Do not distribute course material

You may not and may not allow others to reproduce or distribute lecture notes and course materials publicly whether or not a fee is charged.

Topic 4
Linear Classification & Logistic Regression

- http://cs229.stanford.edu/notes2020fall/notes2020fall/cs229-notes1.pdf
- https://eight2late.wordpress.com/2017/07/11/a-gentle-introduction-to-logistic-regression-and-lasso-regularisation-using-r/

PROF. LINDA SELLIE

Learning objectives

- Know how to use a hyperplane for binary classification
- Use the sigmoid function to scale a number in the range $[-\infty, \infty]$ into $[0,1]$
- Apply the principle of maximum likelihood estimation (MLE) to learn the parameters of a probabilistic model
- Derive the conditional log-likelihood
- How to apply gradient ascent to find the parameters of the the conditional log-likelihood
- Evaluate performance with different measures
- Create more complex models by feature transformation
- Understand how to add L1 and L2 regularization to the objective function
- Know how to interpret the output of soft-max

Outline

- Motivating example: How can we classify?
  How can we use a hyperplane for a classification problem?
- Estimating probabilities
  Can we predict not only which class an example belongs to - but a confidence score of that classification
- Maximum likelihood
  How can we find the most likely hyperplane? Could we write a function to describe how likely a hyperplane was to have generated the dataset?
- Thinking about different types of error
  Some errors are more costly than other errors. Can we modify our predictions to decrease one type of error (and perhaps increase another type of error?
- Transformation of the features
  Extending our algorithm to nonlinear decision boundaries
- Multiple classes
  What if we have more than two classes?

Outline

- Motivating example: How can we classify?

  How can we use a hyperplane for a classification problem?

- E

  Which model

  Can we predict not only which class an example belongs to - but a confidence score of that classification

- Finding an objective function

  How can we find the most likely hyperplane? Could we write a function to describe how likely a hyperplane was to have generated the dataset?

- Thinking about different types of error

  Some errors are more costly than other errors. Can we modify our predictions to decrease one type of error (and perhaps increase another type of error?

- Transformation of the features

  Extending our algorithm to nonlinear decision boundaries

- Multiple classes

  What if we have more than two classes?

Classification vs Regression


https://en.wikipedia.org/wiki/Iris_flower_data_set#/media/File:Kosaciec_szczecinkowaty_Iris_setosa.jpg

- Regression we were given:

$$\{(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),\ldots,(x^{(N)},y^{(N)})\},\;\; x\in\mathbb{R}^d,\;\; y\in\mathbb{R}$$

- Classification we are given:

$$\{(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),\ldots,(x^{(N)},y^{(N)})\},\;\; x\in\mathbb{R}^d,\;\; y\in\{0,1\} \text{ or } y\in\{0,1,2,3,4\},\ldots$$

- Given attributes of a flower: (['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)']

$\mathbf{x}^T=(5.1 \quad 3.5 \quad 1.4 \quad 0.2)$

- If you knew a flower was either a setosa Iris or versicolor Iris can you determine which type it is?


https://commons.wikimedia.org/wiki/File:Iris_versicolor_3.jpg#file

```
1 - setosa

0 - versicolor
```

Classification vs Regression

- Regression we were given:

$$\{(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),\ldots,(x^{(N)},y^{(N)})\},\ \ x\in\mathbb{R}^d,\ \ y\in\mathbb{R}$$

If we have more than two categories we can use an encoding to represent the target. If the target could take be virginica, versicolor, setosa or we could represent virginica as [1,0,0], versicolor as [0,1,0], and setosa as [0,0,1]

e are given:

$$(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),\ldots,(x^{(N)},y^{(N)})\},\ \ x\in\mathbb{R}^d,\ \ y\in\{0,1\}$$

If we have two classes, for example: setosa Iris' and versicolor Iris' we can choose to call one class 1 and the other class 0. It doesn't matter which we choose.

, of a flower: (['sepal length (cm)', 'sepal width (cm)', 'petal lengt h (cm)']

$\mathbf{x}^T$=(5.1 3.5 1.4 0.2)

- If you knew a flower was either a setosa Iris or versicolor Iris can you determine which type it is?

1 – setosa

0 – versicolor

ttps://en.wikipedia.org/wiki/
ris_flower_data_set#/media/
ile:Kosaciec_szczecinkowaty_Iris_setos
.jpg

https://commons.wikimedia.org/wiki/
File:Iris_versicolor_3.jpg#file

Intuition

`x=(sepal width, petal length)`

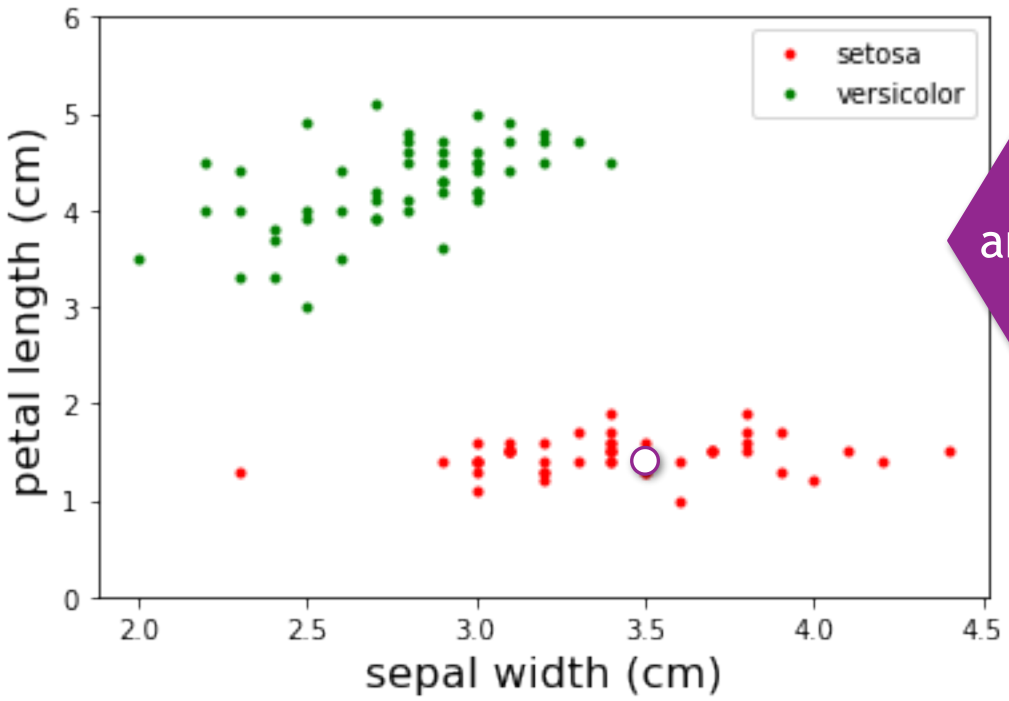
- setosa Iris

```
[ 3.5   1.4 ]
[ 3.    1.4 ]
[ 3.2   1.3 ]
[ 3.1   1.5 ]
[ 3.6   1.4 ]
[ 3.9   1.7 ]
[ 3.4   1.4 ]
[ 3.4   1.5 ]
[ 2.9   1.4 ]
[ 3.1   1.5 ]
```

- versicolor Iris

```
[ 3.2   4.7 ]
[ 3.2   4.5 ]
[ 3.1   4.9 ]
[ 2.3   4.  ]
[ 2.8   4.6 ]
[ 2.8   4.5 ]
[ 3.3   4.7 ]
[ 2.4   3.3 ]
[ 2.9   4.6 ]
[ 2.7   3.9 ]
```

`x=(3.5  1.4)`



The relationship separating the Irises using the features sepal width and petal length is very pronounced. Normally this relationship will not be so clean.

