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

- 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/
- Some slides/approaches used are from Prof. Rangan
- Many approaches used are from CMU 18-661

Linear Classification &

Logistic Regression

PROF. LINDA SELLIE

EDITED BY HAORAN CHEN

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 estimation
- 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 also a confidence score of that classification ?
- ❑Maximum likelihood
  - How can we find the most likely hyperplane ?
  - 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_se

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

- 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:

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

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.

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.

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

Intuition

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

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

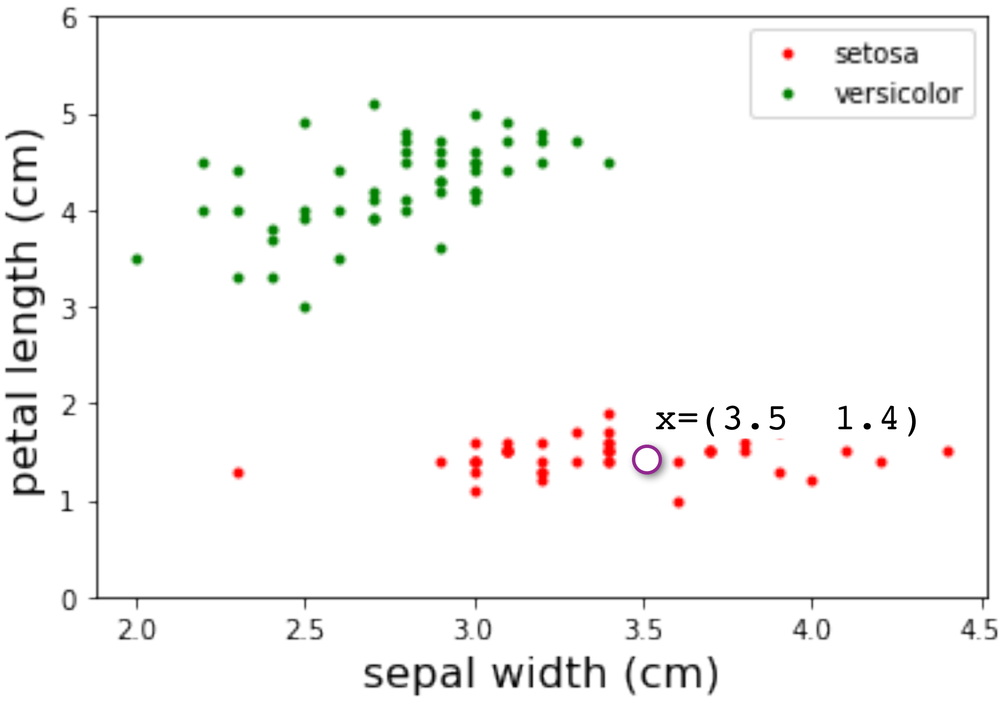
```
x=(sepal width, petal length)
```

❑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 ]
```



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

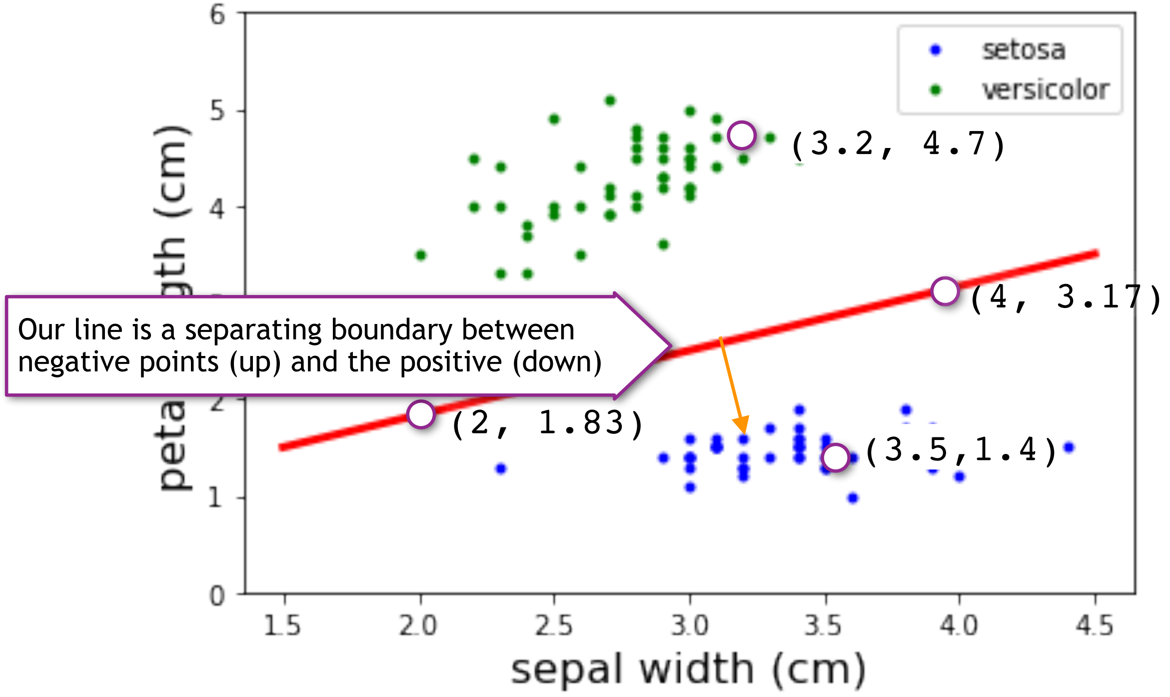
Next

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

Intuition: Decision Boundary

The line is: 0.5 + (2/3)sepal width + (-1) petal length = 0

Data: $\mathbf{x}^{(i)}$ = (sepal width$^{(i)}$, petal length$^{(i)}$)



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

❑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

