

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 3

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?

Classification vs Regression

- Regression we were given:

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

- Classification we are given:

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

- 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?

1 - setosa

0 - versicolor



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



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

Classification vs Regression

□ Regression we were given:

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

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]

we are given:

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

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 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?

1 - setosa

0 - versicolor



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



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

Intuition

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

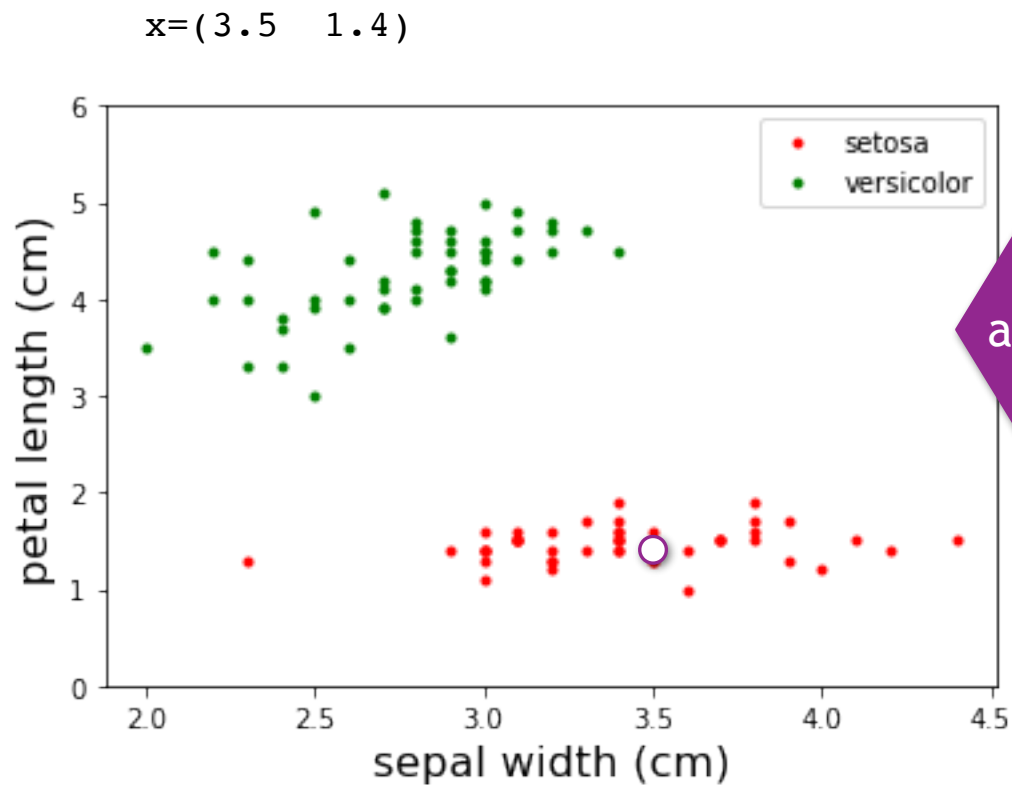
`x=(sepal width, petal length)`

❑ 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 ]
```

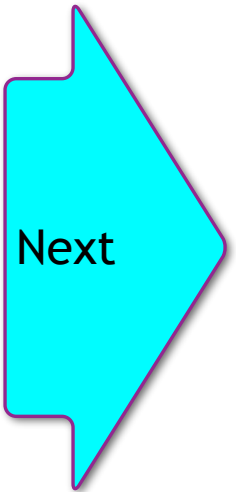
❑ versicolor Iris