To simplify we will only look at **two** features: sepal width and petal length

1. How can we find a line that separates the data

Next

2. Given a line (hyperplane) how can we find which side of the line a point lies on?

Intuition: decision boundary

Data:

$\mathbf{x}_i$= (sepal width$_i$, petal length$_i$)

- setosa Iris

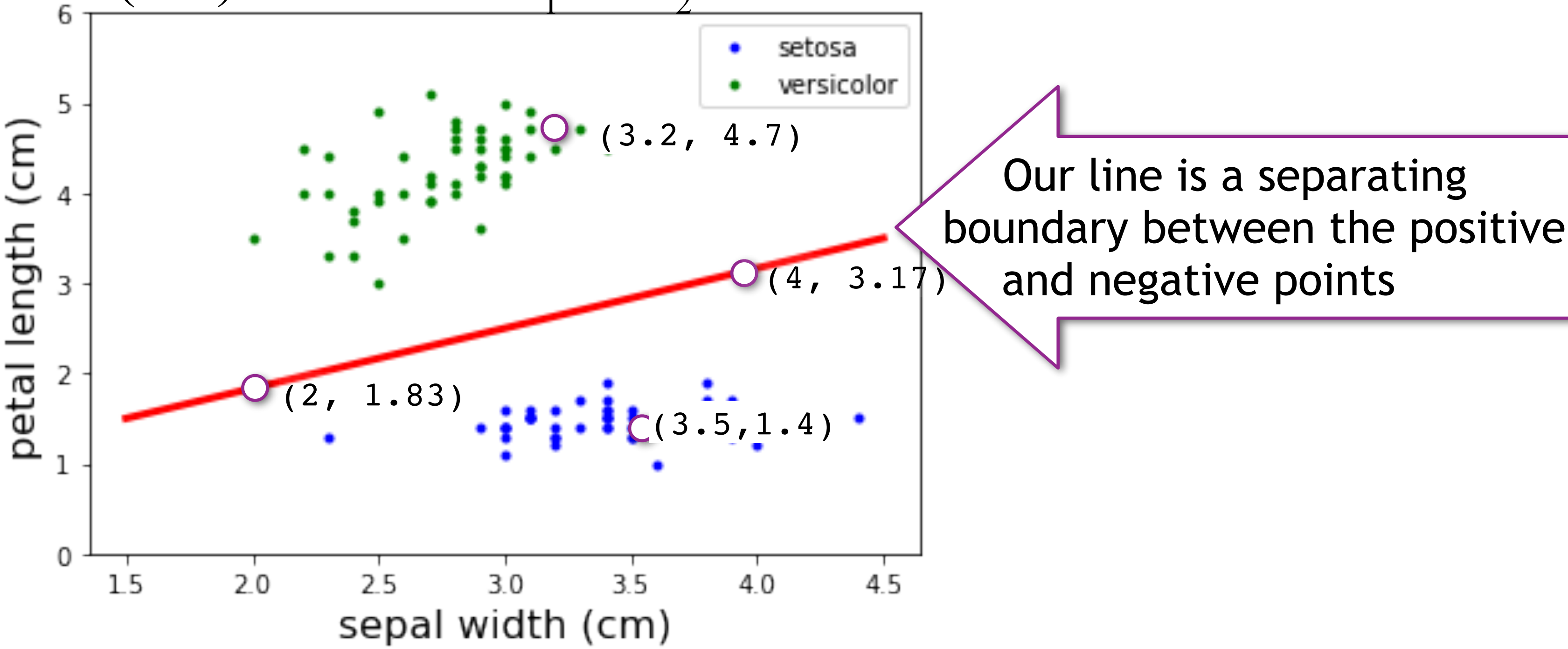
```
[ 3.5   1.4 ]
[ 3.    1.4 ]
[ 3.2   1.3 ]
[ 3.1   1.5 ]
[ 3.6   1.4 ]
[ 3.9   1.7 ]
[ 3.4   1.4 ]
[ 3.4   1.5 ]
[ 2.9   1.4 ]
[ 3.1   1.5 ]
```

- versicolor Iris

```
[ 3.2   4.7 ]
[ 3.2   4.5 ]
[ 3.1   4.9 ]
[ 2.3   4.  ]
[ 2.8   4.6 ]
[ 2.8   4.5 ]
[ 3.3   4.7 ]
[ 2.4   3.3 ]
[ 2.9   4.6 ]
[ 2.7   3.9 ]
```

The line is: 0.5 + 2/3 sepal width - petal length = 0

$$z(\mathbf{x}^{(i)}) = 0.5 + 2/3 x_1^{(i)} - x_2^{(i)}$$



$$z(2,\ 1.83) = 0.5 + 2/3 \cdot 2 - 1.83 \ = 0$$

$$z(4,\ 3.17) = 0.5 + 2/3 \cdot 4 - 3.17 \ = 0$$

$$z(3.5,\ 1.4) = 0.5 + 2/3(3.5) - (1.4) = 2.7$$

$$z(3.2,\ 4.7) = 0.5 + 2/3(3.2) - (4.7) = -2.07$$

Intuition: decision boundary

Data:

$\mathbf{x}_i$= (sepal width$_i$, petal length$_i$)