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

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

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

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

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

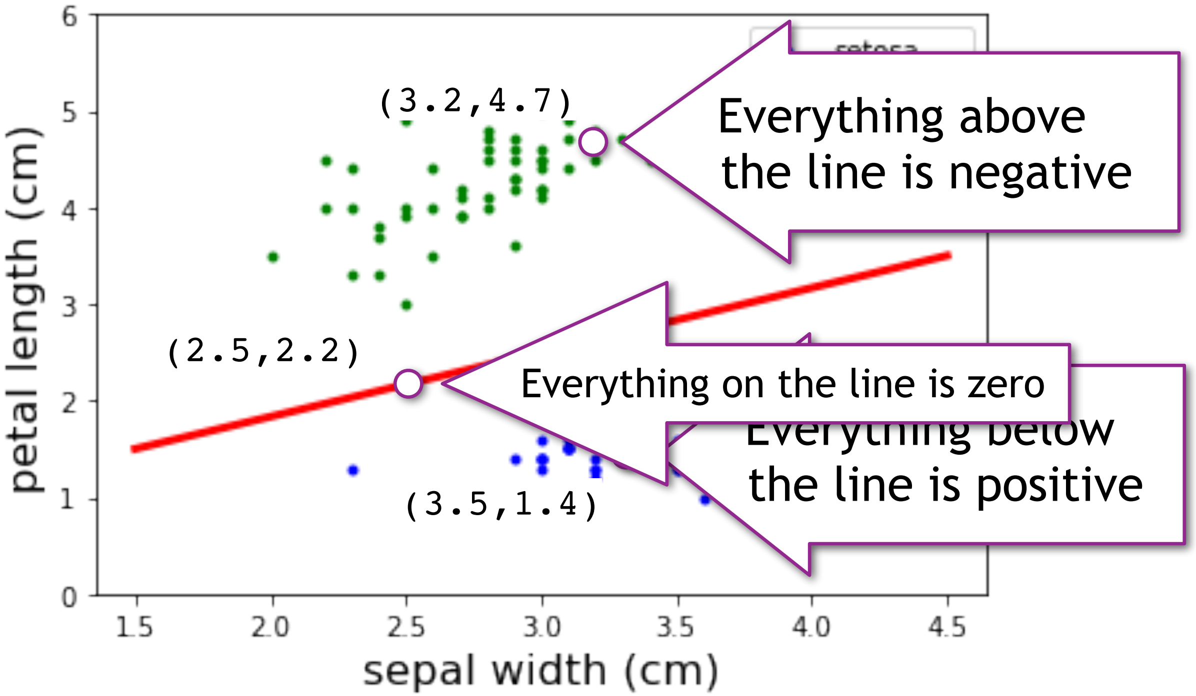
Linear Classifier

The line is: 0.5 + (2/3)sepal width + (-1) petal length = 0

- setosa Iris
- versicolor Iris

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

setosa Iris		versicolor Iris	
3.5	1.4	3.2	4.7
3.	1.4	3.2	4.5
3.2	1.3	3.1	4.9
3.1	1.5	2.3	4.
3.6	1.4	2.8	4.6
3.9	1.7	2.8	4.5
3.4	1.4	3.3	4.7
3.4	1.5	2.4	3.3
2.9	1.4	2.9	4.6
3.1	1.5	2.7	3.9



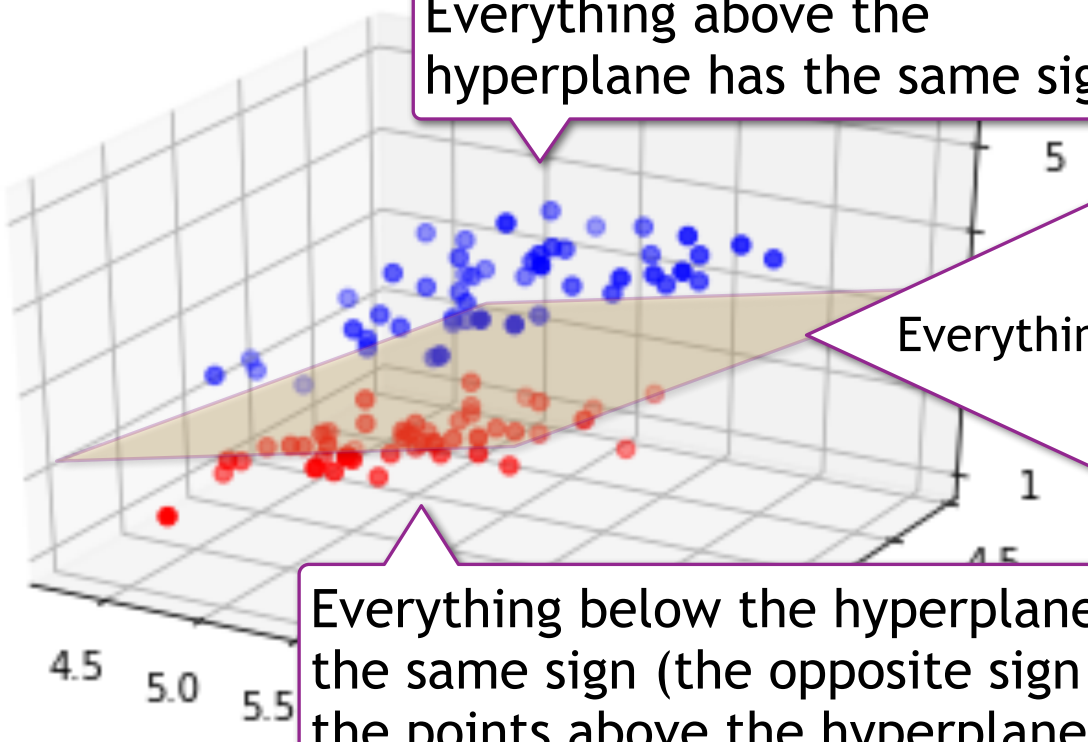
