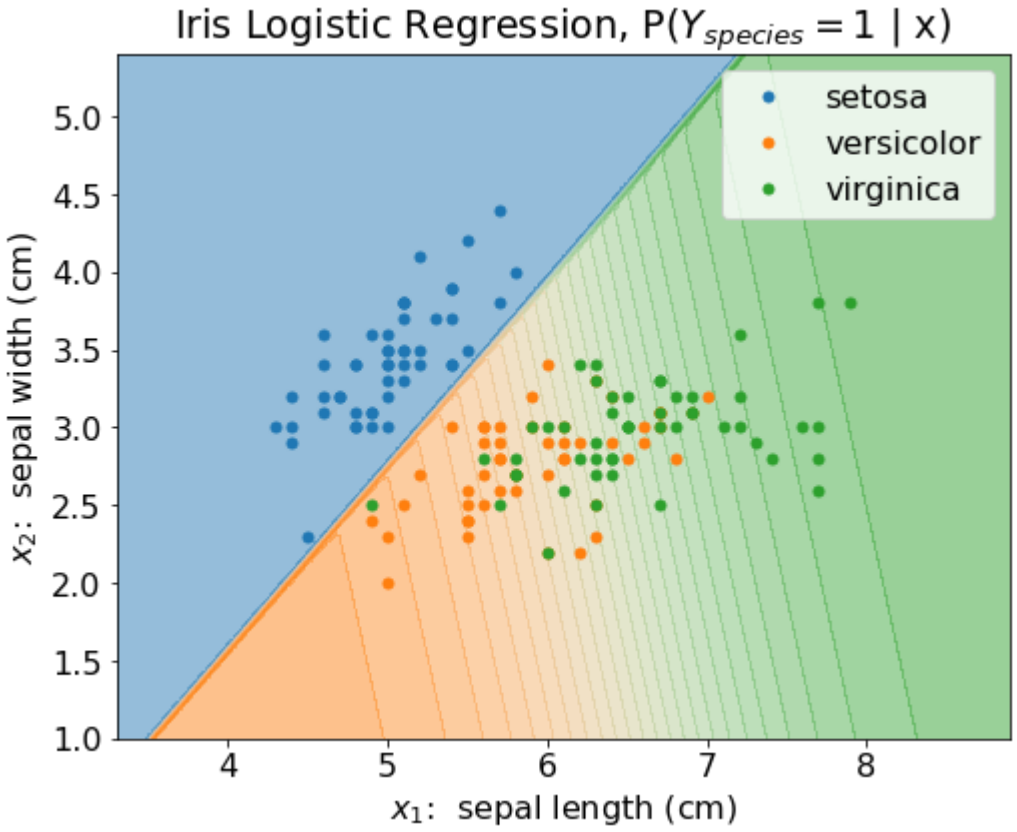
Predicting one specific class is troubling, especially when we know that there is some uncertainty in our prediction



Classification Probability

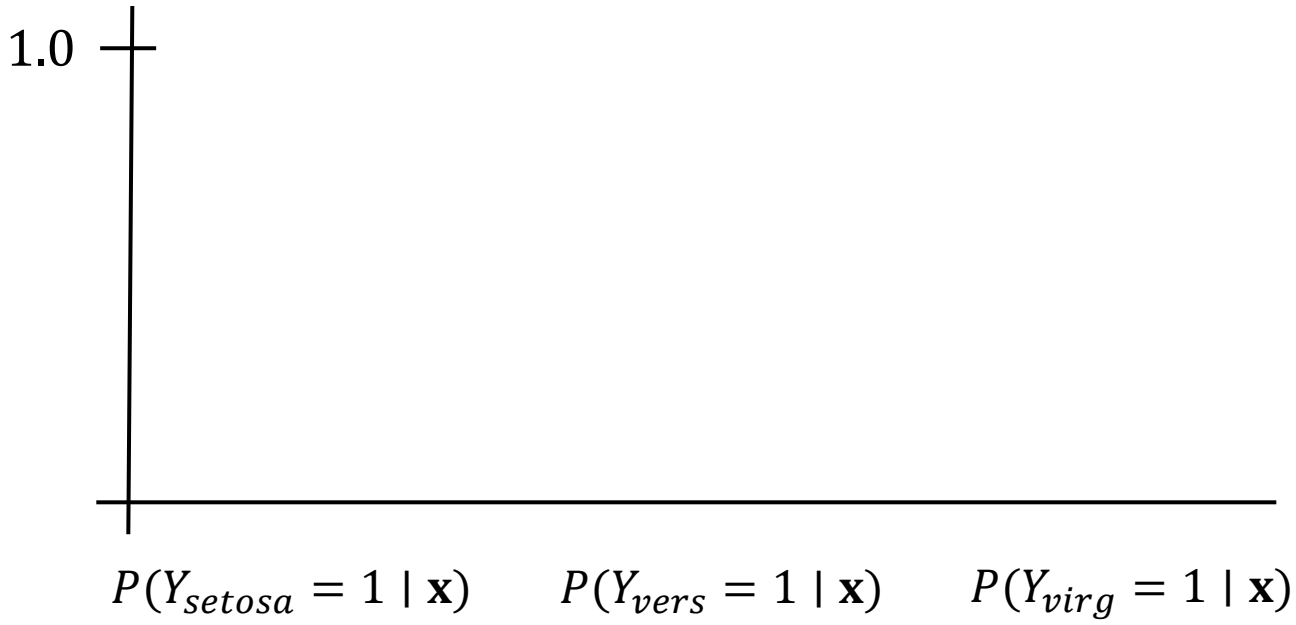
Constructing a model than can return the probability of the output being a specific class could be incredibly useful