```
[ 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 ]
```



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

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



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

Intuition: decision boundary

Pair share: The orange vector normal to the red line (hyperplane), describe it as column vector.

Data:

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

□ setosa Iris

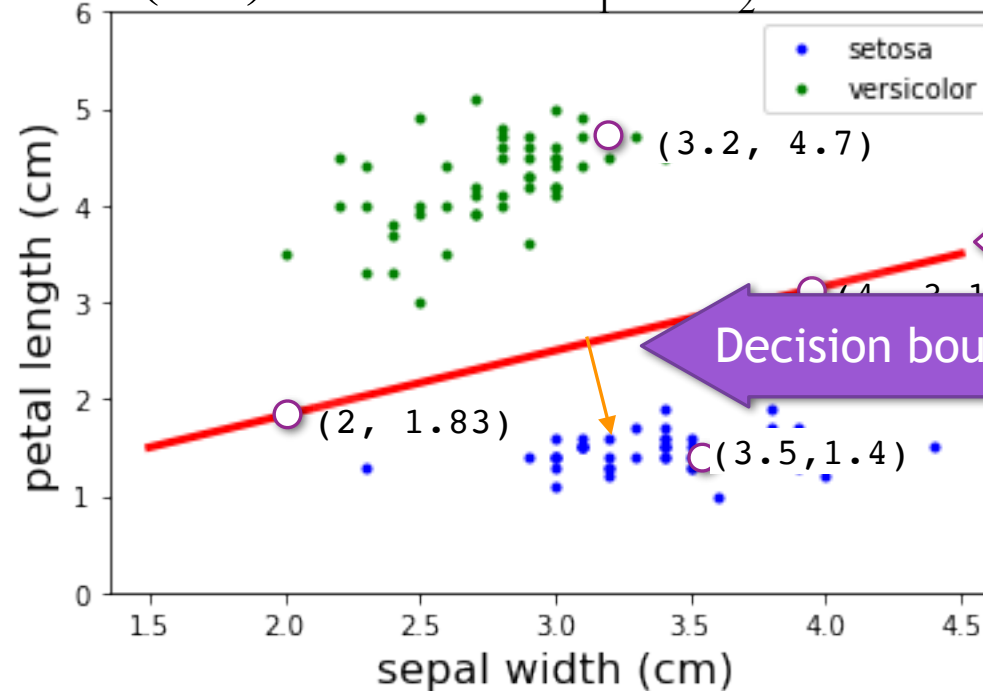
```
[ 3.5  1.4 ]  
[ 3.   1.4 ]  
[ 3.2  1.3 ]  
[ 3.1  1.5 ]  
[ 3.6  1.4 ]  
[ 3.9  1.7 ]  
[ 3.4  1.4 ]  
[ 3.4  1.5 ]  
[ 2.9  1.4 ]  
[ 3.1  1.5 ]
```

□ 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: $0.5 + 2/3 \text{ sepal width} - \text{petal length} = 0$

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



Our line is a separating boundary between the positive and negative points

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

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

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

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

Linear Classifier

□ setosa Iris

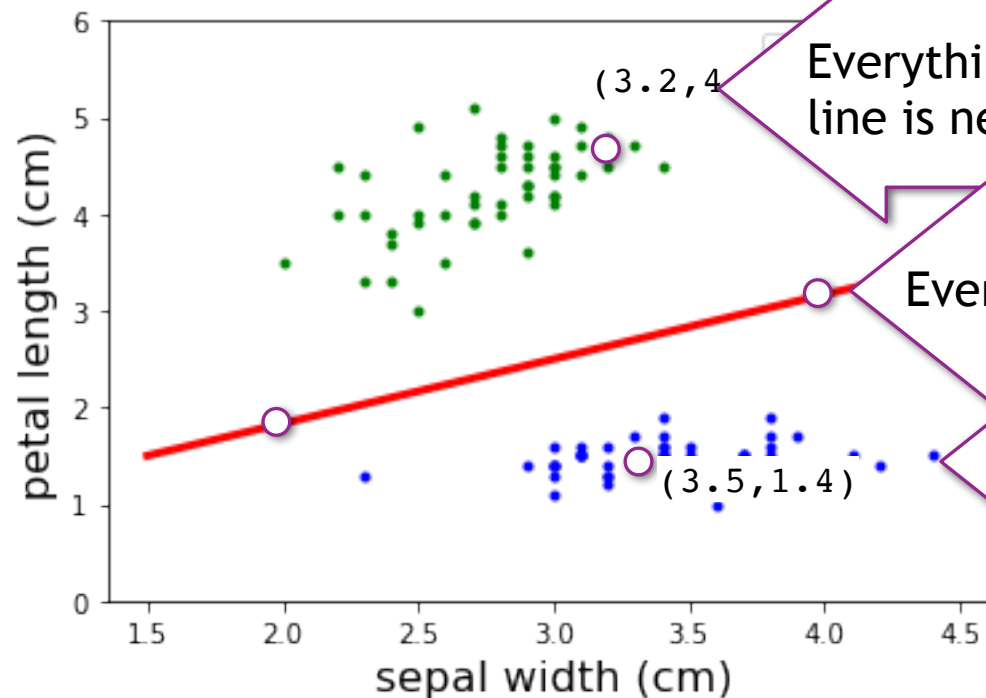
[3.5	1.4]
[3.	1.4]
[3.2	1.3]
[3.1	1.5]
[3.6	1.4]
[3.9	1.7]
[3.4	1.4]
[3.4	1.5]
[2.9	1.4]
[3.1	1.5]

□ 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 $0 = + 0.5 + 2/3 \text{ sepal width} - \text{petal length}$

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

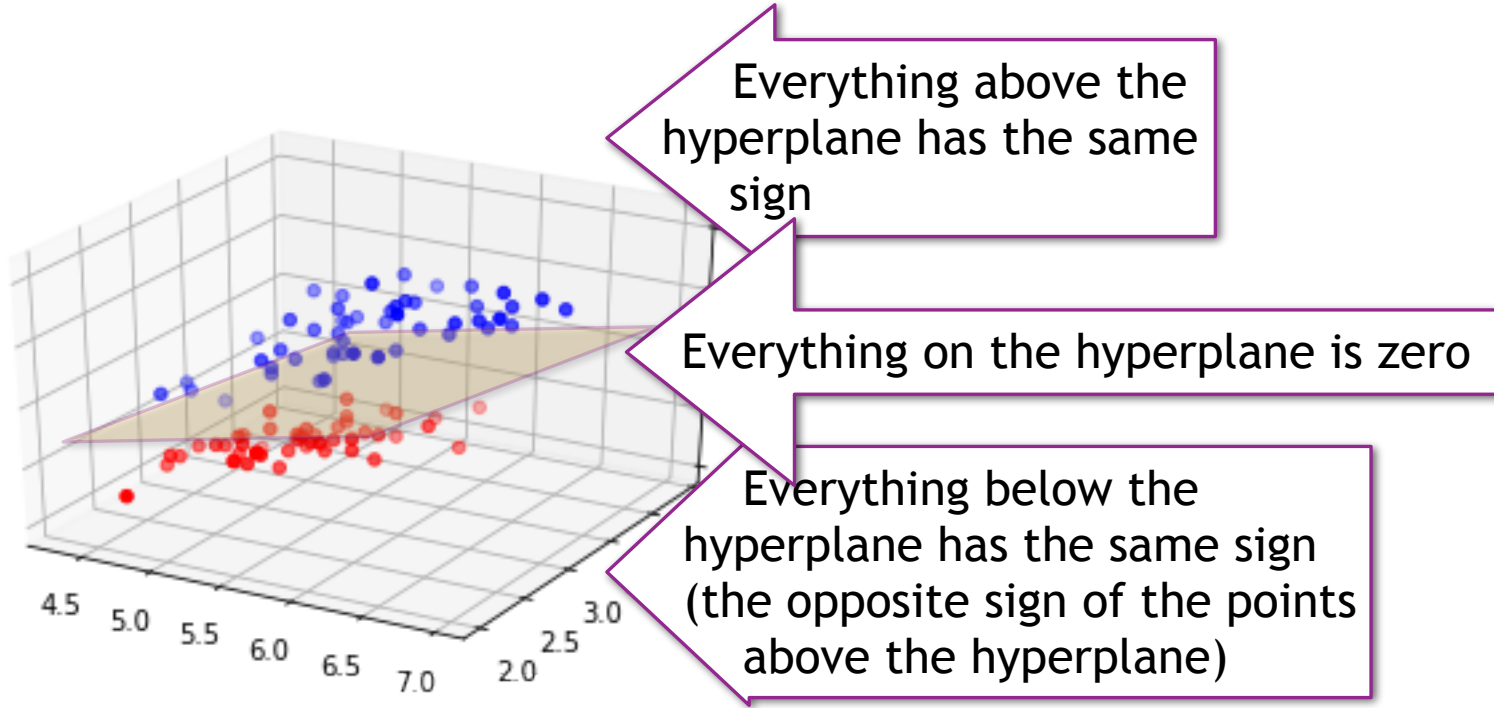


Pair share: What change would you make to have the separating line (hyperplane) in the same place, but to classify all the points labeled 'positive' in the diagram as negative and all the points labeled 'negative' in the diagram as positive?

We will now go back
to adding a 1 to every
example \mathbf{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



Hyperplane:

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

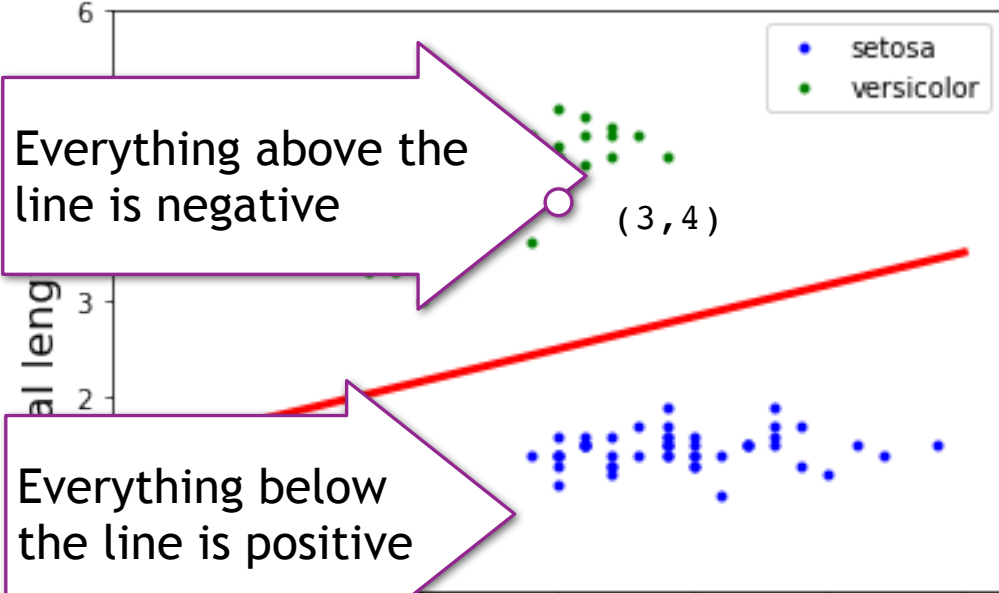
Half-spaces:

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

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