Pair share: What change would you make to have the separating line (hyperplane) in the same place, but to classify all the points labeled `positive' in the diagram as negative and all the points labeled `negative' in the diagram as positive?

Next

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)

Half-spaces:

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

Hyperplane:

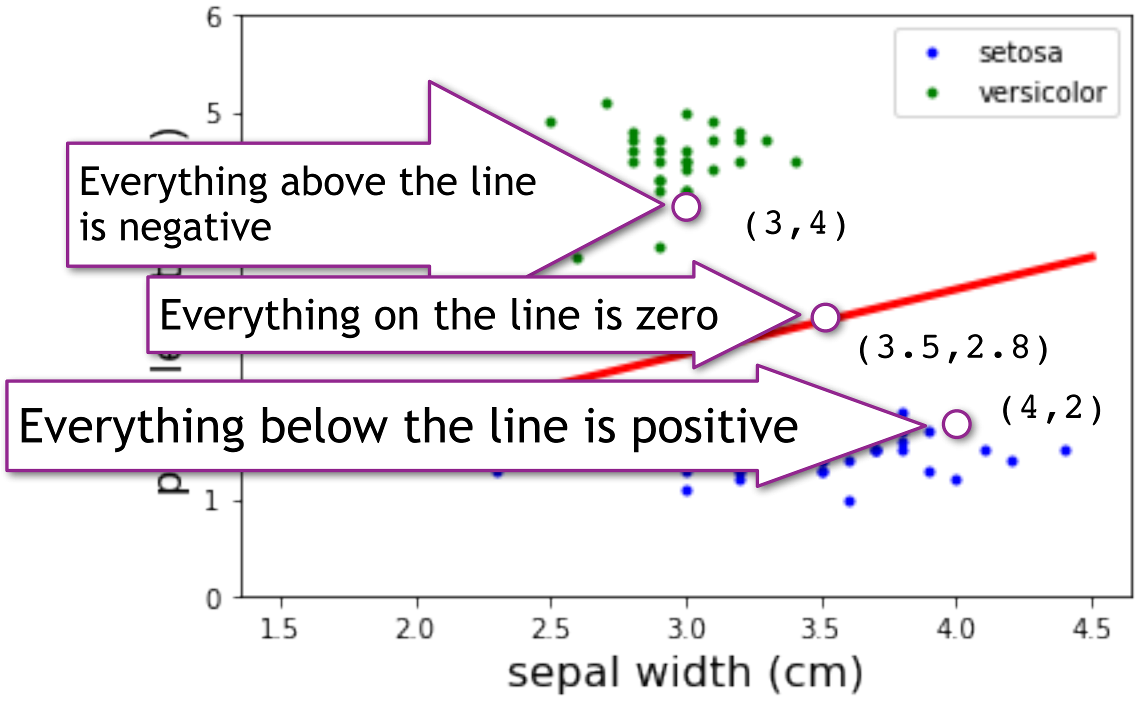
$$\mathcal{H} = \{\mathbf{x} : \mathbf{w}^T\mathbf{x} = 0\}$$

Half-spaces:

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

Prediction using a decision boundary

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



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

Setosa

Versicolor

Pair share:

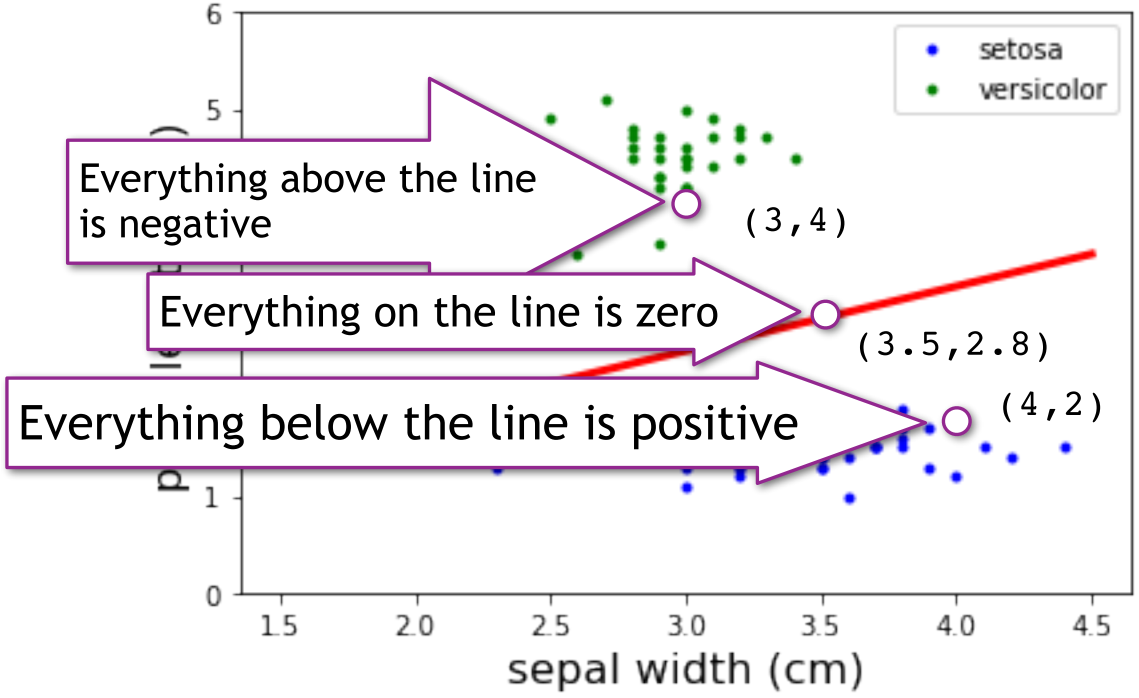
1. Suppose you found an iris with sepal width = 3 and petal length = 4.

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