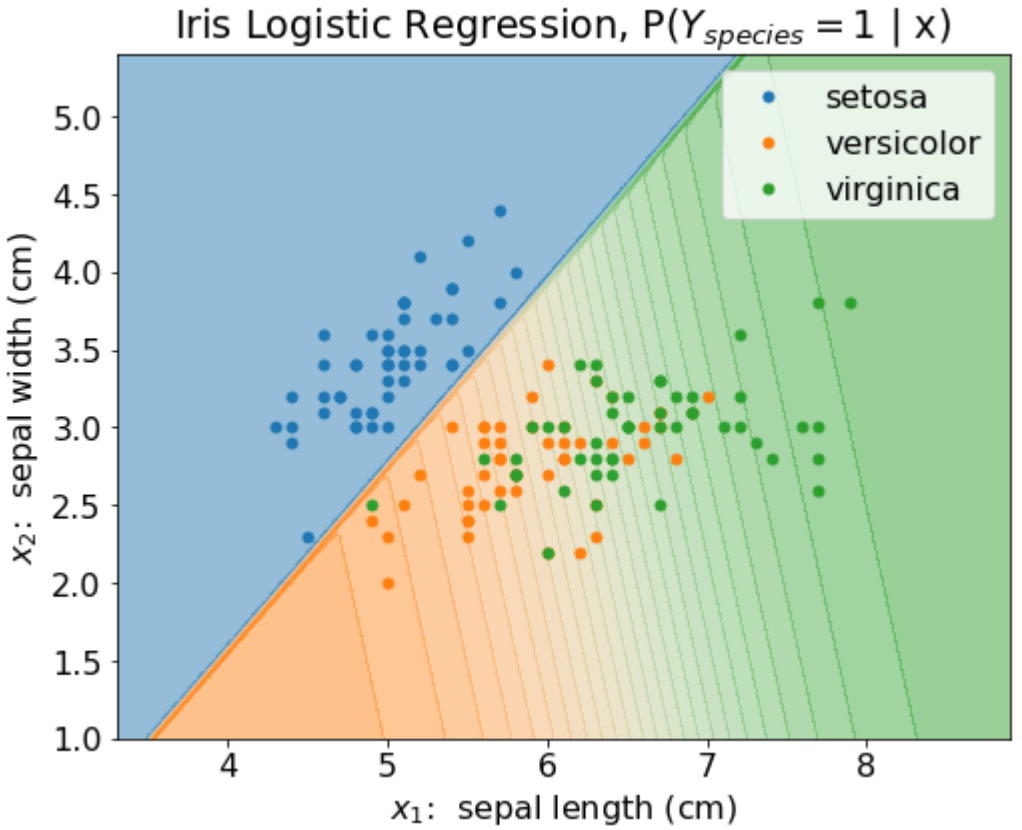




Classification Probability

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



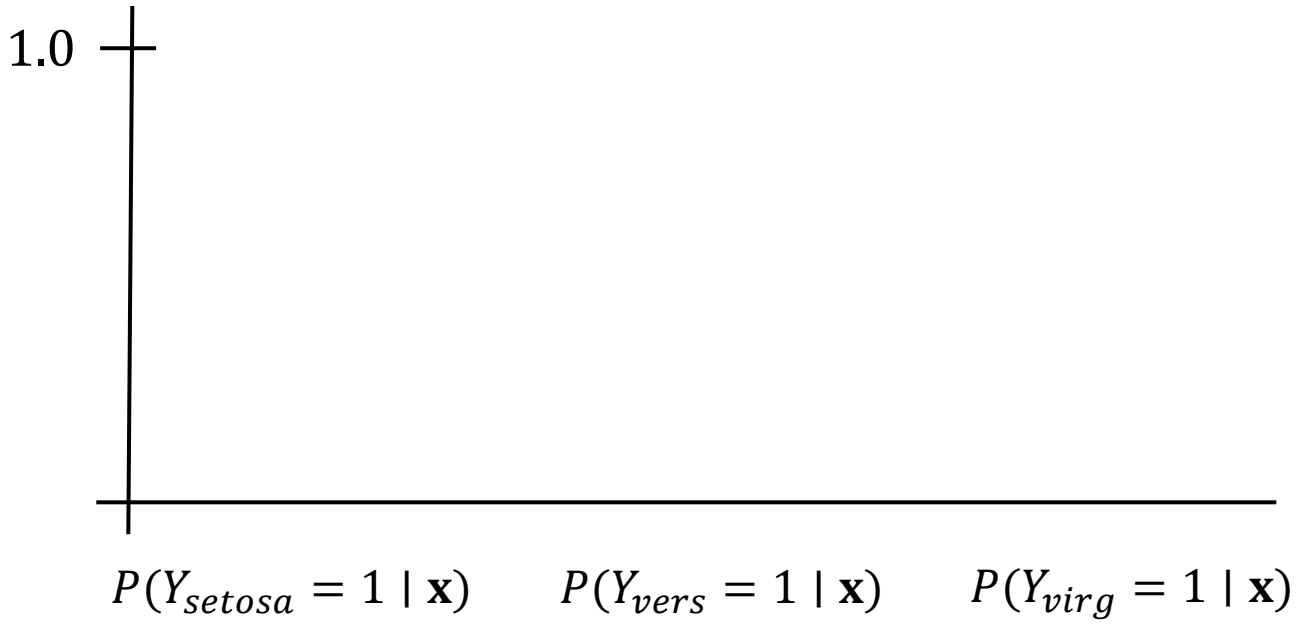


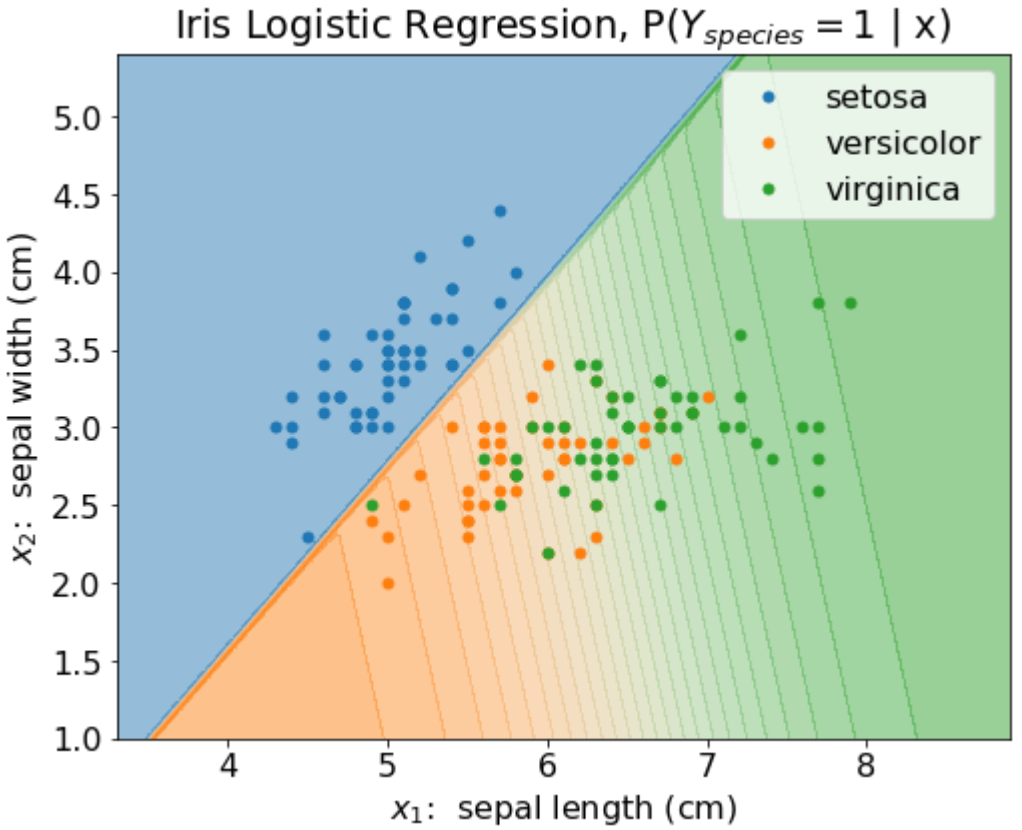
We can still make decisions, .e.g,

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

Loss for Probabilty Disributions

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



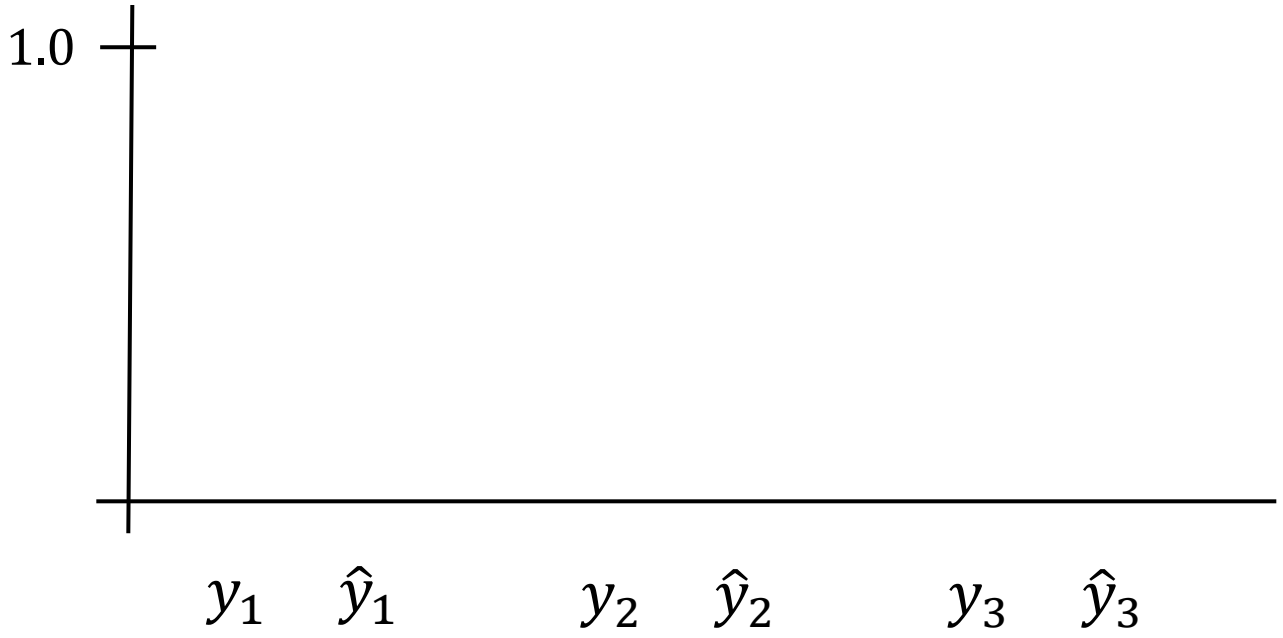


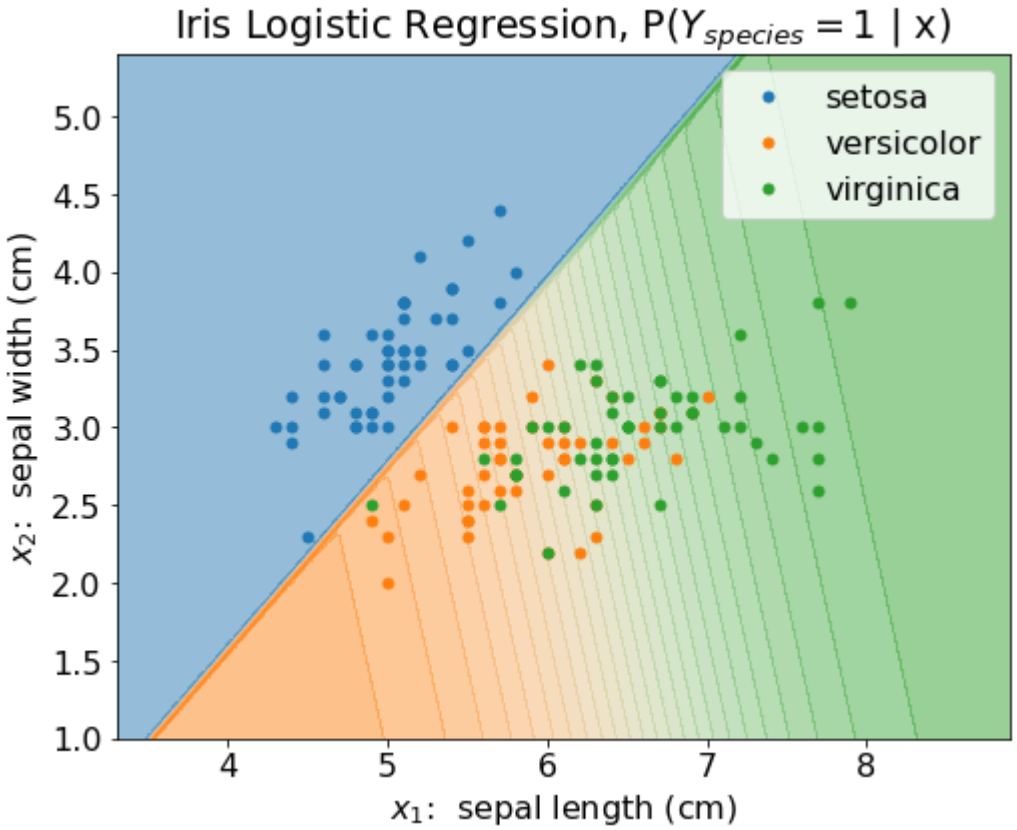
Cross-entropy loss

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

Loss for Probabilty Disributions

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





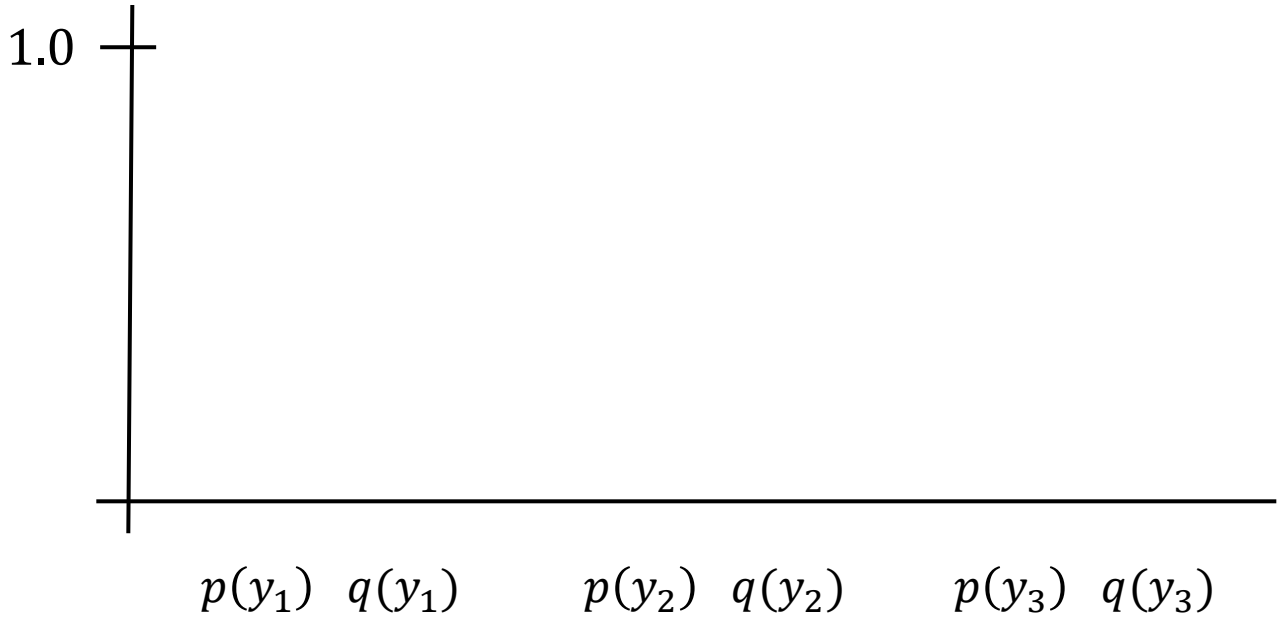
Cross-entropy loss