- setosa Iris

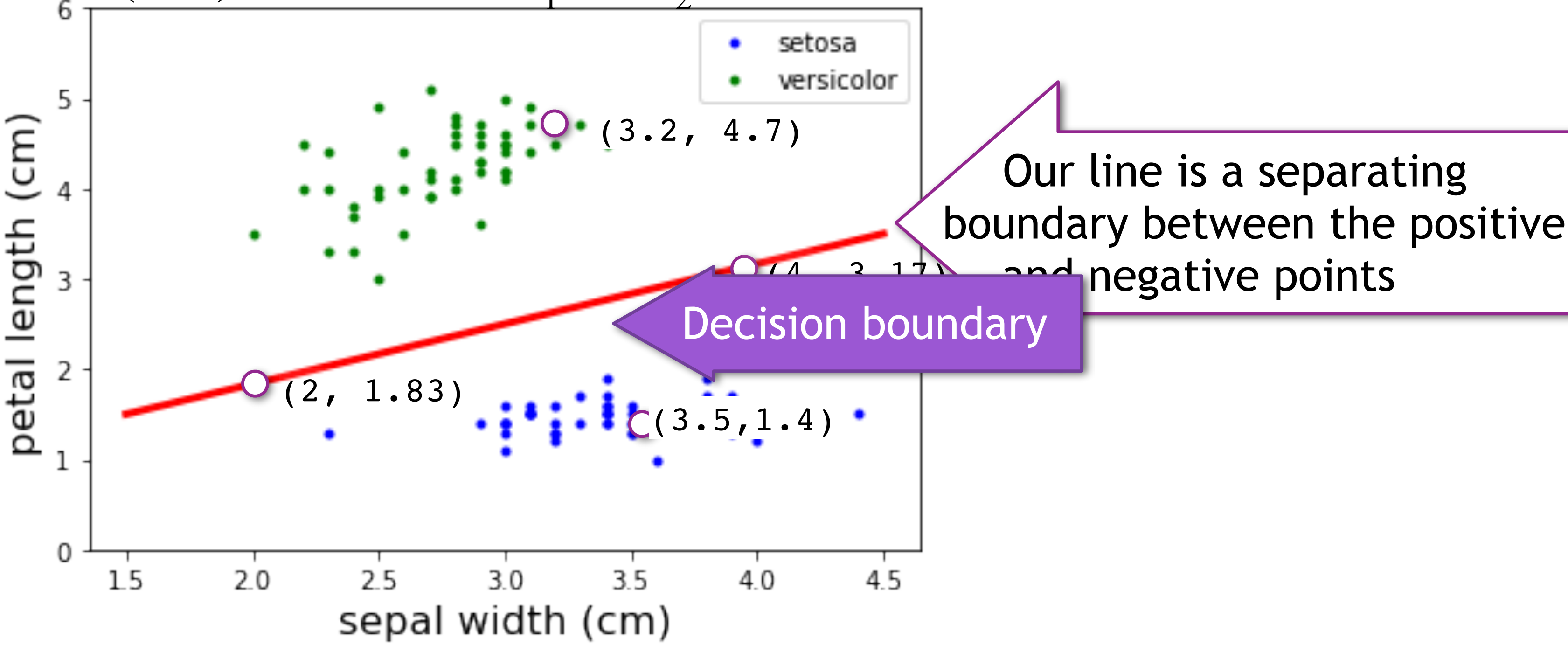
```
[ 3.5   1.4 ]
[ 3.    1.4 ]
[ 3.2   1.3 ]
[ 3.1   1.5 ]
[ 3.6   1.4 ]
[ 3.9   1.7 ]
[ 3.4   1.4 ]
[ 3.4   1.5 ]
[ 2.9   1.4 ]
[ 3.1   1.5 ]
```

- versicolor Iris

```
[ 3.2   4.7 ]
[ 3.2   4.5 ]
[ 3.1   4.9 ]
[ 2.3   4.  ]
[ 2.8   4.6 ]
[ 2.8   4.5 ]
[ 3.3   4.7 ]
[ 2.4   3.3 ]
[ 2.9   4.6 ]
[ 2.7   3.9 ]
```

The line is: 0.5 + 2/3 sepal width - petal length = 0

$$z(\mathbf{x}^{(i)}) = 0.5 + 2/3x_1^{(i)} - x_2^{(i)}$$



$$z(2,\ 1.83) = 0.5 + 2/3 \cdot 2 - 1.83 \ = 0$$

$$z(4,\ 3.17) = 0.5 + 2/3 \cdot 4 - 3.17 \ = 0$$

$$z(3.5,\ 1.4) = 0.5 + 2/3(3.5) - (1.4) = 2.7$$

$$z(3.2,\ 4.7) = 0.5 + 2/3(3.2) - (4.7) = -2.07$$

Linear Classifier

- setosa Iris

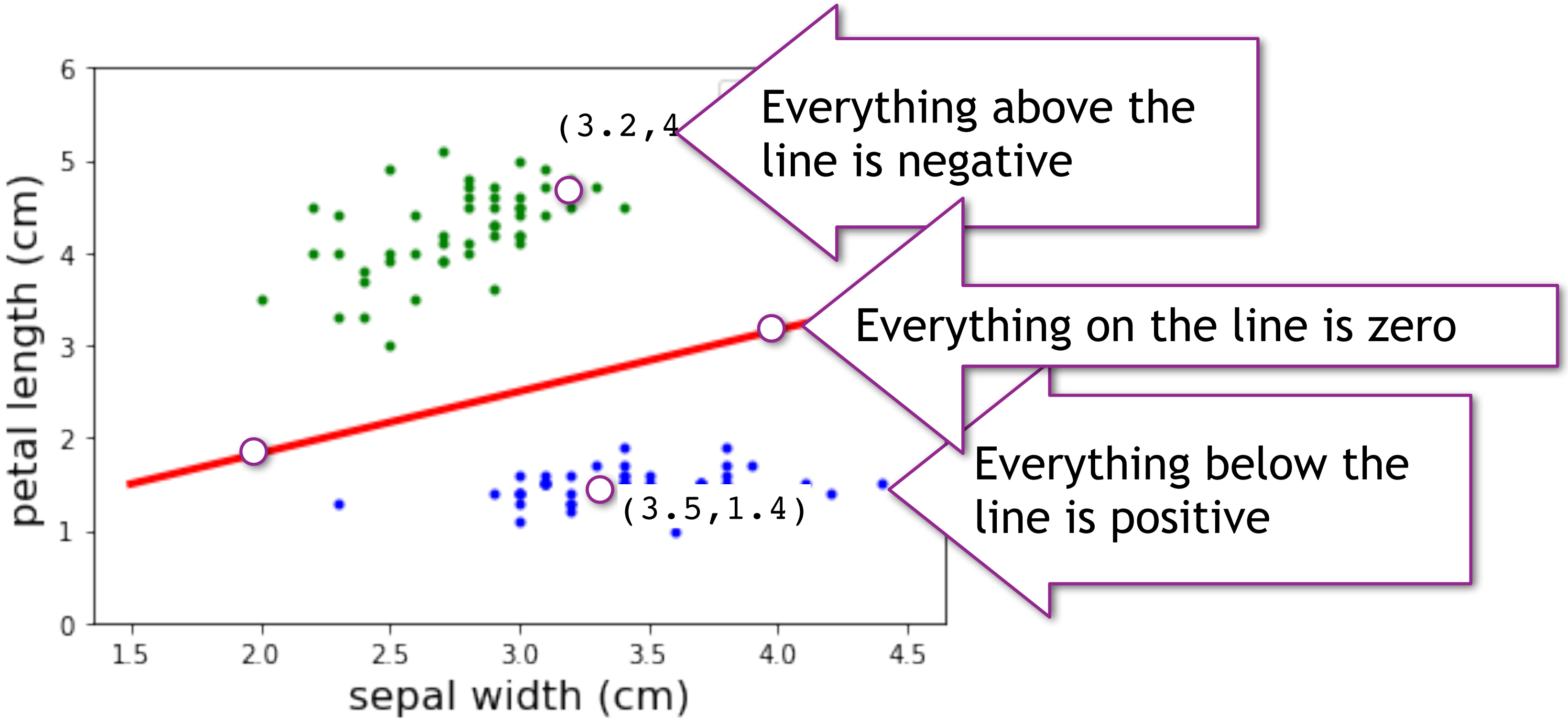
```
[ 3.5   1.4 ]
[ 3.    1.4 ]
[ 3.2   1.3 ]
[ 3.1   1.5 ]
[ 3.6   1.4 ]
[ 3.9   1.7 ]
[ 3.4   1.4 ]
[ 3.4   1.5 ]
[ 2.9   1.4 ]
[ 3.1   1.5 ]
```