Prediction using a decision boundary

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



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

Setosa

Versicolor

Pair share:

2. How can we predict the label of a new example ?

$$\mathbf{w} = \begin{bmatrix} 0.5 \\ 2/3 \\ -1 \end{bmatrix}$$

<u>Examples:</u> (3,4) (4,2)

$$\mathbf{x} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} \quad \mathbf{w}^T\mathbf{x}_2 = \begin{bmatrix} 0.5 & 2/3 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} = -1.5$$

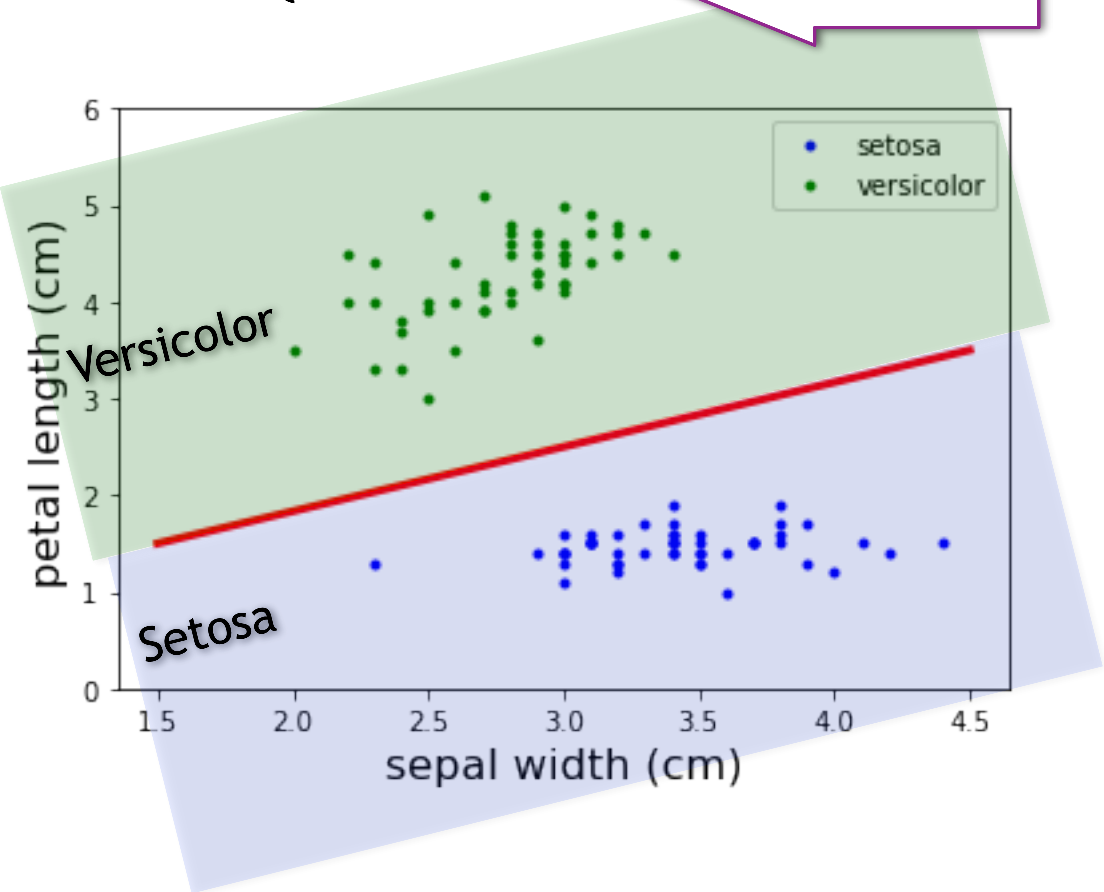
$$\mathbf{x} = \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} \quad \mathbf{w}^T\mathbf{x}_1 = \begin{bmatrix} 0.5 & 2/3 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 2 \end{bmatrix} = 7/6$$

Visualizing a linear classifier