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

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

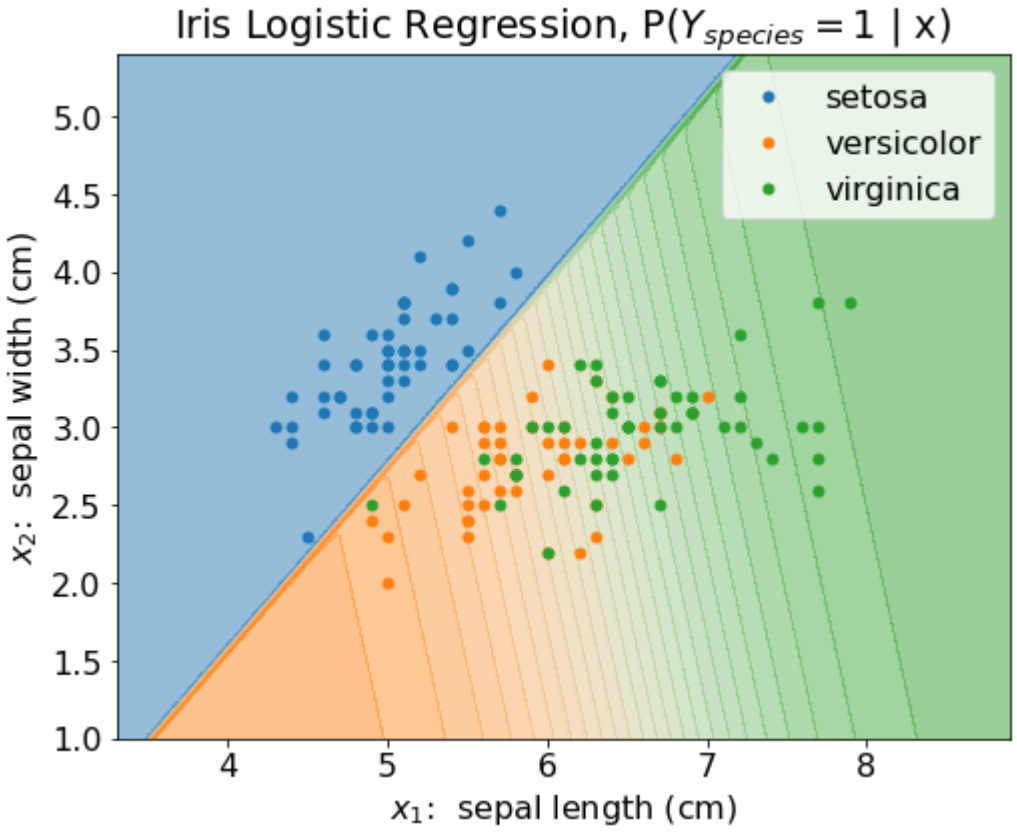
$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \hat{R}(h)$$

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

Cross-entropy loss

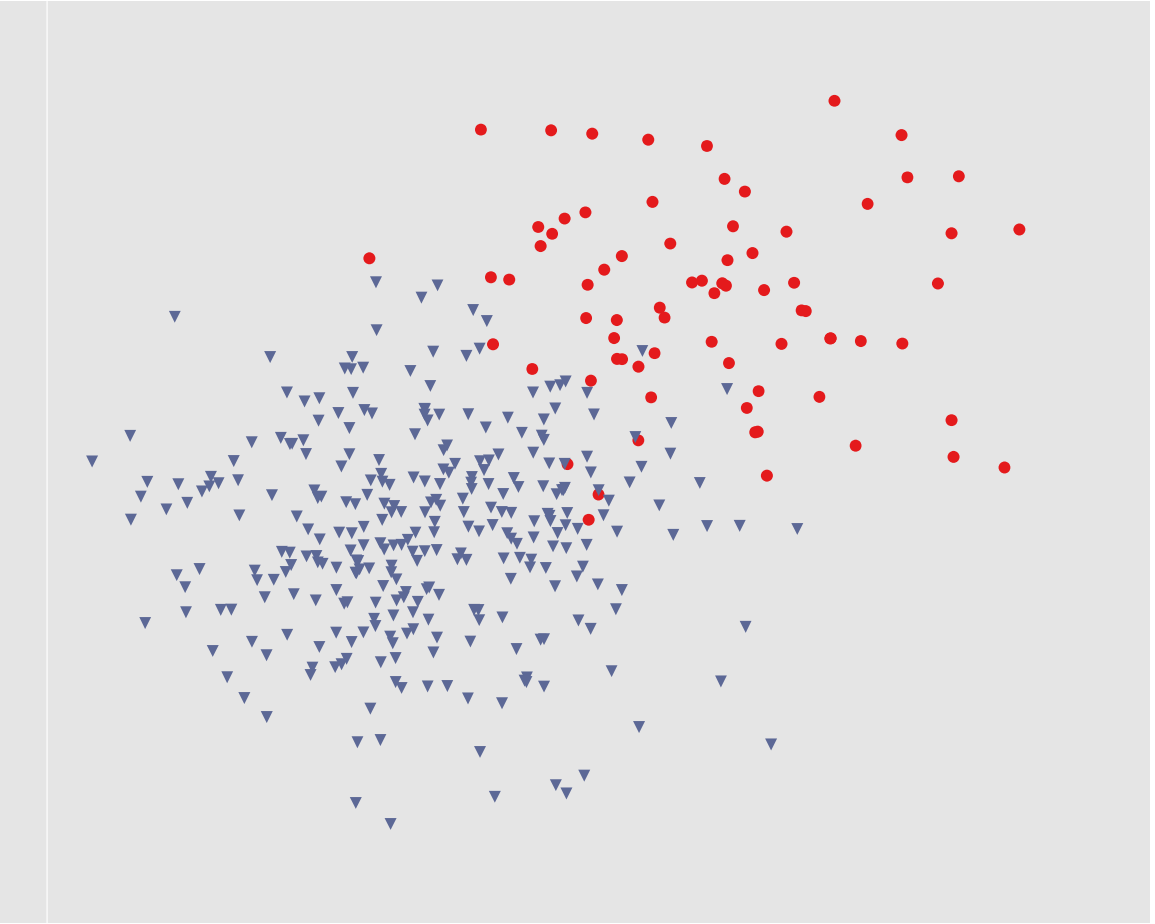
$$\ell(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{k=1}^{K} y_k \log \hat{y}_k$$

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.

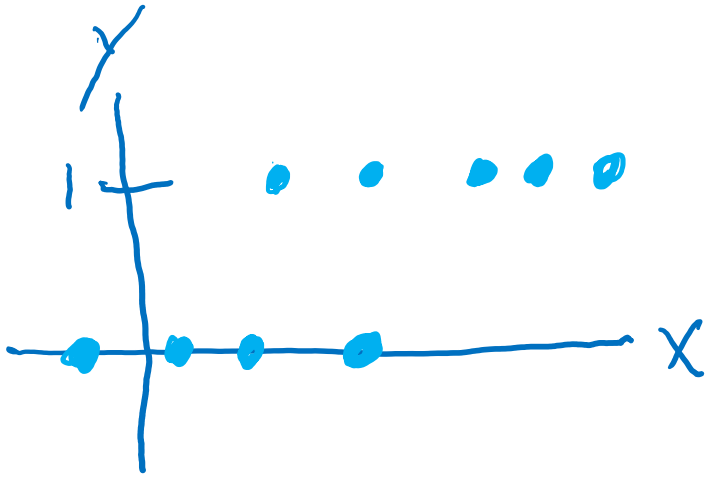


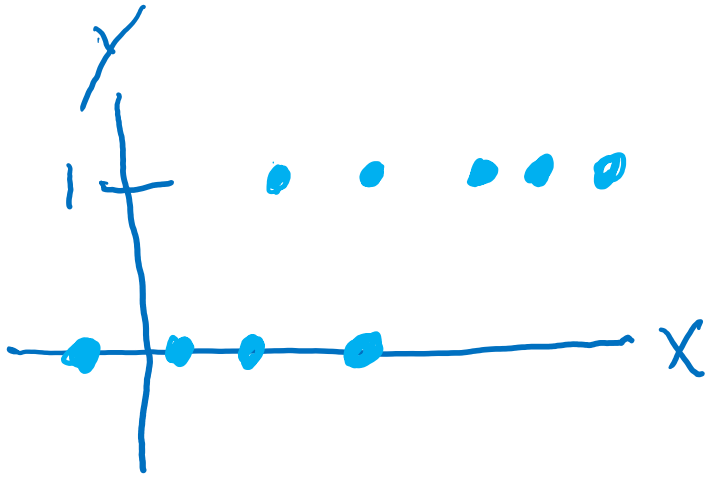
Prediction for Cancer Diagnosis

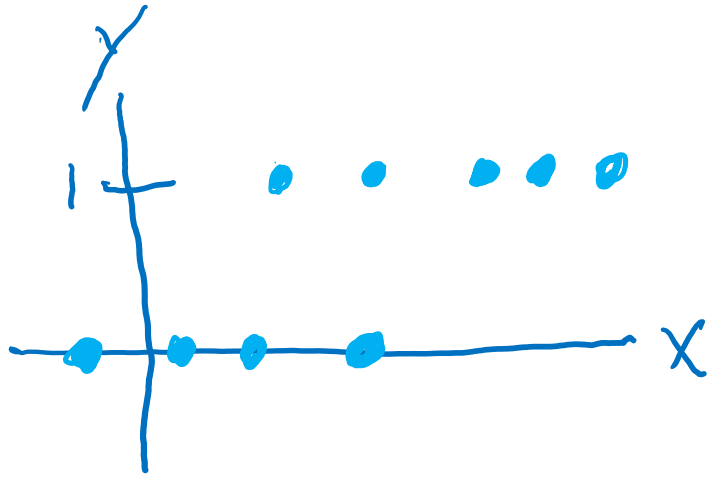
Learn to predict if a patient has cancer ($Y = 1$) or not ($Y = 0$) given the input of just one test result, X_A.