- versicolor Iris

```
[ 3.2   4.7 ]
[ 3.2   4.5 ]
[ 3.1   4.9 ]
[ 2.3   4.  ]
[ 2.8   4.6 ]
[ 2.8   4.5 ]
[ 3.3   4.7 ]
[ 2.4   3.3 ]
[ 2.9   4.6 ]
[ 2.7   3.9 ]
```

The line is 0 = + 0.5 + 2/3 sepal width - petal length

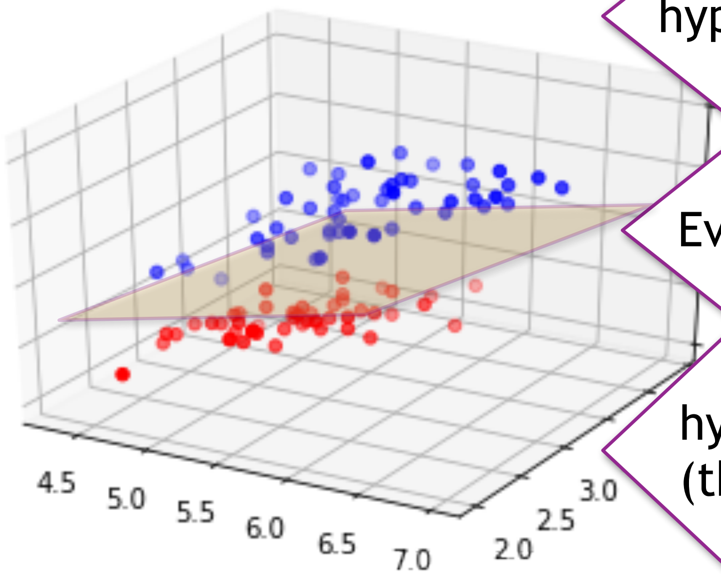
$$z(\mathbf{x}^{(i)}) = 0.5 + 2/3x_1^{(i)} - x_2^{(i)}$$



We will now go back to adding a 1 to every example x

$$\mathbf{x} = \begin{bmatrix} 3 \\ 2.5 \end{bmatrix} \rightarrow \mathbf{x} = \begin{bmatrix} 1 \\ 3 \\ 2.5 \end{bmatrix}$$

Linear classifier in higher dimensions



Everything above the hyperplane has the same sign

Everything on the hyperplane is zero

Everything below the hyperplane has the same sign (the opposite sign of the points above the hyperplane)

Hyperplane:

$$\mathcal{H} = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} = 0\}$$

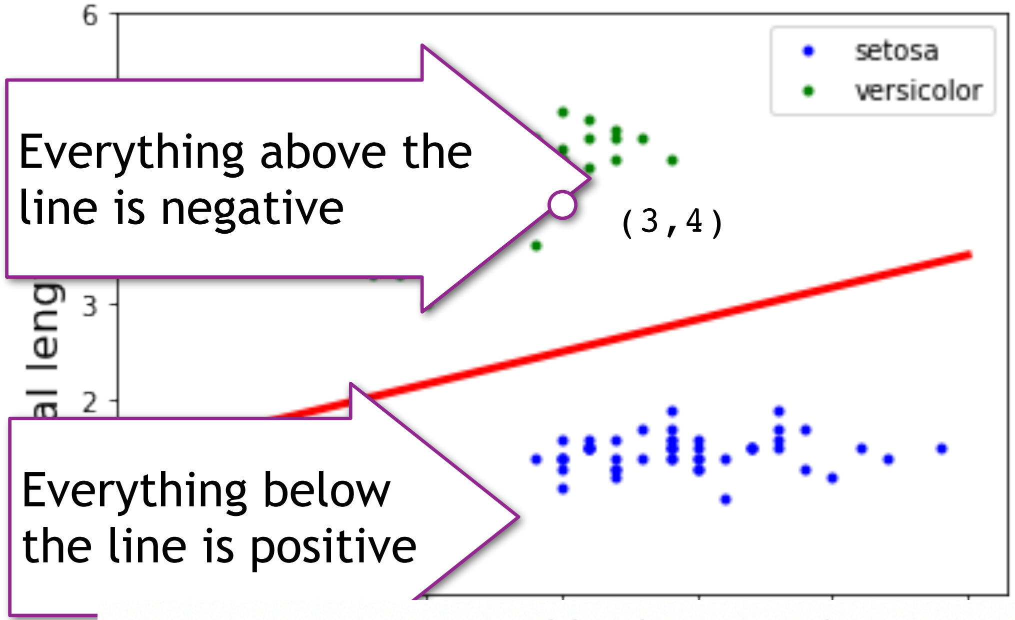
Half-spaces:

$$\mathcal{H}^+ = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} > 0\}$$

$$\mathcal{H}^- = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} < 0\}$$

Prediction using a decision boundary

The line is 0 = 0.5 + 2/3 sepal width - petal length



1. Suppose you found an iris with sepal width = 3 and petal length = 4.

If you knew it was either a setosa iris or a versicolor iris, could you predict which type it was?

- setosa iris
- versicolor
- cannot predict using the information that is given

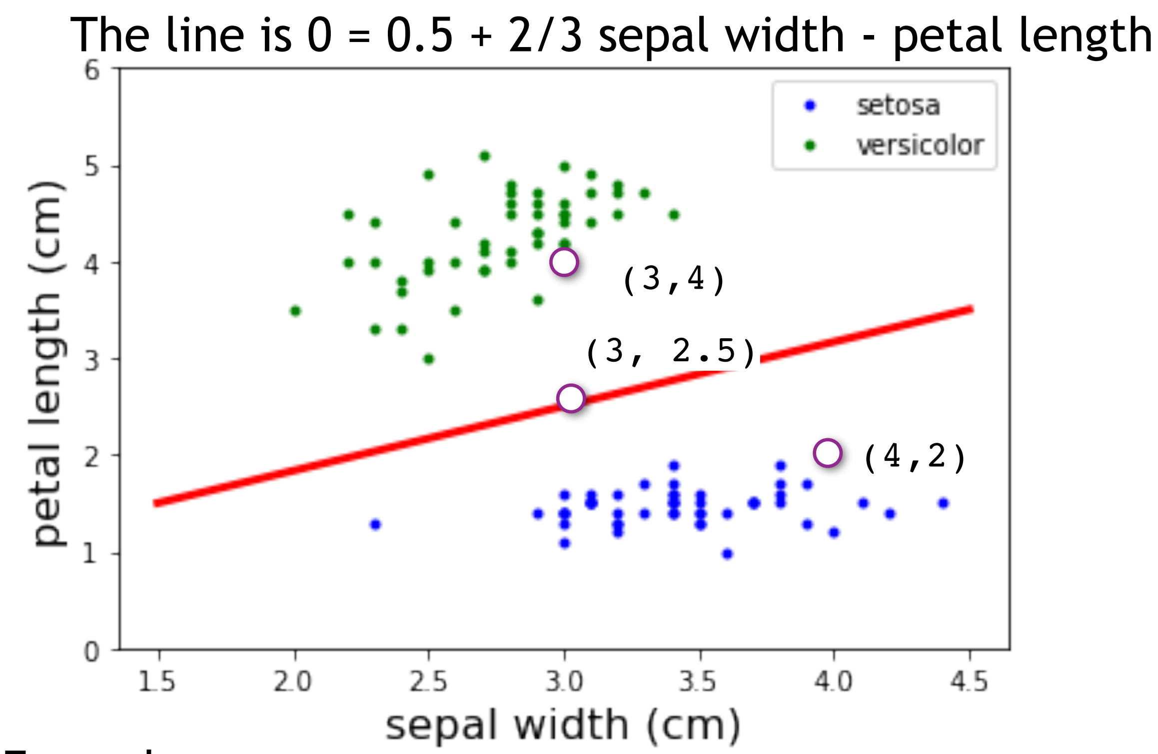
How can we predict the label of a new example when $\mathbf{w} = \begin{bmatrix} 0.5 \\ 2/3 \\ -1 \end{bmatrix}$?