Prediction using a decision boundary

The line is $0 = 0.5 + 2/3 \text{ sepal width} - \text{petal length}$



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

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

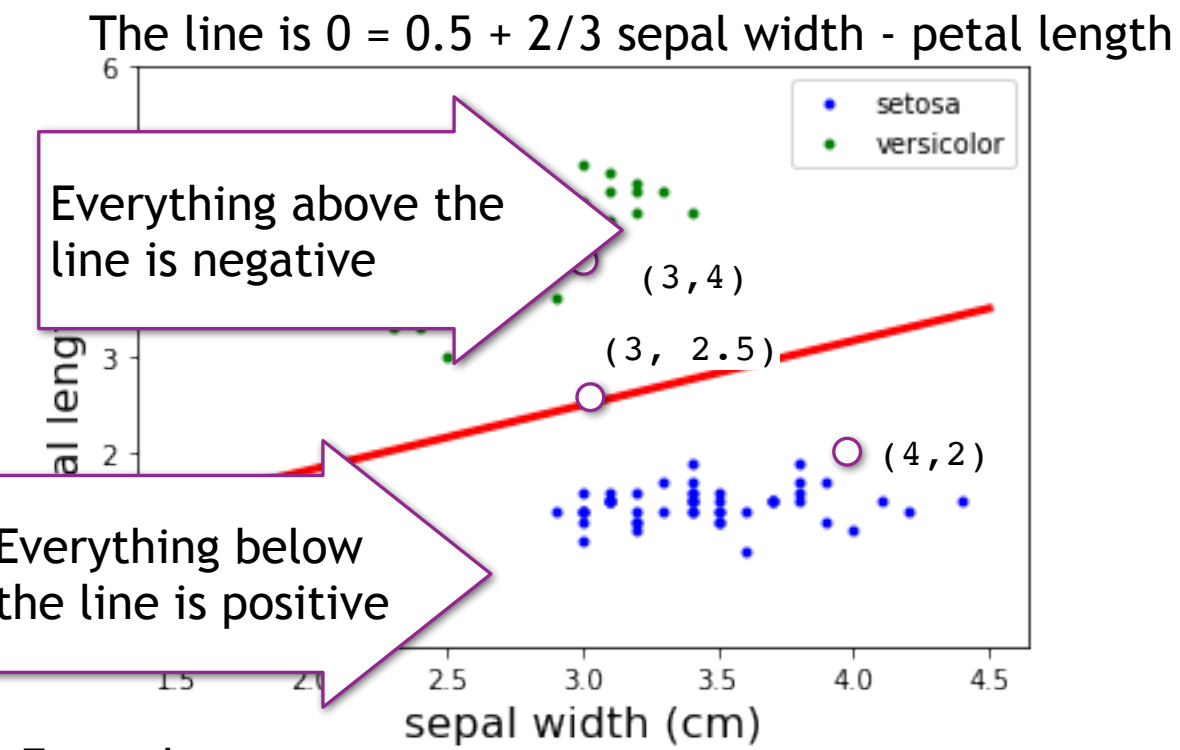
☐ cannot predict using the information that is given

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

Setosa

Versicolor

Prediction using a decision boundary



Examples

(3,4)

$$\mathbf{x} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}$$

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

(4,2)

$$\mathbf{x} = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix}$$

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

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

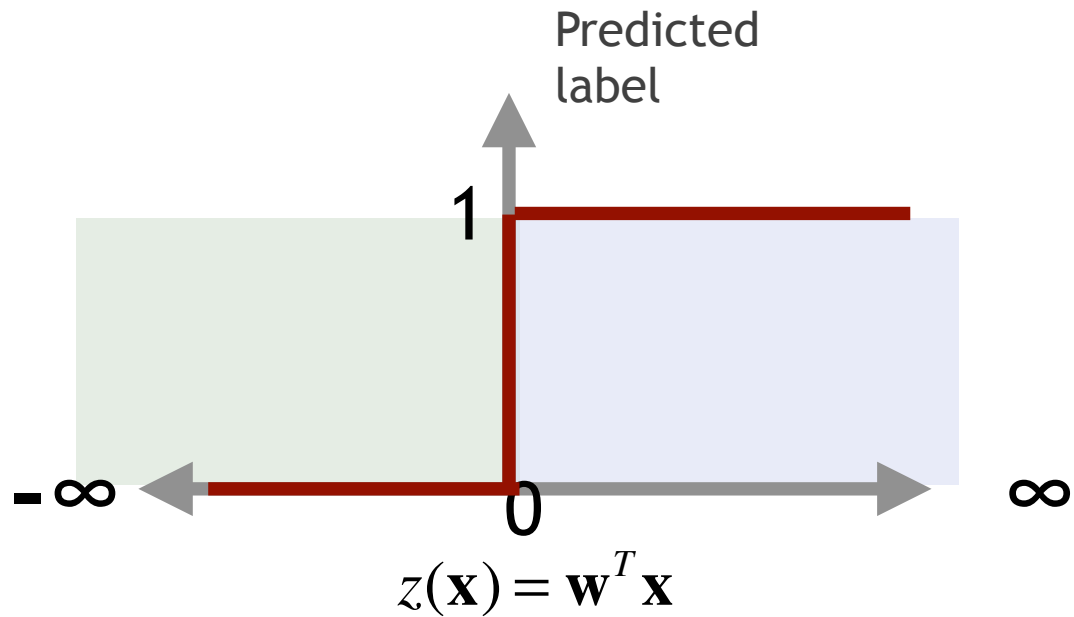
Visualizing a linear classifier

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



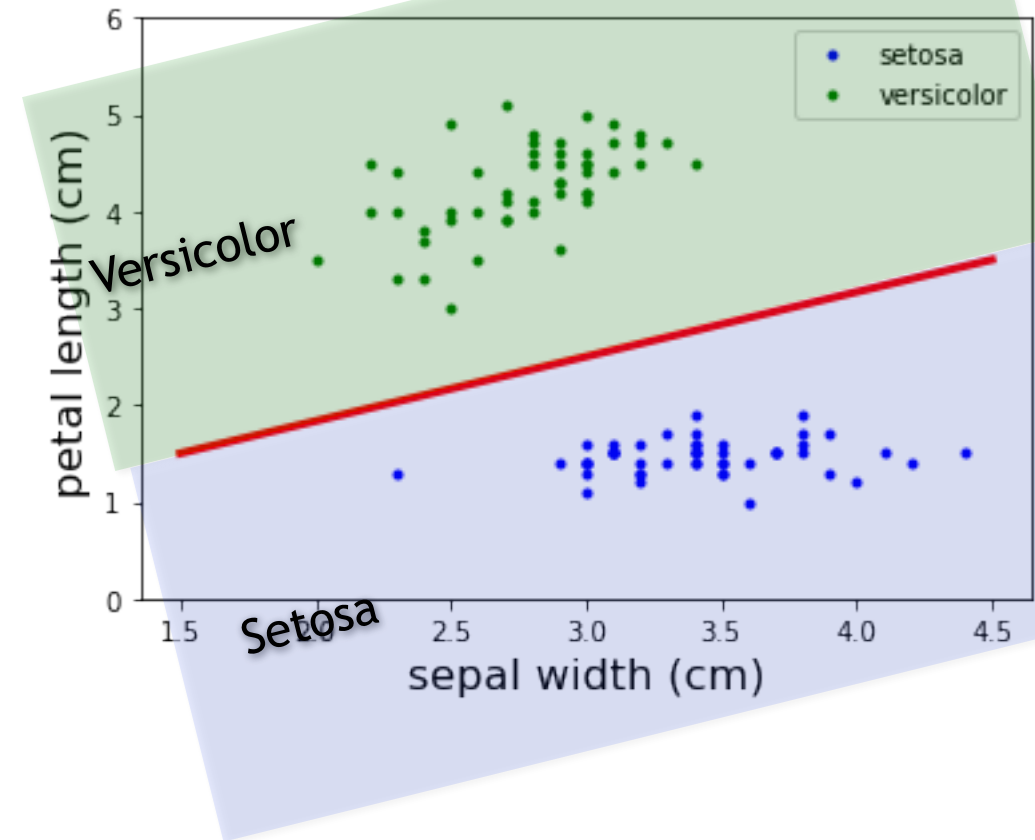
Hyperplane:

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


Half-spaces:

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


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



Outline

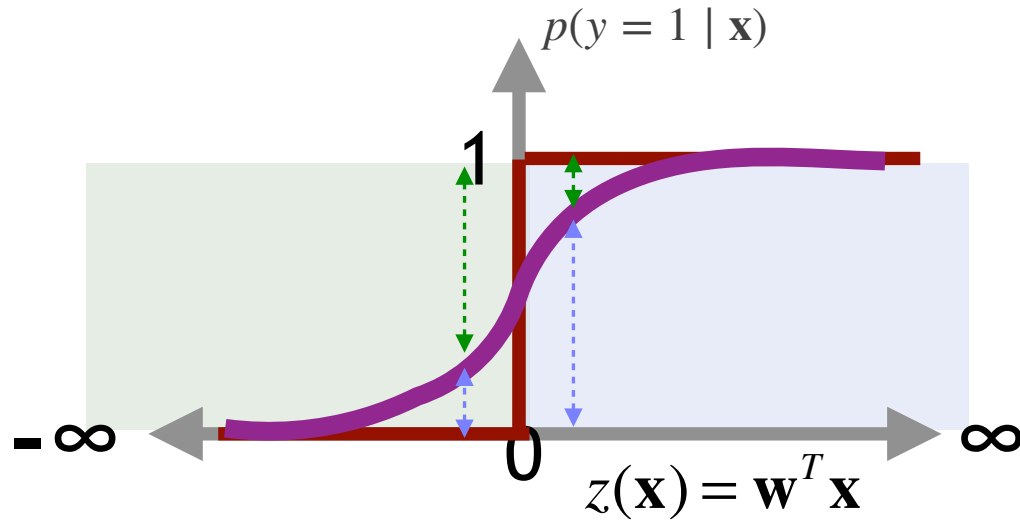
- ❑ Motivating example: How can we classify? } How can we use a hyperplane for a classification problem?
-  ❑ Estimating probabilities } Can we predict not only which class an example belongs to - but a confidence score of that classification
- ❑ Maximum likelihood } How can we find the most likely hyperplane? 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?

Outline

- ❑ Motivating example: How can we classify? } How can we use a hyperplane for a classification problem?
-  Which model } 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 function 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

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


$-\infty$ 0 ∞

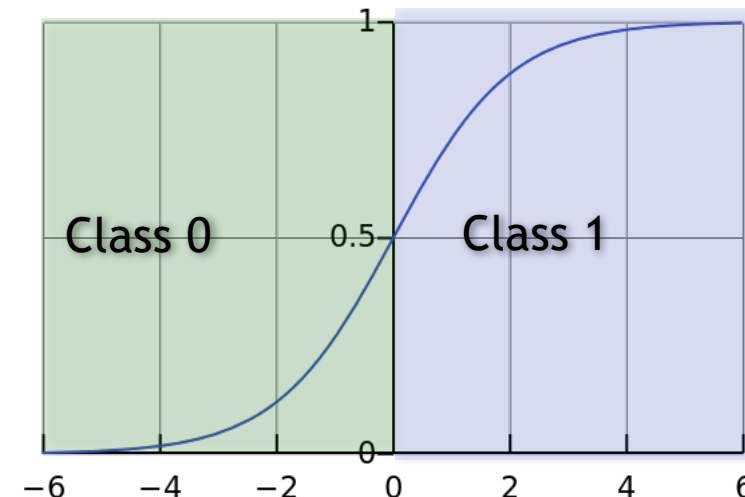
$$\sigma(z(\mathbf{x})) = \frac{1}{1 + e^{-z(\mathbf{x})}} = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x})}}$$

0 1

Monotonically increasing