Building on a Linear Model

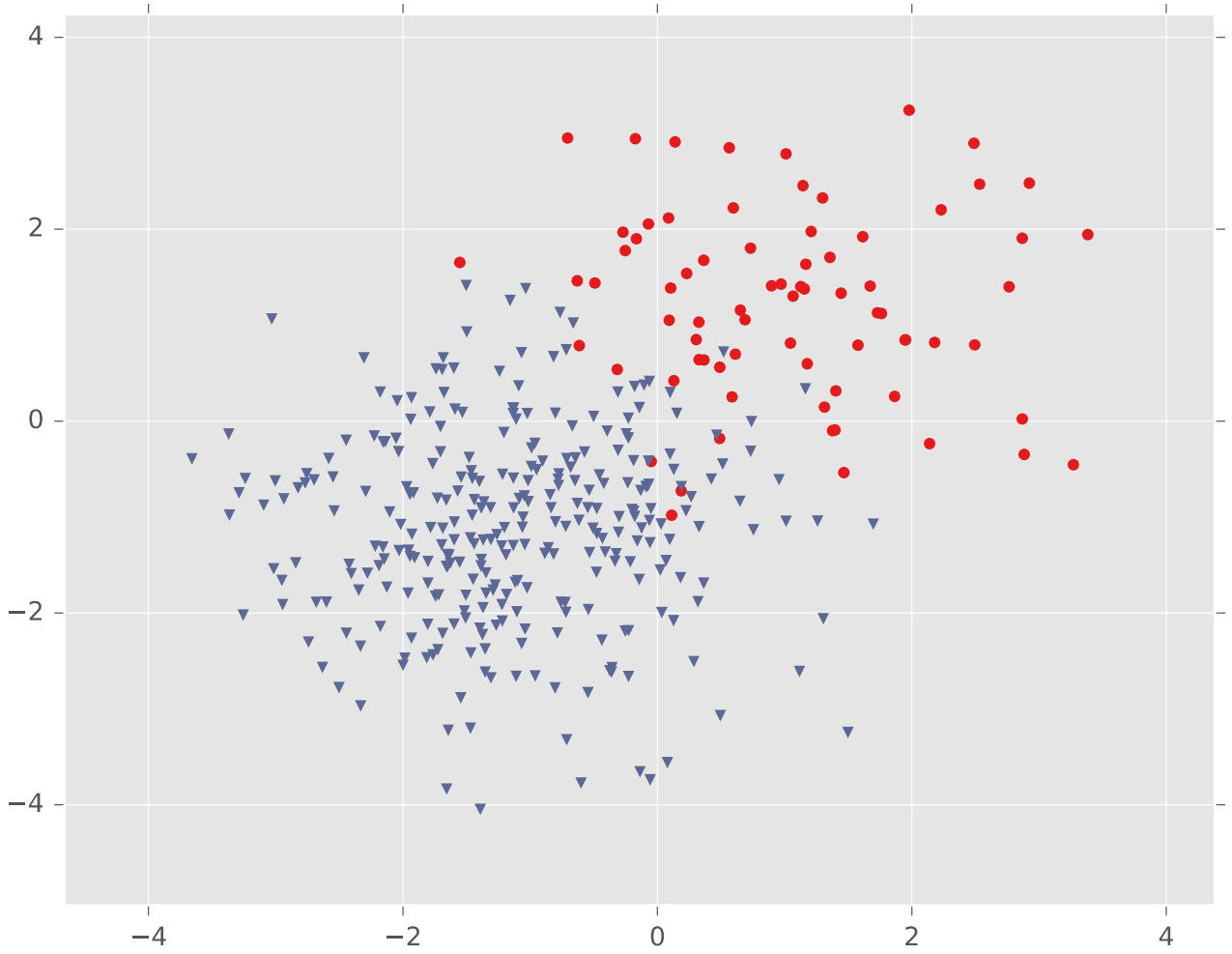
Linear vs Thresholded Linear vs Logistic Linear