$$h(\mathbf{x}) = \begin{cases} 1 & \mathbf{w}^T\mathbf{x} \geq 0 \\ 0 & \mathbf{w}^T\mathbf{x} < 0 \end{cases}$$

Setosa

Versicolor

Prediction using a decision boundary



How can we predict the label of a new example when $\mathbf{w} = \begin{bmatrix} 0.5 \\ 2/3 \\ -1 \end{bmatrix}$?

$$h(\mathbf{x}) = \begin{cases} 1 & \mathbf{w}^T\mathbf{x} \geq 0 \\ 0 & \mathbf{w}^T\mathbf{x} < 0 \end{cases}$$

Setosa

Versicolor

<u>Examples</u>

(3,4)

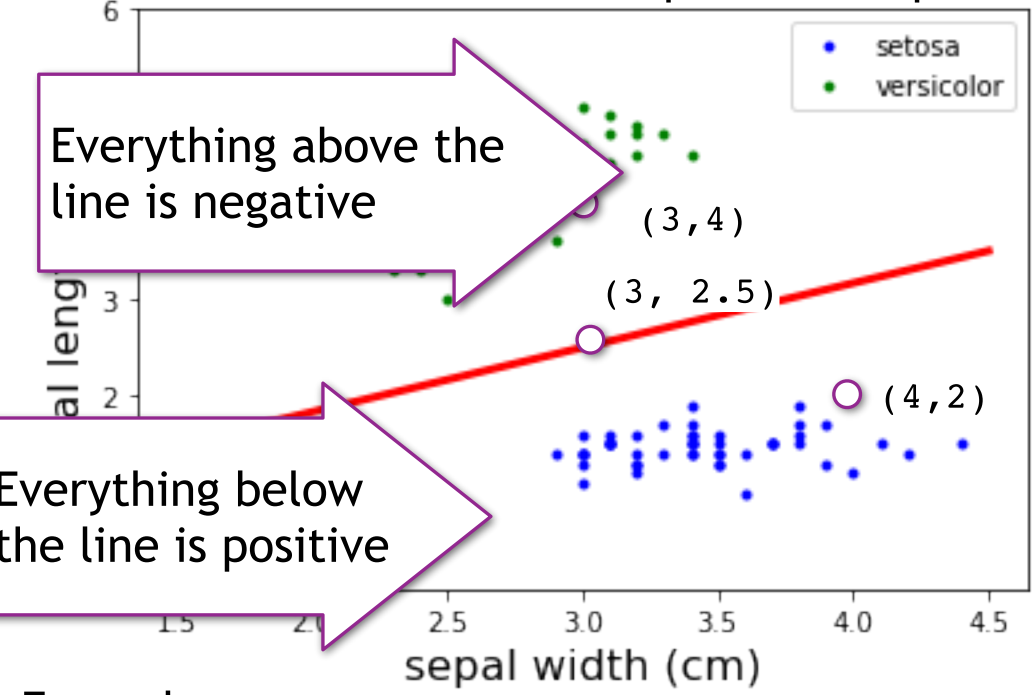
$$\mathbf{x} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} \qquad \mathbf{w}^T\mathbf{x} = \begin{bmatrix} 0.5 & 2/3 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} = -1.5$$

(4,2)

$$\mathbf{x} = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} \qquad \mathbf{w}^T\mathbf{x} = \begin{bmatrix} 0.5 & 2/3 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} = 1/2 + 8/3 - 2$$

Prediction using a decision boundary

The line is 0 = 0.5 + 2/3 sepal width - petal length



How can we predict the label of a new example when $\mathbf{w} = \begin{bmatrix} 0.5 \\ 2/3 \\ -1 \end{bmatrix}$?

$$h(\mathbf{x}) = \begin{cases} 1 & \mathbf{w}^T\mathbf{x} \geq 0 \\ 0 & \mathbf{w}^T\mathbf{x} < 0 \end{cases}$$

Setosa

Versicolor

<u>Examples</u>

(3,4)

$$\mathbf{x} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} \qquad \mathbf{w}^T\mathbf{x} = \begin{bmatrix} 0.5 & 2/3 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} = -1.5$$

(4,2)

$$\mathbf{x} = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} \qquad \mathbf{w}^T\mathbf{x} = \begin{bmatrix} 0.5 & 2/3 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} = 1/2 + 8/3 - 2$$