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

Squashing function



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

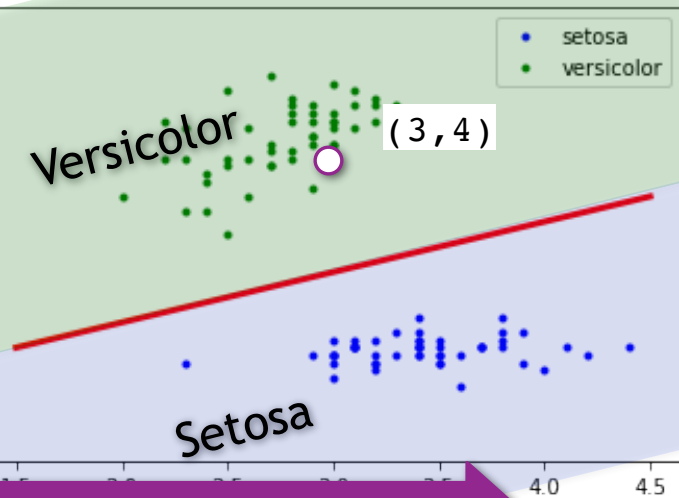
$$\sigma(-\infty) = \frac{1}{1 + e^{\infty}} = 0$$

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

Pair share: Why is the output of σ is always in the interval (0, 1)?
Why can't it equal 0 or equal 1?
For what value of z does $\sigma(z) = 0.5$?

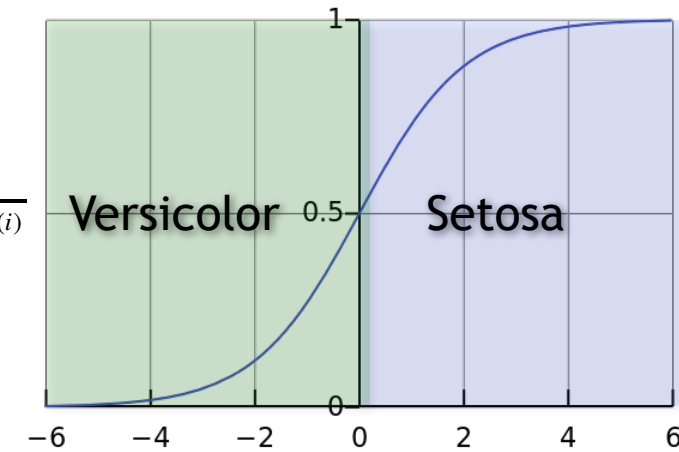
Note that:
 $\sigma(-z) = 1 - \sigma(z)$

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



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

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



Probability of the label

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

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)

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

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

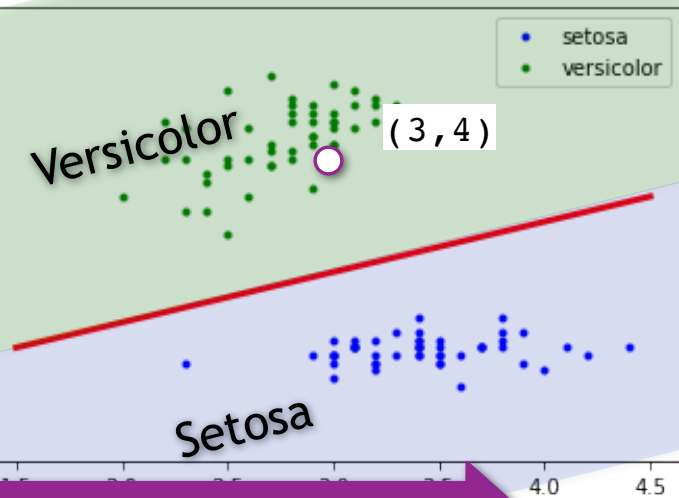
$$p(y = 1 | [1,3,4]^T; \mathbf{w}) = .182 = (.182)^1 (1 - .182)^{1-1}$$

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

Pair share

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

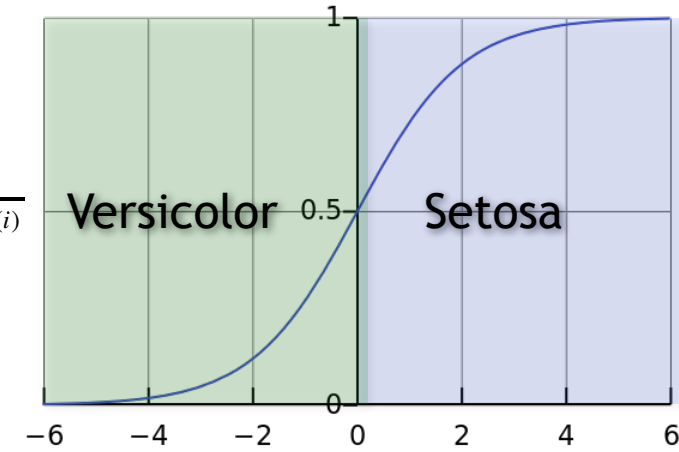
Example of estimating the prob. of (\mathbf{x}, y) belonging to class 1 using $\sigma(\cdot)$



Probability of being class 1

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

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



Probability of the label

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

Examples: $z(\mathbf{x}^{(i)}) = 0.5 + 2/3x_1^{(i)} - x_2^{(i)}$
(3,4)

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

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

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

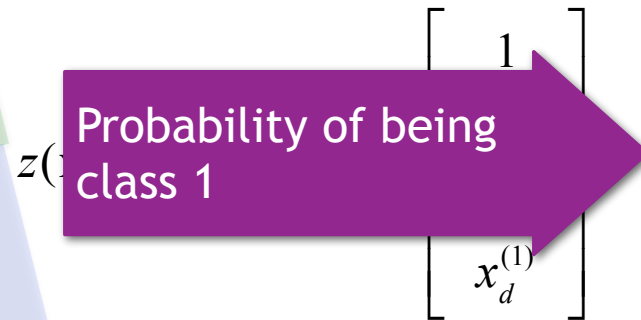
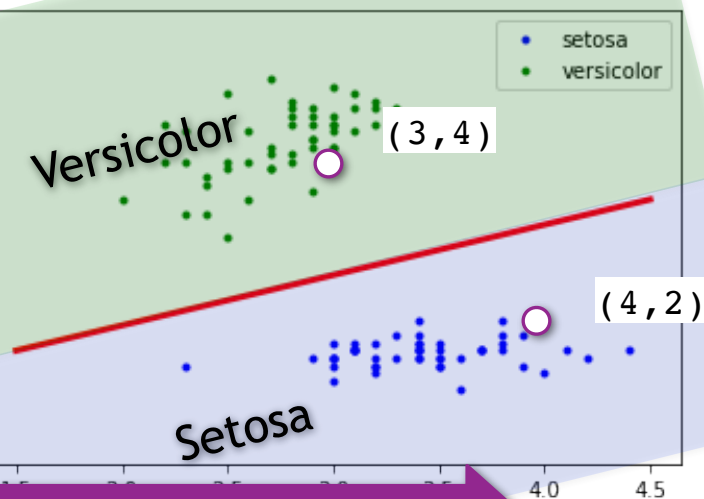
$$p(y = 1 | [1,3,4]^T; \mathbf{w}) = .182 = (.182)^1 (1 - .182)^{1-1}$$

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

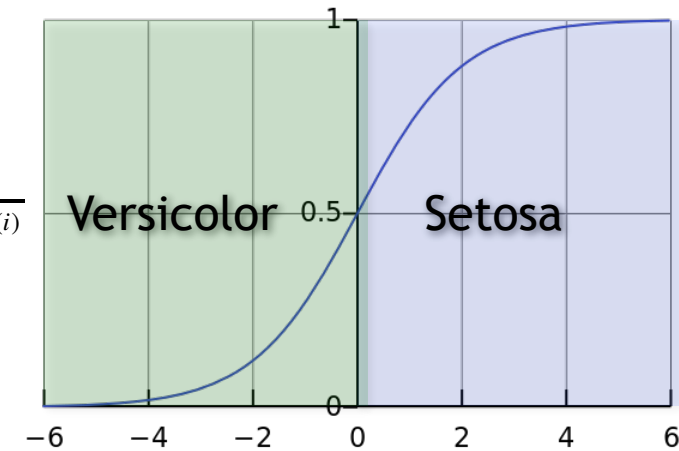
Pair share