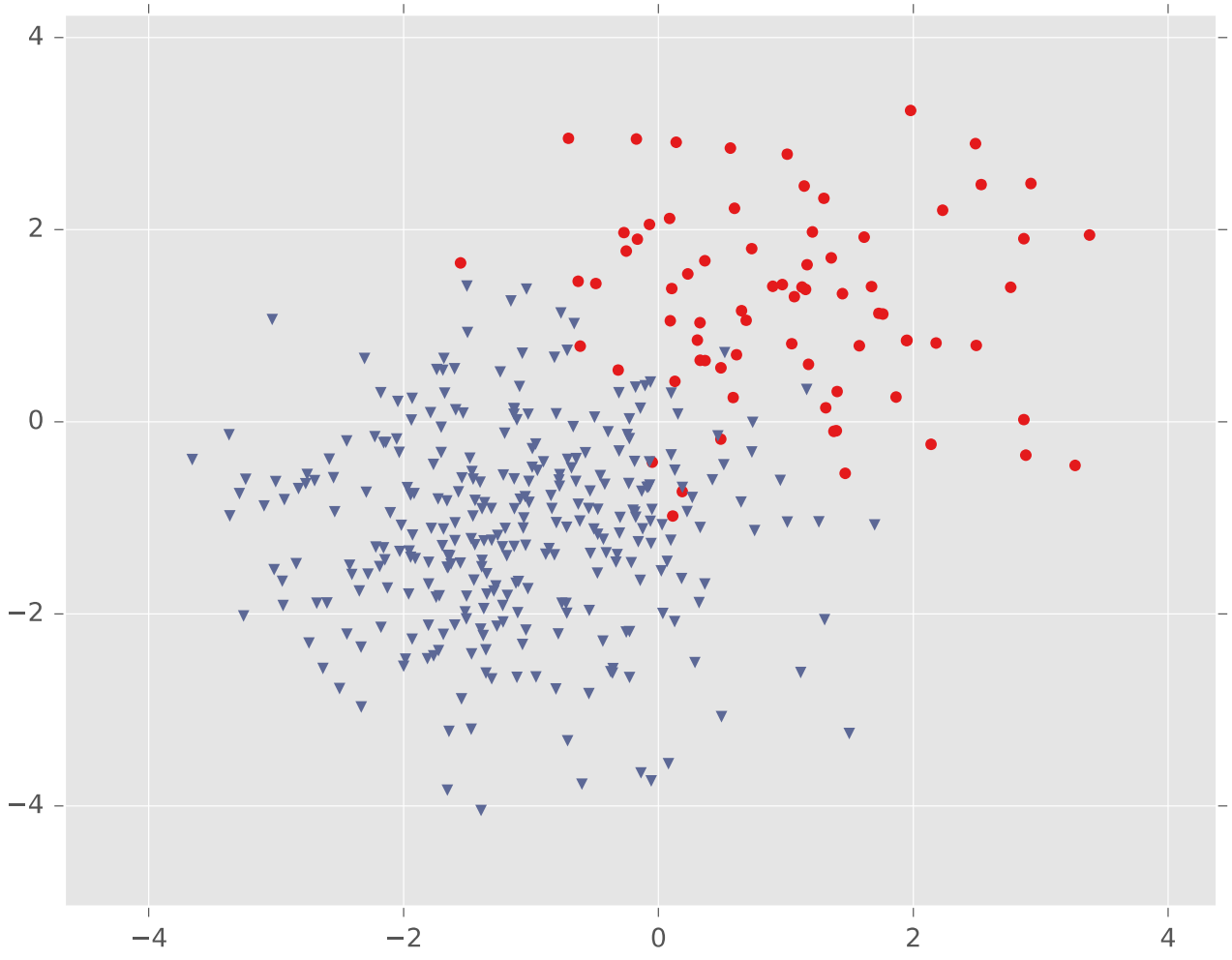


Building on a Linear Model



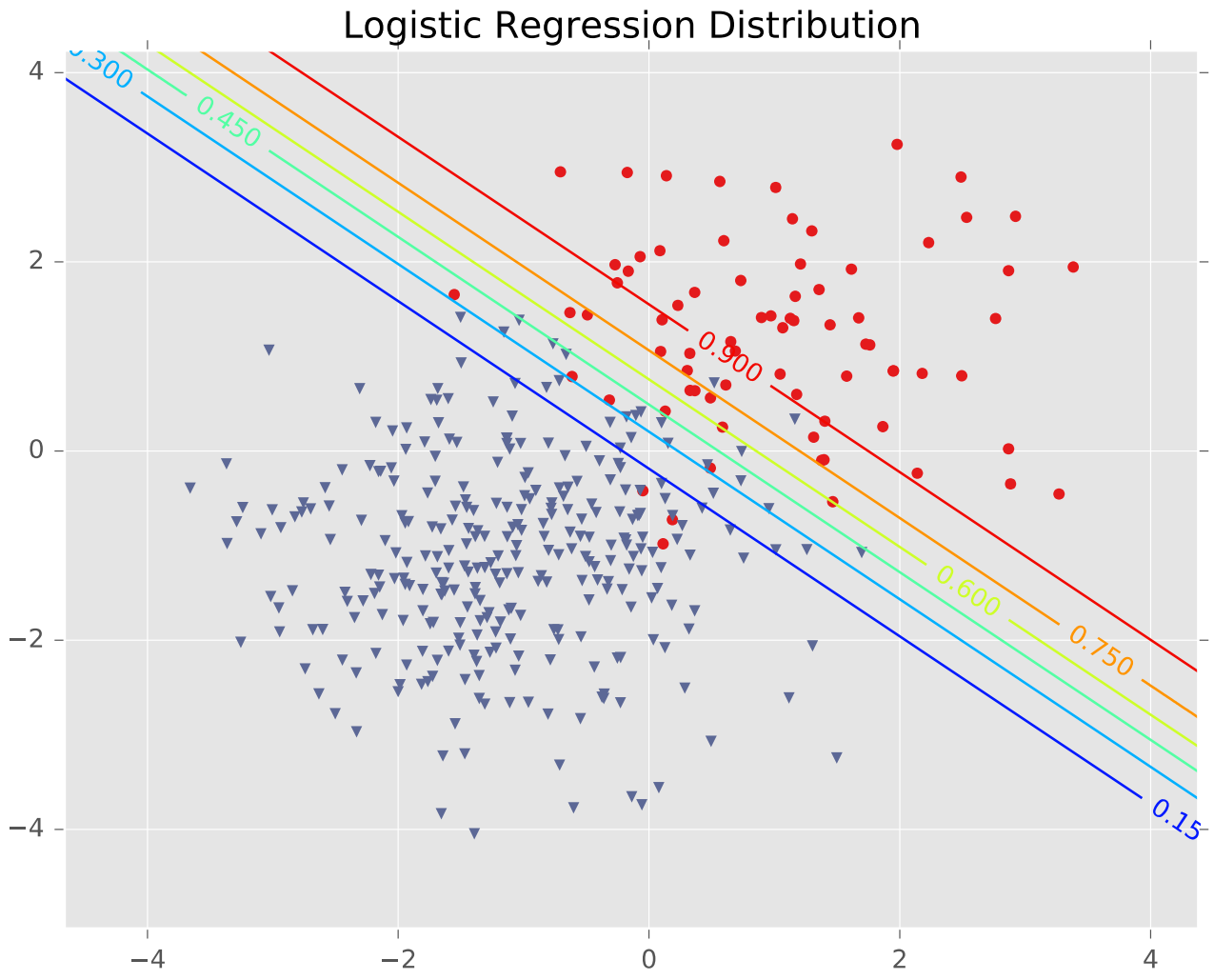
Building on a Linear Model

Logistic Regression



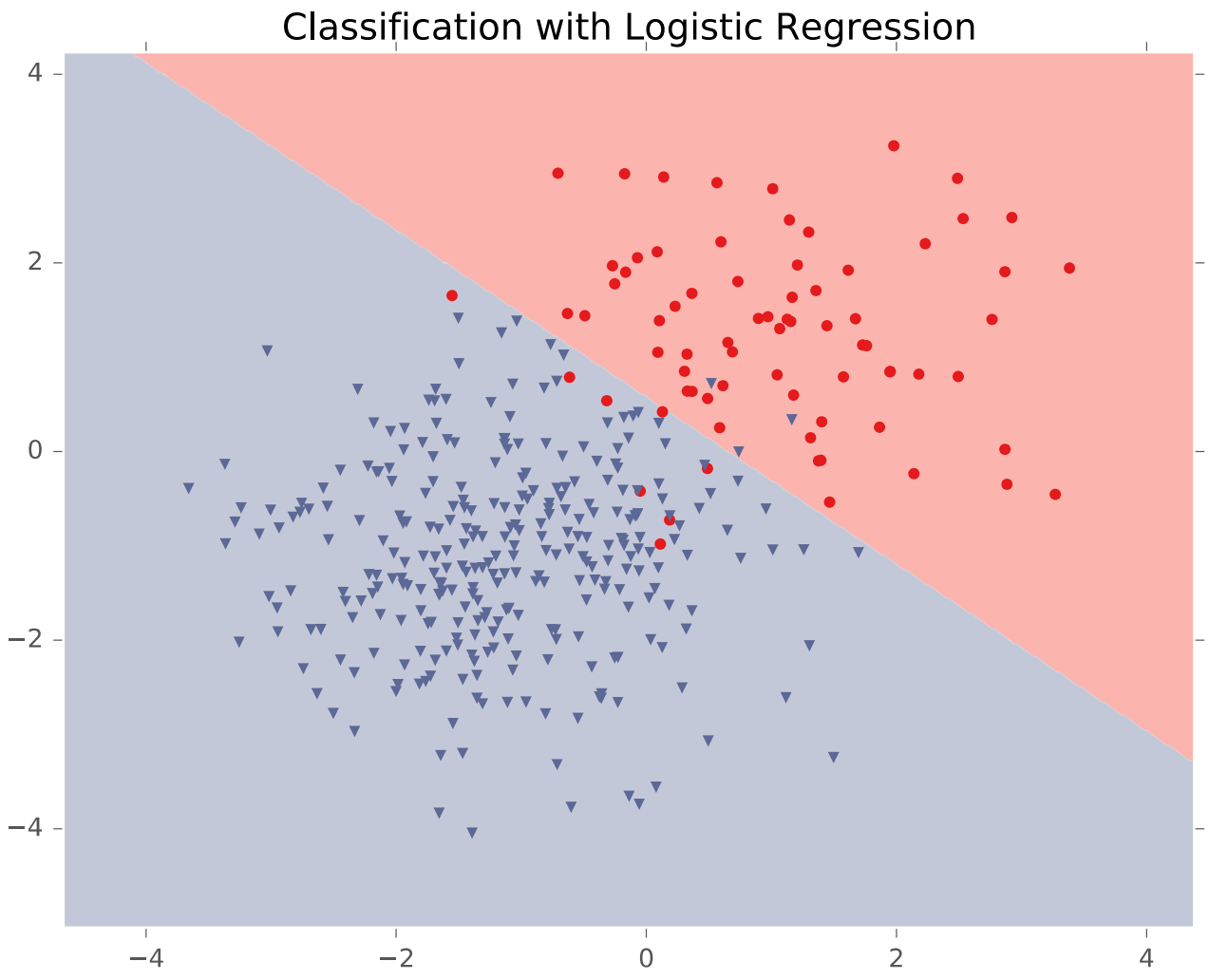
Slide credit: CMU MLD Matt Gormley

Logistic Regression



Slide credit: CMU MLD Matt Gormley

Logistic Regression Decision Boundary

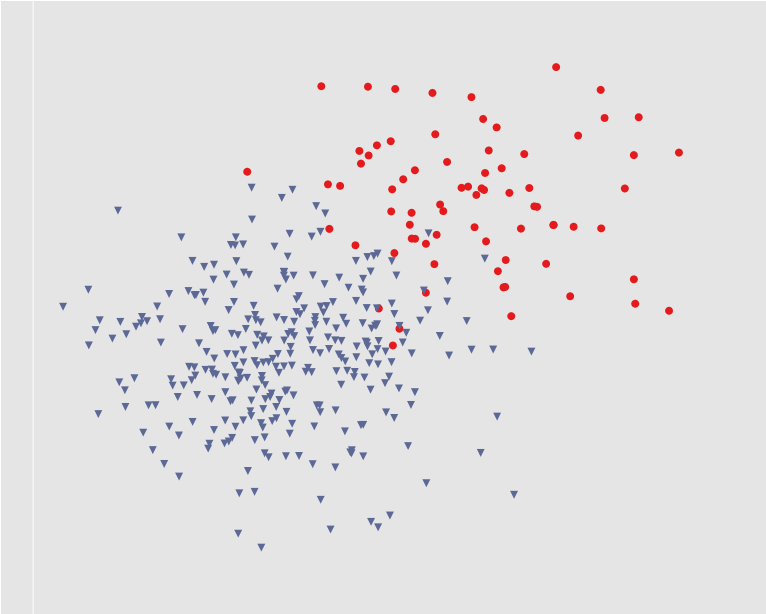


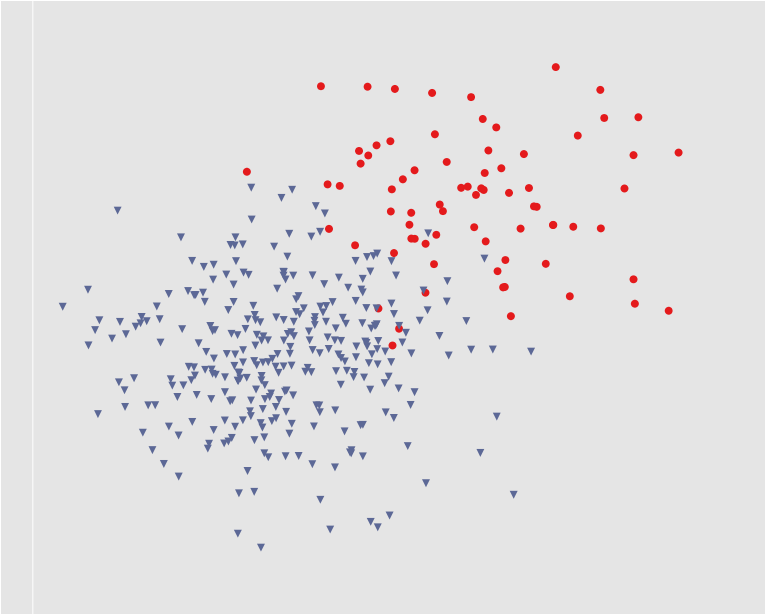
Slide credit: CMU MLD Matt Gormley

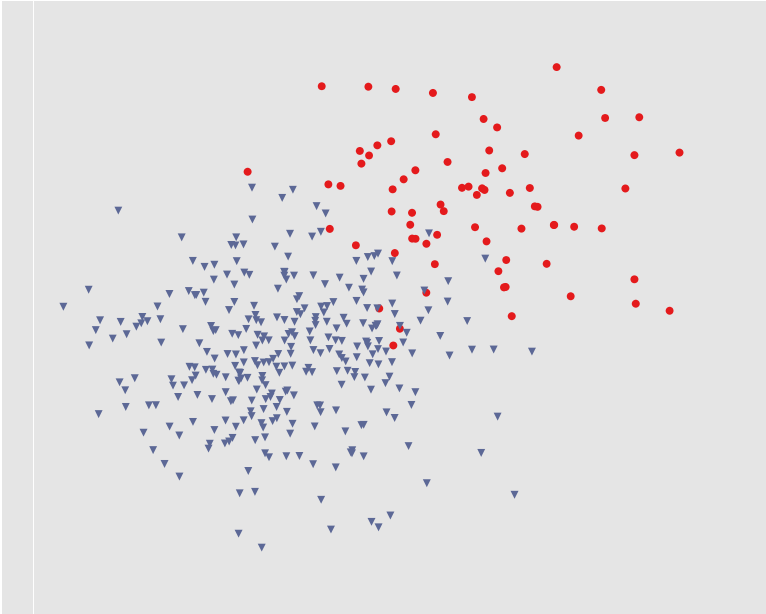
Optimizing a Model for Cancer Diagnosis

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

1) Model

2) Objective function

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

Binary Logistic Regression

Gradient

Solve Logistic Regression