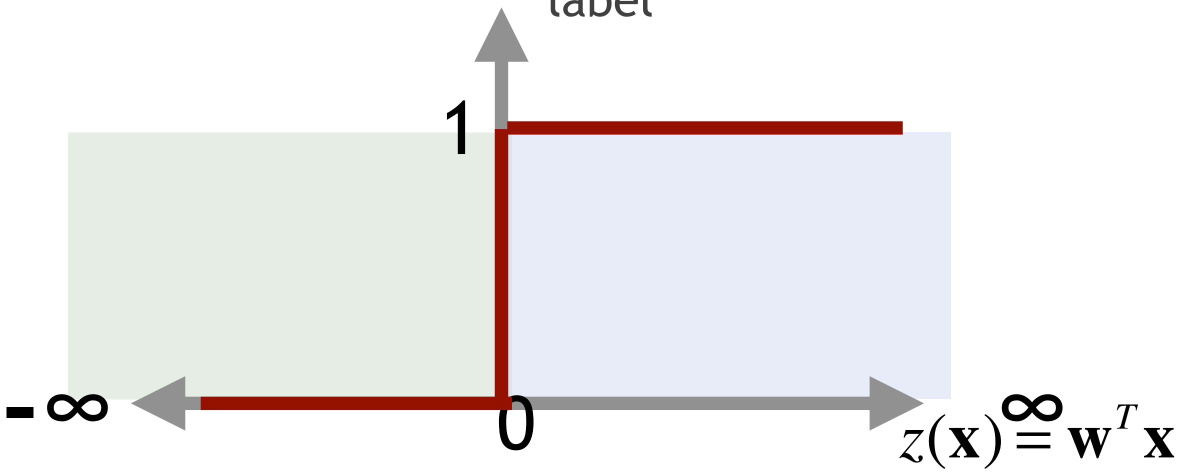
Visualizing a linear classifier

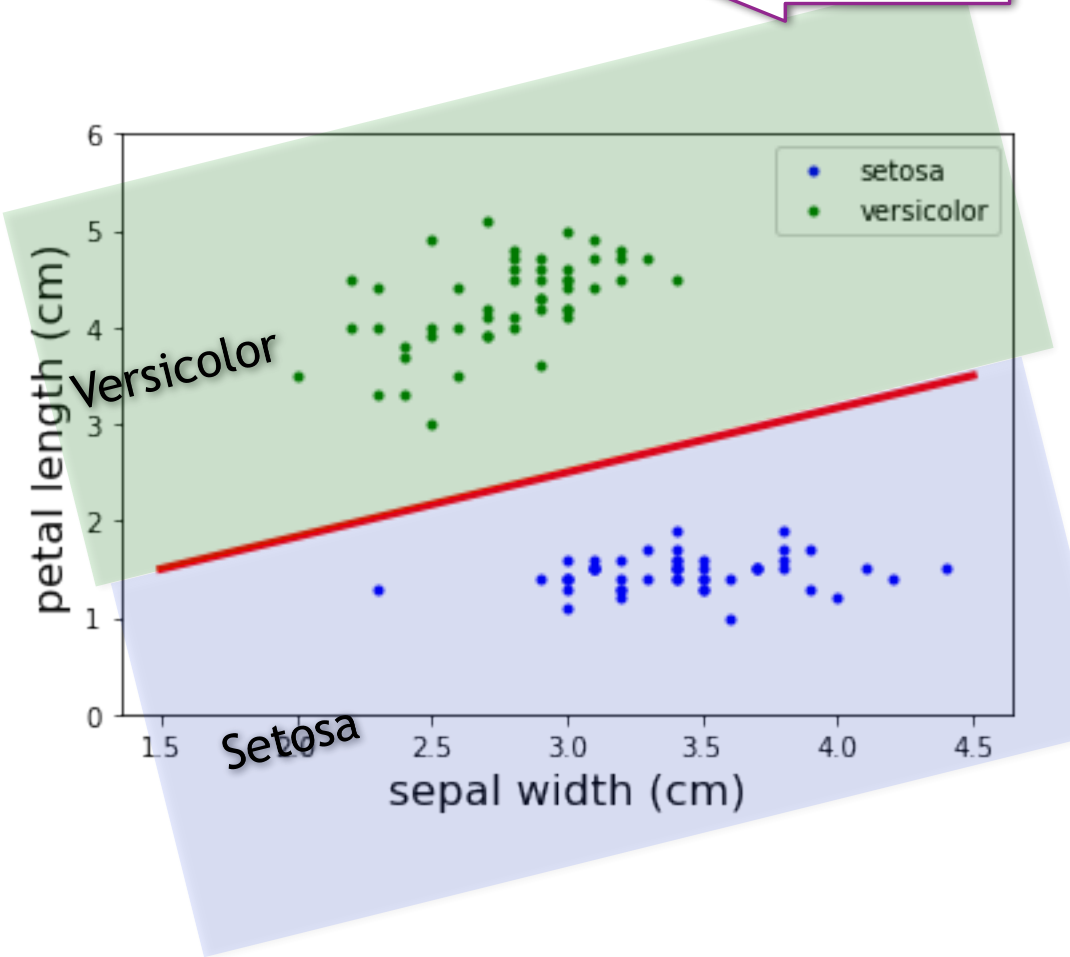
For a feature vector $\mathbf{x} = [1, x_1, x_2]^T$

$$h(\mathbf{x}) = \begin{cases} 1 & \mathbf{w}^T\mathbf{x} \geq 0 \\ 0 & \mathbf{w}^T\mathbf{x} < 0 \end{cases}$$

Setosa

Versicolor





Hyperplane:

$$\mathcal{H} = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} = 0\}$$

Half-spaces:

$$\mathcal{H}^+ = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} > 0\}$$

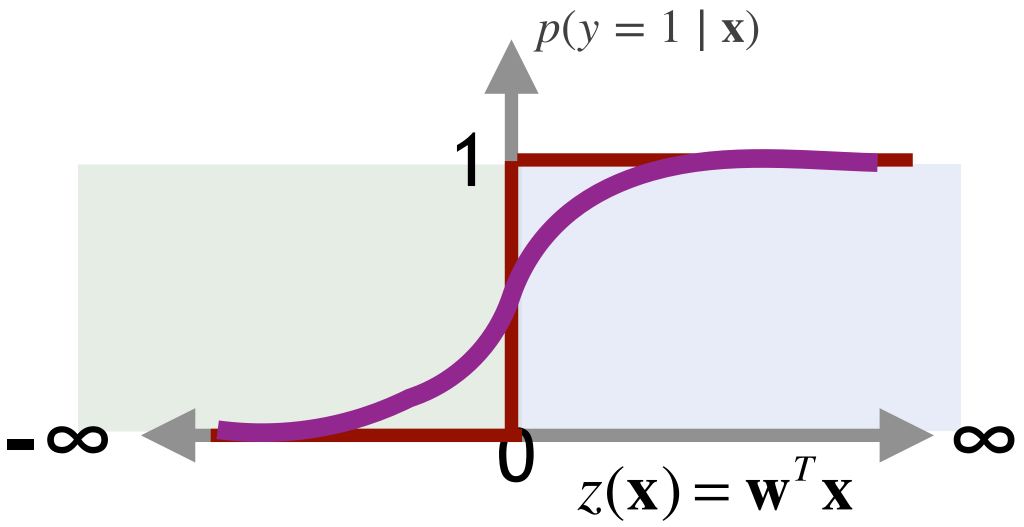
$$\mathcal{H}^- = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} < 0\}$$

Outline

- Motivating example: How can we classify? — How can we use a hyperplane for a classification problem?
- Estimating probabilities — Can we predict not only which class an example belongs to - but a confidence score of that classification
- Maximum likelihood — How can we find the most likely hyperplane? Could we write a function to describe how likely a hyperplane was to have generated the dataset?
- Thinking about different types of error — Some errors are more costly than other errors. Can we modify our predictions to decrease one type of error (and perhaps increase another type of error?
- Transformation of the features — Extending our algorithm to nonlinear decision boundaries
- Multiple classes — What if we have more than two classes?

Could we modify the hypothesis to give more information about how confident we are in our prediction….

Intuition: Logistic Regression



How confident are we of our prediction?

Instead of returning a label, let us return a probability.