“Notational note: In the expression $p(y|\mathbf{x}; \mathbf{w})$ the semicolon indicates that \mathbf{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 $\sigma(\cdot)$



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



Probability of the label

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

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

(3,4)

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

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

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

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

(4,2)

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

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

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

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

Logistic Regression

Data: $(\mathbf{x}^{(i)}, y^{(i)}), i = 1, 2, \dots, N$ where $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{0, 1\}$

model: Logistic function applied to $\mathbf{w}^T \mathbf{x}$

$$p(y = 1 \mid \mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x})}$$

We just developed an intuition on why this makes sense

Next we will show why this is true

And find an optimizer to find the "best" \mathbf{w}


Learning: find parameters that maximizes the **objective function**:

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} \left(\sum_{i=1}^N y^{(i)} \ln(\sigma(\mathbf{w}^T \mathbf{x}^{(i)})) + (1 - y^{(i)}) \ln(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)})) \right)$$

$$\text{where } \sigma(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x})}$$

Prediction: $\hat{y} = \arg \max_{y \in \{0, 1\}} p(y \mid \mathbf{x}; \mathbf{w})$ or $\hat{y} = p(y \mid \mathbf{x}; \mathbf{w})$

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?
 - Iterative approach - gradient ascent } Maximizing the function
- 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?
- 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?
 - Its Optimizer } gradient ascent } Maximizing the function
- 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?

Given $D = \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(N)}, y^{(N)})\}$