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

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

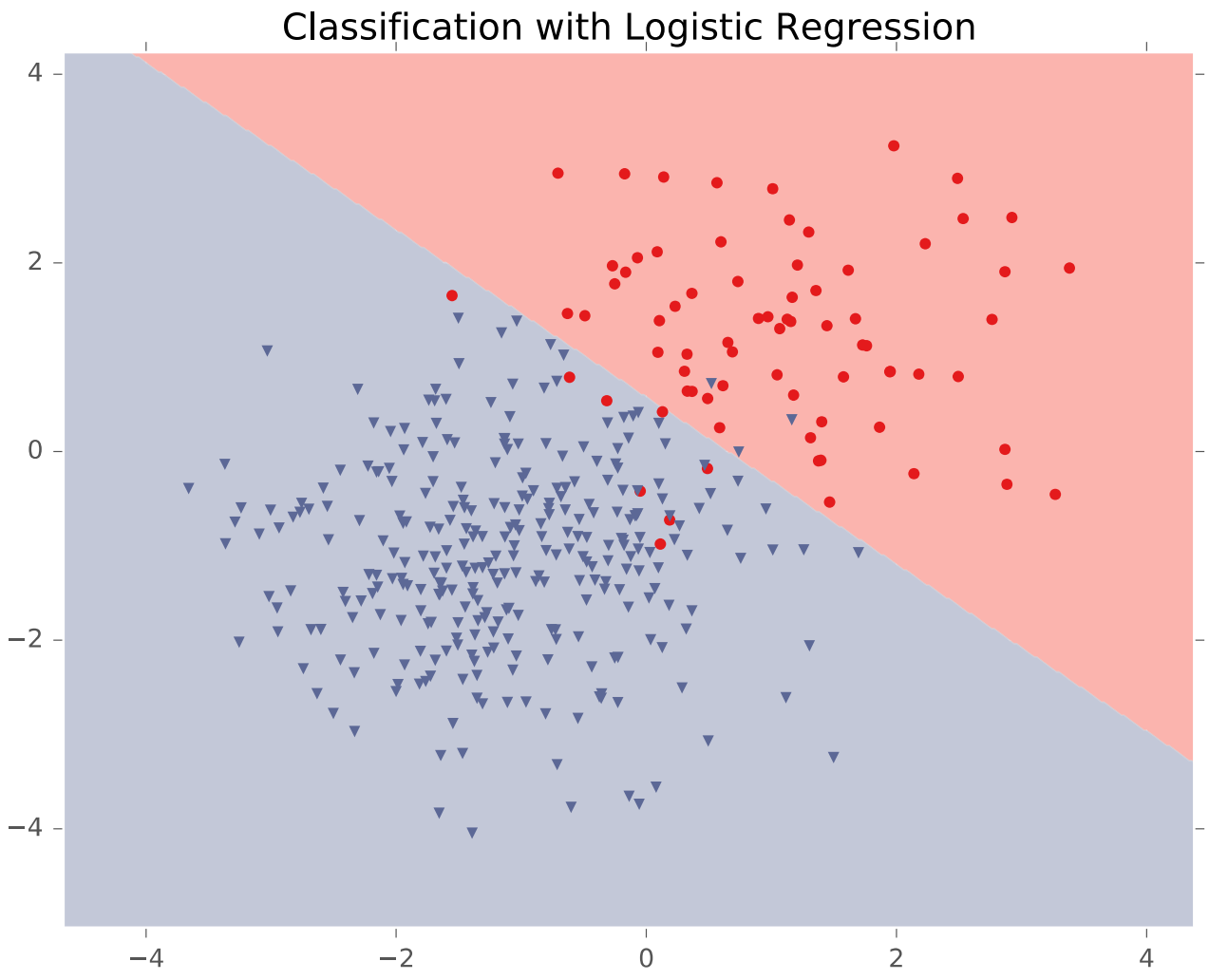
$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = -\frac{1}{N}\sum_i \left(y^{(i)} - \hat{y}^{(i)}\right) \mathbf{x}^{(i)}$$

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

No closed form solution ☹

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

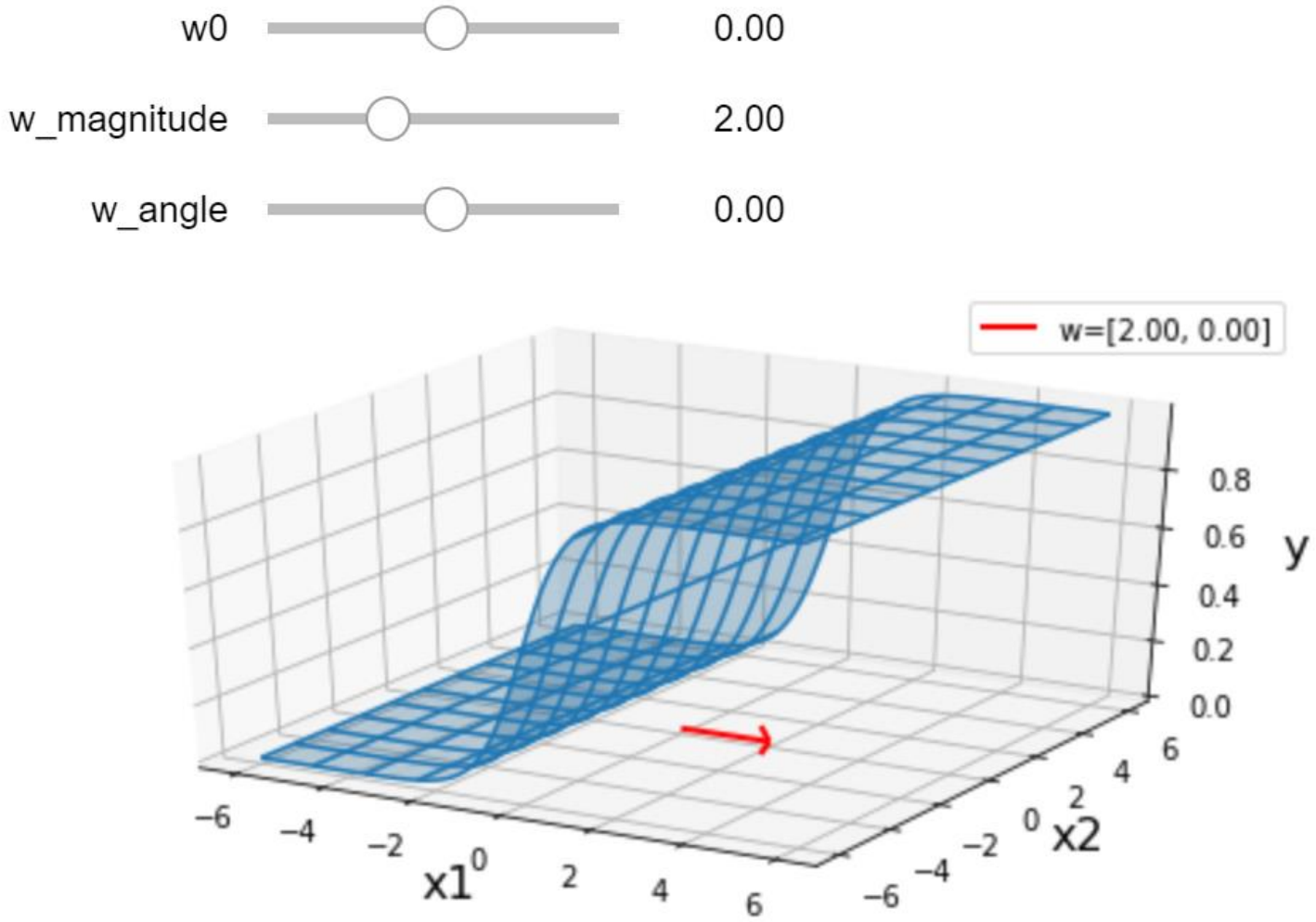
Logistic Regression Decision Boundary



Slide credit: CMU MLD Matt Gormley

Exercise

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



Linear in Higher Dimensions

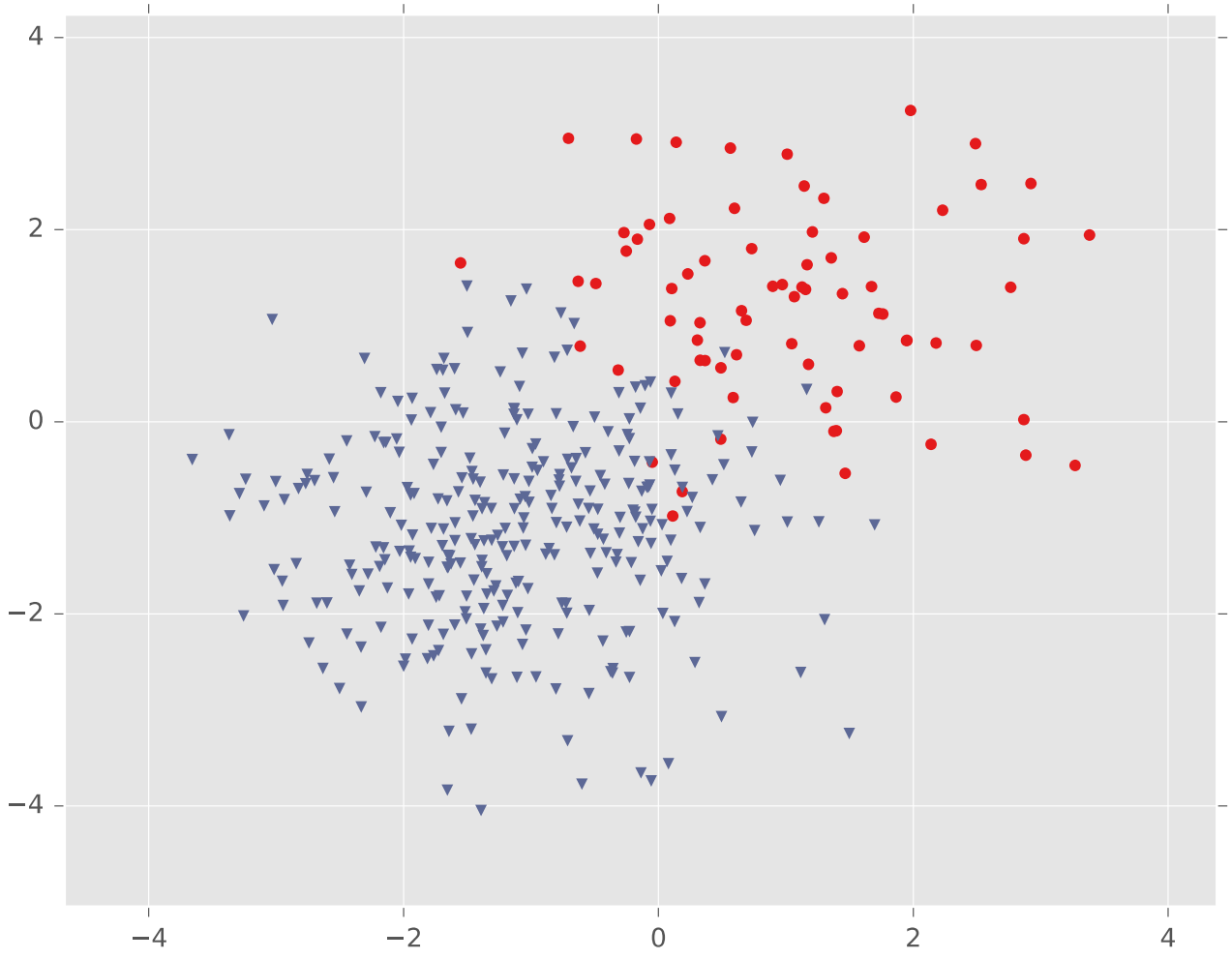
1-D $y = wx + b$

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?