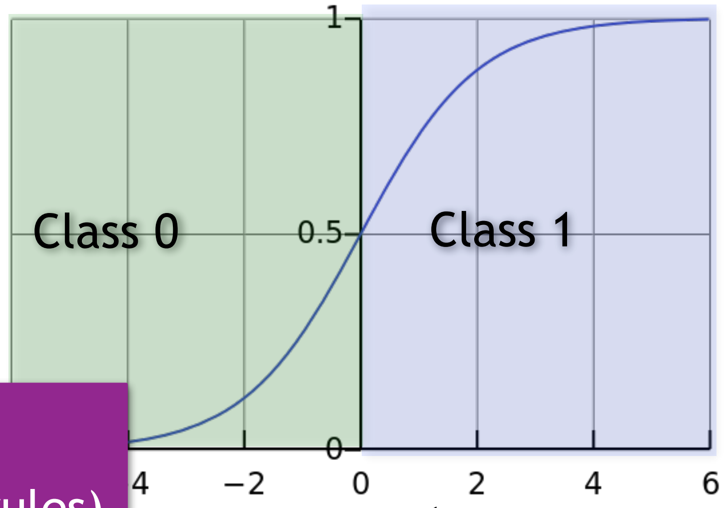
We need a <u>function</u> that takes $\mathbf{w}^T\mathbf{x}$ and returns a number between 0 and 1.

Note: We still have to find $\mathbf{w}$

$\sigma(\cdot)$ Logistic function (sigmoid function)

Other functions could be used - but this works well

Squashing function



$$-\infty < \quad z(\mathbf{x}) = \mathbf{w}^T\mathbf{x} \quad < \infty$$

-∞ 0 ∞

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

0 1

Monotonically increasing
Thus we can derive classification rules)

$\sigma(z)$ bounded between 0 and 1
Thus we can interpret as probability

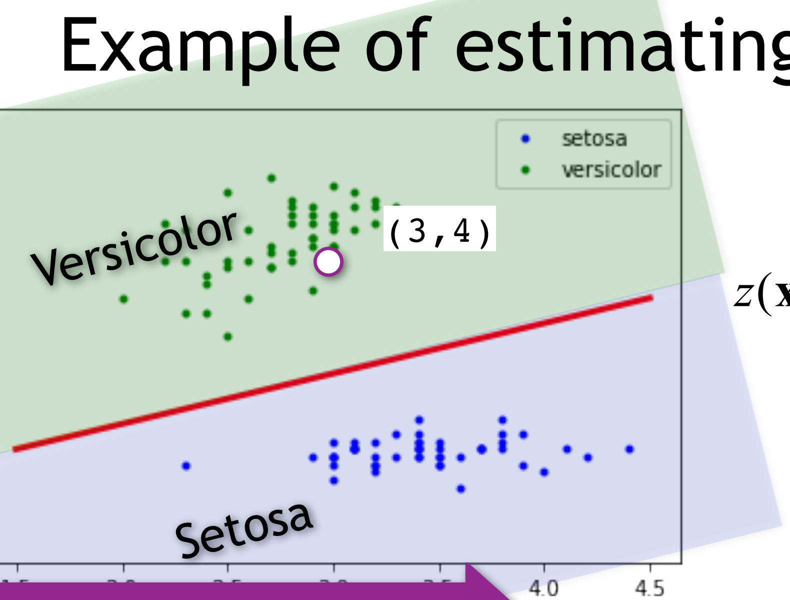
$$\sigma(\infty) = \frac{1}{1+e^{-\infty}} = 1$$

$$\sigma(-\infty) = \frac{1}{1+e^{\infty}} = 0$$

$$\sigma(0) = \frac{1}{1+e^{0}} = \frac{1}{2} = 0.5$$

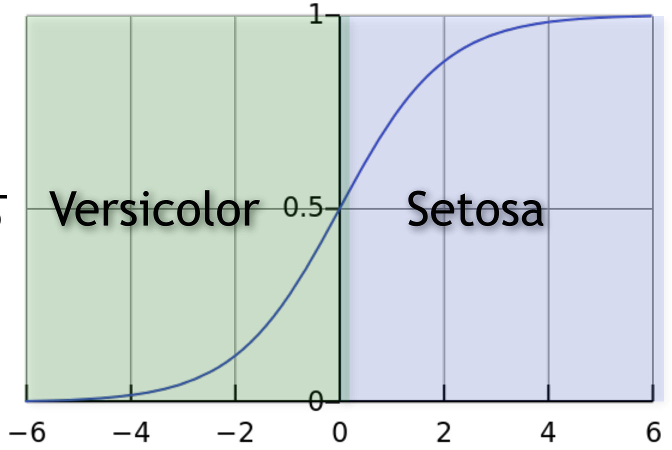
Note that:
$\sigma(-z) = 1 - \sigma(z)$

Example of estimating the prob. of $(\mathbf{x}, y)$ belonging to class 1 using $\sigma(\,\cdot\,)$



$$z(\mathbf{x}^{(i)}) = \mathbf{w}^T\mathbf{x}^{(i)} = \mathbf{w}^T \begin{bmatrix} 1 \\ x_1^{(i)} \\ \vdots \\ x_d^{(1)} \end{bmatrix}$$

$$\sigma(\mathbf{w}^T\mathbf{x}^{(i)}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}^{(i)}}}$$



Probability of the label

$$p(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}^{(i)})^{y^{(i)}}\left(1-\sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)^{1-y^{(i)}} = \begin{cases} \sigma(\mathbf{w}^T\mathbf{x}^{(i)}) & \text{for } y^{(i)} = 1 \\ 1-\sigma(\mathbf{w}^T\mathbf{x}^{(i)}) & \text{for } y^{(i)} = 0 \end{cases}$$

Exploiting the fact that $y^{(i)}$ is 0 or 1

Examples: $z(\mathbf{x}^{(i)}) = 0.5 + 2/3x_1^{(i)} - x_2^{(i)}$

$(3,4)$

$$z([1,3,4]; \mathbf{w}) = -1.5$$

$$\sigma(z(1,3,4)) = \frac{1}{1+e^{1.5}} = .182$$

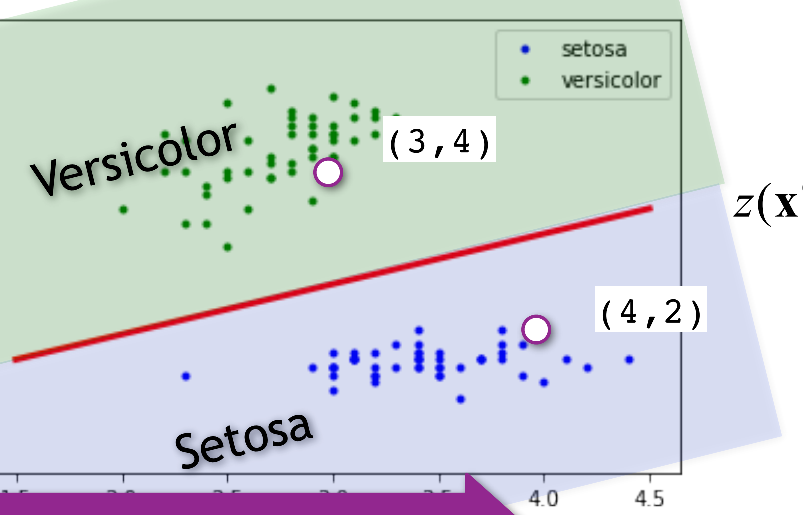
$$p(y = 1 \mid [1,3,4]^T; \mathbf{w}) = .182 = (.182)^1(1-.182)^{1-1}$$