how can we find the “best”
hyperplane, \mathbf{w} ?

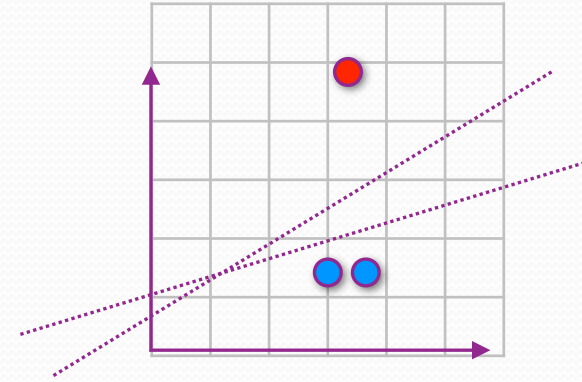


Optimize \mathbf{w}

We first need to decide what makes
one hyperplane better than another
(i.e. an objective function) **Pair share**

MLE!

Likelihood of seeing the data



- Our model: $p(y^{(i)} | \mathbf{x}^{(i)}; \mathbf{w}) = \begin{cases} \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) & \text{for } y^{(i)} = 1 \\ 1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) & \text{for } y^{(i)} = 0 \end{cases}$

- Given the following data:

$\mathbf{x}^{(1)} = [1, 3.2 \quad 4.7]$	$y^{(1)} = 0$
$\mathbf{x}^{(2)} = [1, 3.5 \quad 1.4]$	$y^{(2)} = 1$
$\mathbf{x}^{(3)} = [1, 3. \quad 1.4]$	$y^{(3)} = 1$

- How likely were we to see the data if the line was:

$$\mathbf{w} = \begin{bmatrix} 1/2 \\ 2/3 \\ -1 \end{bmatrix} \quad \underbrace{\left(1 - \frac{1}{1 + e^{-(1/2 + (2/3)3.2 - 4.7)}}\right)}_{1-0.11} \quad \underbrace{\left(\frac{1}{1 + e^{-(1/2 + (2/3)3.5 - 1.4)}}\right)}_{0.81} \quad \underbrace{\left(\frac{1}{1 + e^{-(1/2 + (2/3)3 - 1.4)}}\right)}_{0.75} = 0.54$$

$$\mathbf{w} = \begin{bmatrix} 1 \\ 1/3 \\ -1 \end{bmatrix} \quad \left(1 - \frac{1}{1 + e^{-(1 + (1/3)3.2 - 4.7)}}\right) \quad \left(\frac{1}{1 + e^{-(1 + (1/3)3.5 - 1.4)}}\right) \quad \left(\frac{1}{1 + e^{-(1 + (1/3)3 - 1.4)}}\right) = 0.41$$

Pair share: Write the conditional likelihood function for these three examples

Our model:

$$p(y^{(i)} | \mathbf{x}^{(i)}; \mathbf{w}) = \begin{cases} \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) & \text{for } y^{(i)} = 1 \\ 1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) & \text{for } y^{(i)} = 0 \end{cases}$$
$$p(y = 1 | \mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x})}}$$
$$p(y = 0 | \mathbf{x}; \mathbf{w}) = 1 - \sigma(\mathbf{w}^T \mathbf{x}) = 1 - \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x})}}$$

$$D = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), (\mathbf{x}^{(3)}, y^{(3)})\}$$

setosa

$$\begin{aligned} \mathbf{x}^{(2)} &= [1, 3.5, 1.4] & y^{(2)} &= 1 \\ \mathbf{x}^{(3)} &= [1, 3., 1.4] & y^{(3)} &= 1 \end{aligned}$$