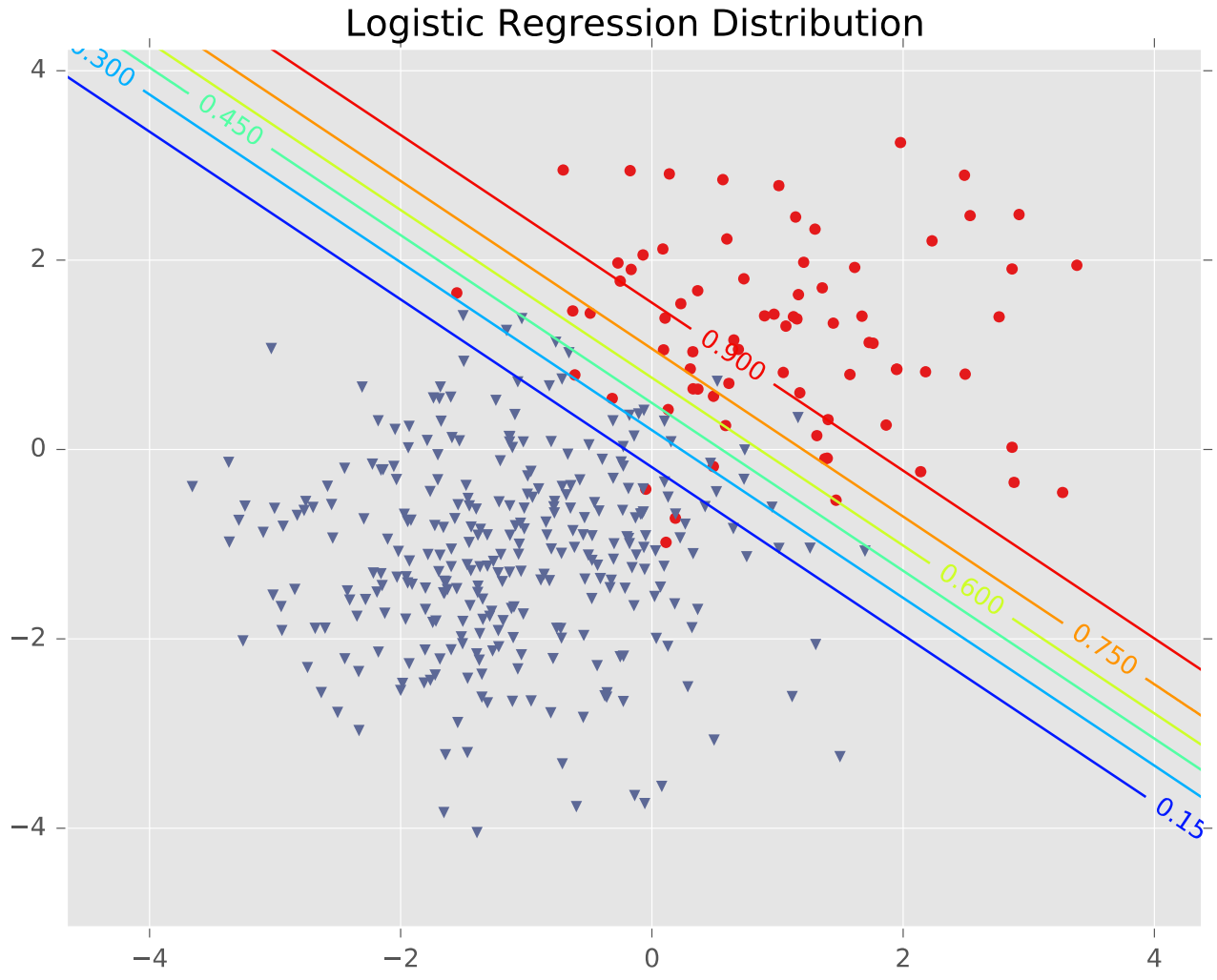
	$x \in \mathbb{R}$	$x \in \mathbb{R}^2$	$x \in \mathbb{R}^3$	$x \in \mathbb{R}^M$
→ $y = w^T x + b$	line	plane	hyperplane	hyperplane
$w^T x + b = 0$	point	line	plane	hyperplane
$w^T x + b \geq 0$	halfline	halfplane	halfspace	halfspace

Logistic Regression



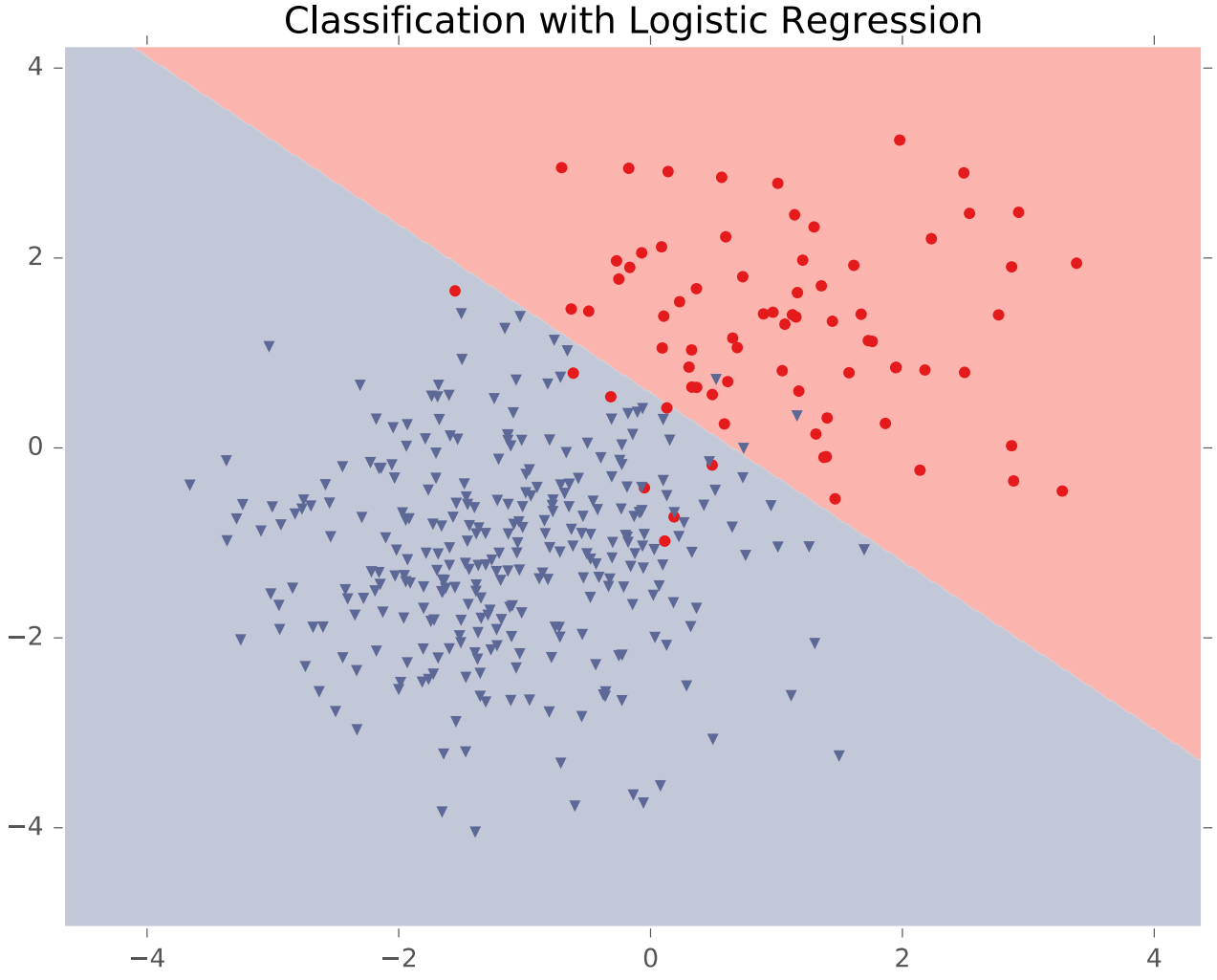
Slide credit: CMU MLD Matt Gormley

Logistic Regression



Slide credit: CMU MLD Matt Gormley

Logistic Regression

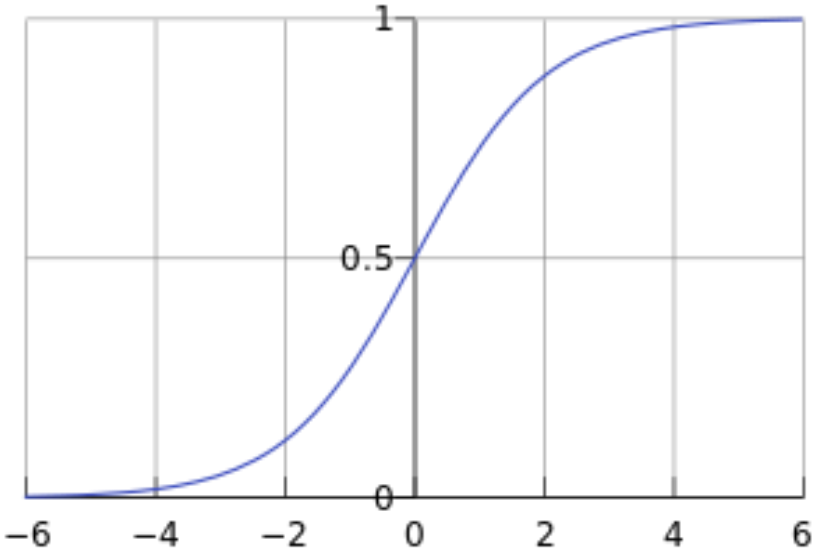


Slide credit: CMU MLD Matt Gormley

Poll 1

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

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

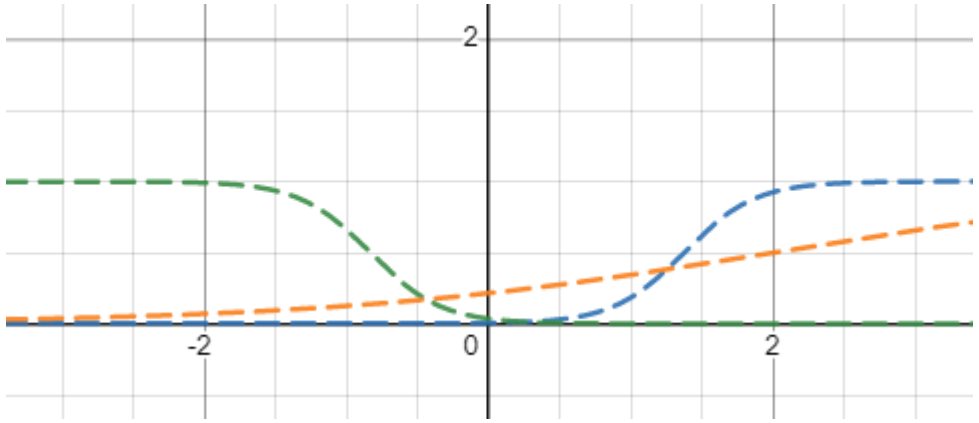


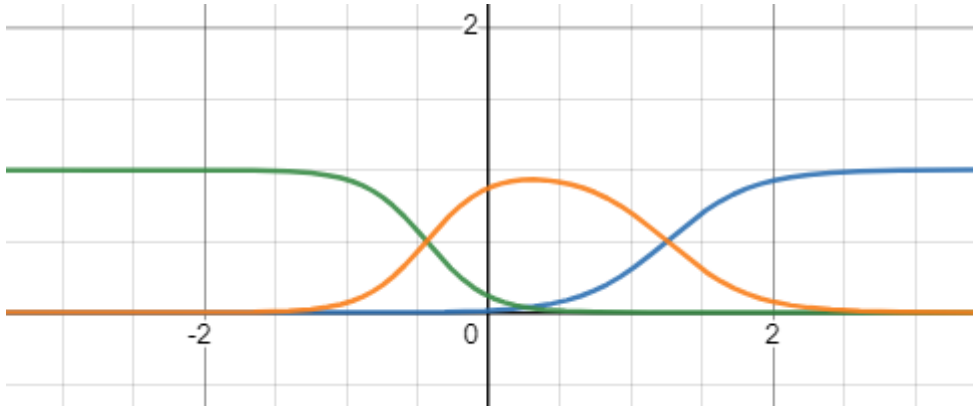
Multi-class Logistic Regression

Multi-class Logistic Regression

Desmos Demo:

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



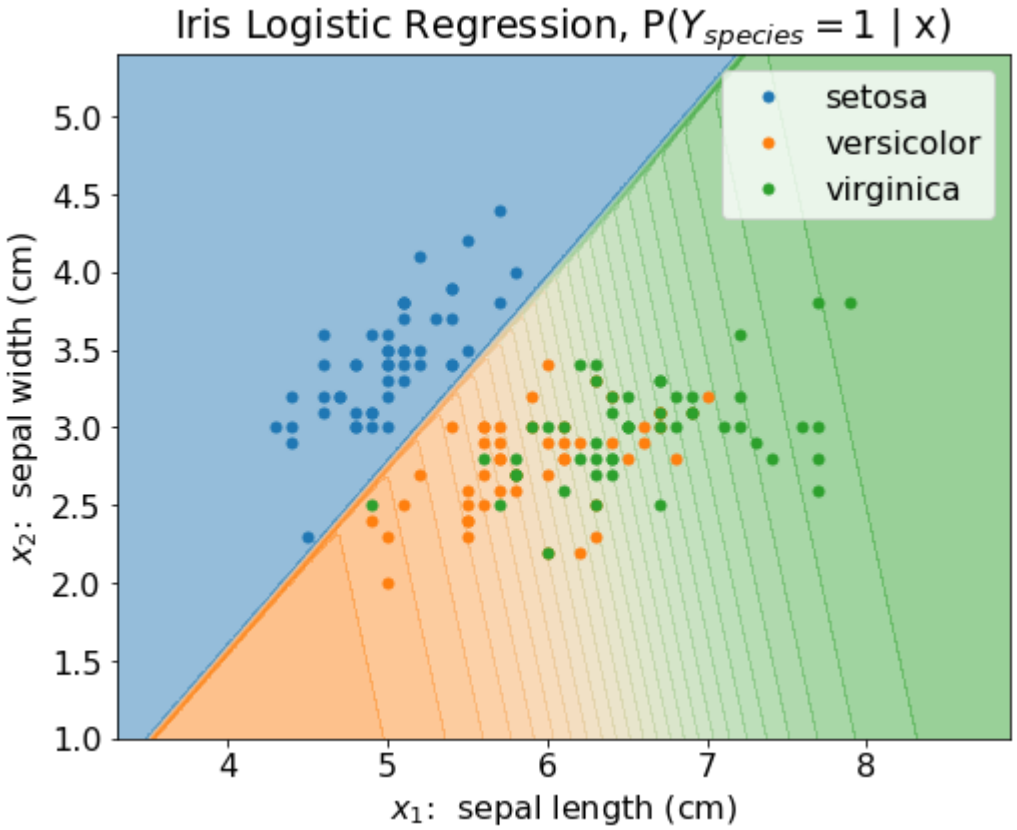


Multi-class Logistic Regression

Cross-entropy loss

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

Model



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 $\rightarrow g(2) = 0.88, \quad 1\text{-}g(2) = 0.12$

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

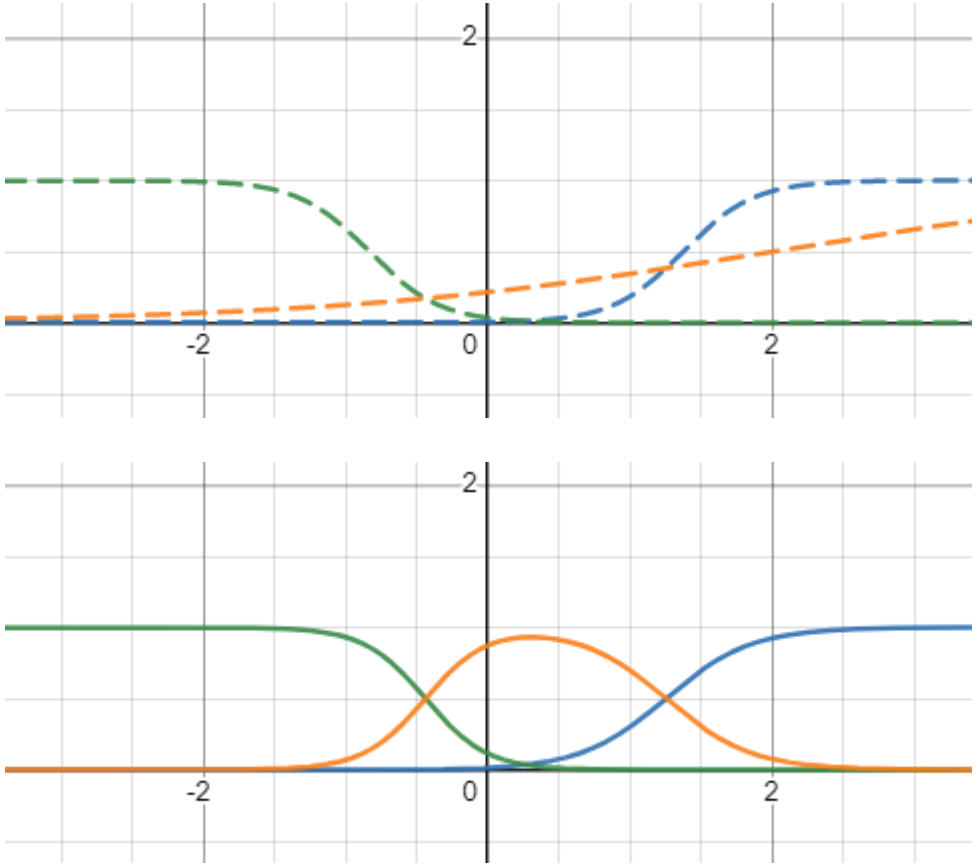
Multiclass Predicted Probability

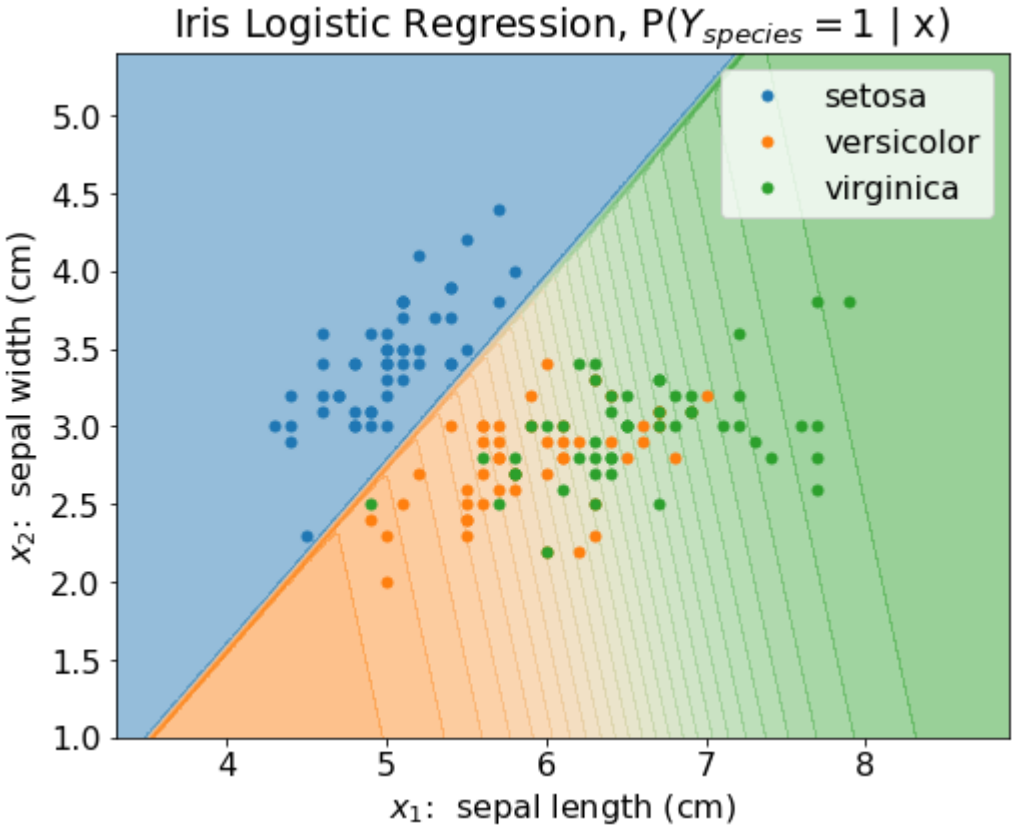
Multiclass logistic regression uses the parameters learned across all K classes to predict the discrete conditional probability distribution of the output Y given a specific input vector $\mathbf{x}$

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

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 $\mathbf{x}$

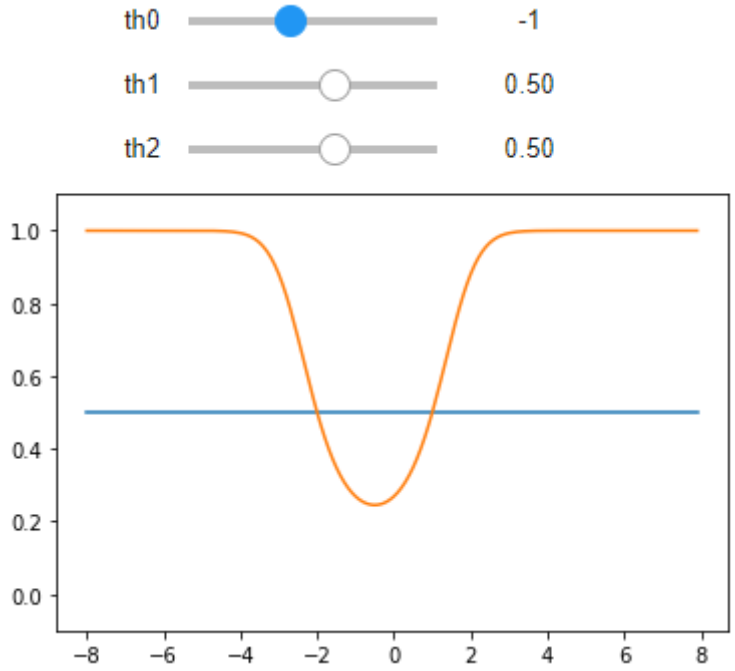




Logistic Regression with Polynomial Features

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