$$p(y = 0 \mid [1,3,4]^T; \mathbf{w})$$

Pair share

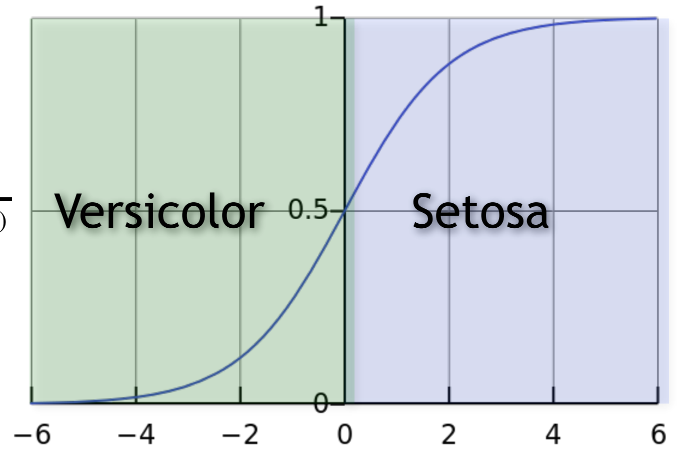
"Notational note: In the expression p(y|**x**; **w**) the semicolon indicates that **w** is a parameter, not a random variable that is being conditioned on, even though it is to the right of the vertical bar. "

Example of estimating the prob. of $(\mathbf{x}, y)$ belonging to class 1 using $\sigma(\cdot)$



$$z(\mathbf{x}^{(i)}) = \mathbf{w}^T\mathbf{x}^{(i)} = \mathbf{w}^T \begin{bmatrix} 1 \\ x_1^{(i)} \\ \vdots \\ x_d^{(1)} \end{bmatrix}$$

$$\sigma(\mathbf{w}^T\mathbf{x}^{(i)}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}^{(i)}}}$$



Probability of the label

$$p(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}^{(i)})^{y^{(i)}}\left(1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)^{1-y^{(i)}} = \begin{cases} \sigma(\mathbf{w}^T\mathbf{x}^{(i)}) & \text{for } y^{(i)} = 1 \\ 1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)}) & \text{for } y^{(i)} = 0 \end{cases}$$

Examples: $z(\mathbf{x}^{(i)}) = 0.5 + 2/3x_1^{(i)} - x_2^{(i)}$

$(3,4)$

$$z([1,3,4]; \mathbf{w}) = -1.5$$

$$\sigma(z(1,3,4)) = \frac{1}{1 + e^{1.5}} = .182$$

$$p(y = 1 \mid [1,3,4]^T; \mathbf{w}) = .182 = (.182)^1(1 - .182)^{1-1}$$

$$p(y = 0 \mid [1,3,4]^T; \mathbf{w}) = 1 - .182 = (.182)^0(1 - .182)^{1-0}$$

$(4,2)$

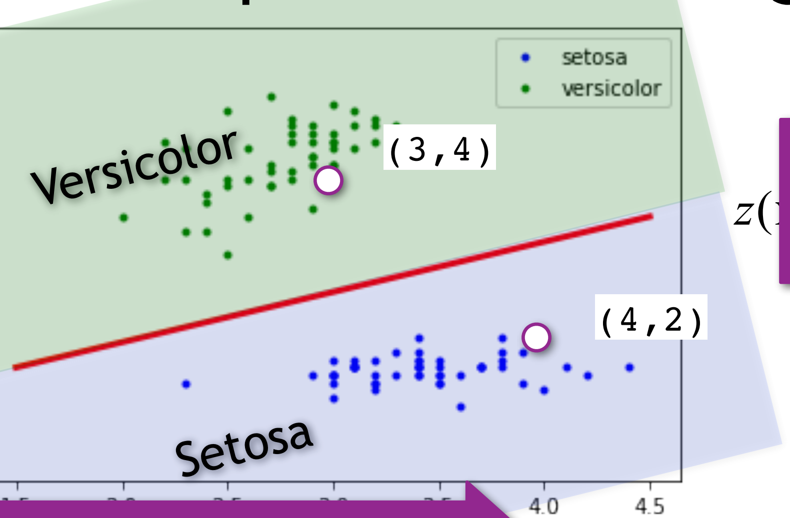
$$z([1,4,2]; \mathbf{w}) = 1.67$$

$$\sigma(z(1,4,2)) = \frac{1}{1 + e^{-1.67}} = .763$$

$$p(y = 1 \mid [1,4,2]^T; \mathbf{w}) = .763$$

$$p(y = 0 \mid [1,4,2]^T; \mathbf{w})$$

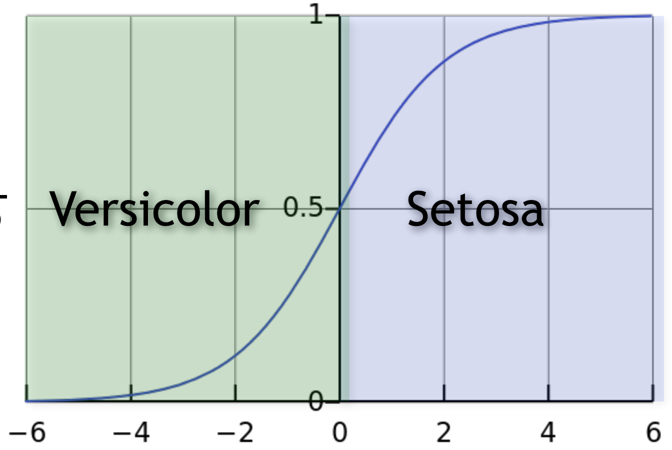
Example of estimating the prob. of $(\mathbf{x}, y)$ belonging to class 1 using $\sigma(\cdot)$



$z($ … $\begin{bmatrix} 1 \\ \vdots \\ x_d^{(1)} \end{bmatrix}$

Probability of being class 1

$$\sigma(\mathbf{w}^T\mathbf{x}^{(i)}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}^{(i)}}}$$



Probability of the label

$$p(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}^{(i)})^{y^{(i)}}\left(1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)^{1-y^{(i)}} = \begin{cases} \sigma(\mathbf{w}^T\mathbf{x}^{(i)}) & \text{for } y^{(i)} = 1 \\ 1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)}) & \text{for } y^{(i)} = 0 \end{cases}$$

Examples: $z(\mathbf{x}^{(i)}) = 0.5 + 2/3x_1^{(i)} - x_2^{(i)}$

$(3,4)$

$$z([1,3,4]; \mathbf{w}) = -1.5$$

$$\sigma(z(1,3,4)) = \frac{1}{1 + e^{1.5}} = .182$$

$$p(y = 1 \mid [1,3,4]^T; \mathbf{w}) = .182 = (.182)^1(1 - .182)^{1-1}$$

$$p(y = 0 \mid [1,3,4]^T; \mathbf{w}) = 1 - .182 = (.182)^0(1 - .182)^{1-0}$$

$(4,2)$

$$z([1,4,2]; \mathbf{w}) = 1.67$$

$$\sigma(z(1,4,2)) = \frac{1}{1 + e^{-1.67}} = .763$$

$$p(y = 1 \mid [1,4,2]^T; \mathbf{w}) = .763$$

$$p(y = 0 \mid [1,4,2]^T; \mathbf{w})$$