https://en.wikipedia.org/wiki/Iris_flower_data_set#/media/



<https://commons.wikimedia.org/>

versicolor

$$\mathbf{x}^{(1)} = [1, 3.2, 4.7] \quad y^{(1)} = 0$$

The conditional likelihood function

versicolor

$$\mathbf{x}^{(1)} = [1 \quad 3.2 \quad 4.7] \quad y^{(1)} = 0$$

setosa

$$\mathbf{x}^{(2)} = [1 \quad 3.5 \quad 1.4] \quad y^{(2)} = 1$$

$$\mathbf{x}^{(3)} = [1 \quad 3. \quad 1.4] \quad y^{(3)} = 1$$

$$p(y = 1 \mid \mathbf{x}; \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}$$

$$p(y = 0 \mid \mathbf{x}; \mathbf{w}) = 1 - \sigma(\mathbf{w}^T \mathbf{x}) = 1 - \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}$$

$$L(\mathbf{w}) = \left(1 - \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x}^{(1)})}}\right) \left(\frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x}^{(2)})}}\right) \left(\frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x}^{(3)})}}\right)$$

$$L(\mathbf{w}) = \left(1 - \frac{1}{1 + e^{-(w_0 + w_1 3.2 + w_2 4.7)}}\right) \left(\frac{1}{1 + e^{-(w_0 + w_1 3.5 + w_2 1.4)}}\right) \left(\frac{1}{1 + e^{-(w_0 + w_1 3 + w_2 1.4)}}\right)$$

$$L(\mathbf{w}) = (1 - p(y = 1 \mid \mathbf{x}^{(1)}; \mathbf{w})) p(y = 1 \mid \mathbf{x}^{(2)}; \mathbf{w}) p(y = 1 \mid \mathbf{x}^{(3)}; \mathbf{w}) = \prod_{i=1}^N p(y^{(i)} \text{ correctly predicted} \mid \mathbf{x}^{(i)}; \mathbf{w})$$

$$L(\mathbf{w}) = \prod_{i: y^{(i)}=1} p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w}) \prod_{i: y^{(i)}=0} (1 - p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w}))$$

The conditional likelihood function

Conditional likelihood function (conditioned on \mathbf{x}). Larger value means more likely

$$L(\mathbf{w}) = \prod_{i:y^{(i)}=1} p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w}) \prod_{i:y^{(i)}=0} (1 - p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w}))$$

Here we assume all the examples are independent

$$\prod_{i:y^{(i)}=1} p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w})^{y^{(i)}} (1 - p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w}))^{1-y^{(i)}} \prod_{i:y^{(i)}=0} (1 - p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w}))^{1-y^{(i)}} p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w})^{y^{(i)}}$$

$$L(\mathbf{w}) = \prod_{i=1}^N p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w})^{y^{(i)}} (1 - p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w}))^{1-y^{(i)}}$$

Define: $p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x}^{(i)})$

$$= \prod_{i=1}^N \sigma(\mathbf{w}^T \mathbf{x}^{(i)})^{y^{(i)}} (1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}))^{1-y^{(i)}}$$

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

Pair share: how do we find the \mathbf{w} that maximizes this function

$$\text{Maximize } L(\mathbf{w}) = \prod_{i=1}^N \sigma(\mathbf{w}^T \mathbf{x}^{(i)})^{y^{(i)}} (1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}))^{1-y^{(i)}}$$

The log-likelihood function

Define: $p(y^{(i)} = 1 \mid \mathbf{x}^{(i)}; \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x})$
 $= \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}$

□ We wanted to maximize $L(\mathbf{w}) = \prod_{i=1}^N \sigma(\mathbf{w}^T \mathbf{x}^{(i)})^{y^{(i)}} (1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}))^{1-y^{(i)}}$

□ This is the same as maximizing $\ell(\mathbf{w}) = \ln(L(\mathbf{w}))$

$$= \ln \left[\prod_{i=1}^N \sigma(\mathbf{w}^T \mathbf{x}^{(i)})^{y^{(i)}} (1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}))^{1-y^{(i)}} \right]$$

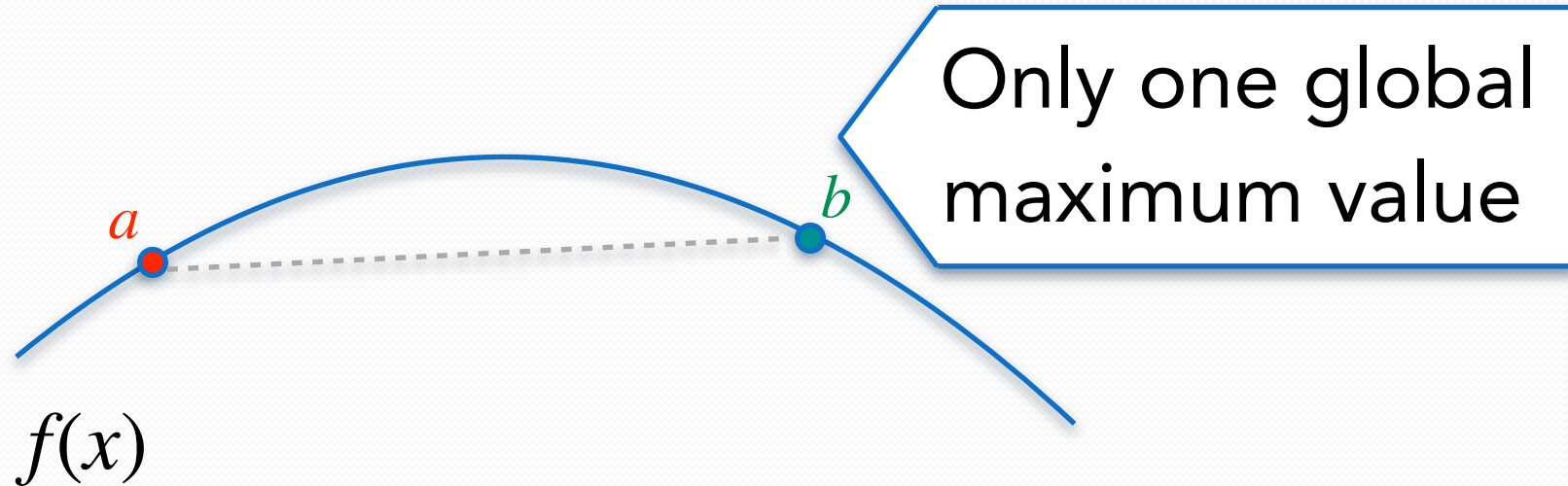
$$= \sum_{i=1}^N \ln \left[\underbrace{\sigma(\mathbf{w}^T \mathbf{x}^{(i)})^{y^{(i)}}}_{a^c} \underbrace{(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}))^{1-y^{(i)}}}_{b^d} \right]$$

$$\log a^c b^d = c \log a + d \log b$$

$$= \sum_{i=1}^N \left[\underbrace{y^{(i)} \ln \sigma(\mathbf{w}^T \mathbf{x}^{(i)})}_{c \log a} + \underbrace{(1 - y^{(i)}) \ln (1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}))}_{d \log b} \right]$$

□ How do we maximize the conditional likelihood?

Concave function



Not a concave function