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

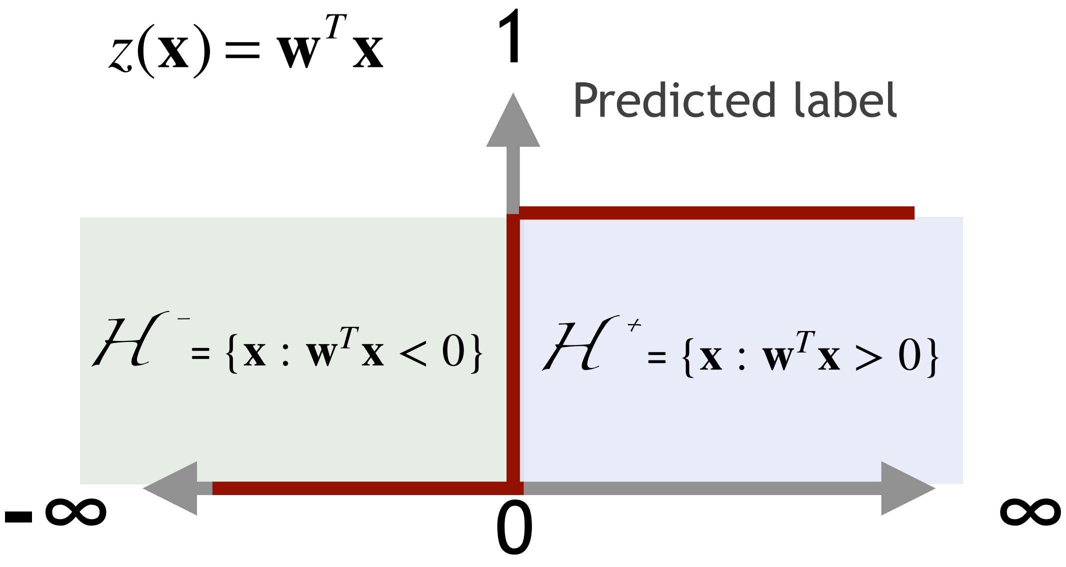
Setosa

Versicolor



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

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



Hyperplane:

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

Outline

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

Next

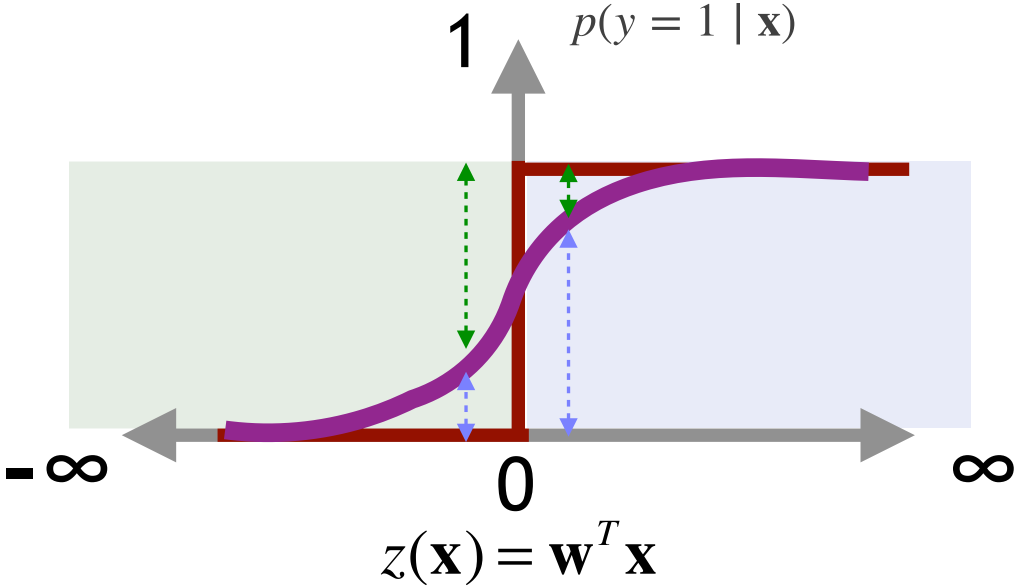
Could we modify the hypothesis to give more information about how confident we are in our prediction ?

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 $\mathbf{w}$

Logistic(Sigmoid) $\sigma(\cdot)$

Other functions could be used
- but this works well

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

-∞ 0 ∞

Squashing function

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

0 1

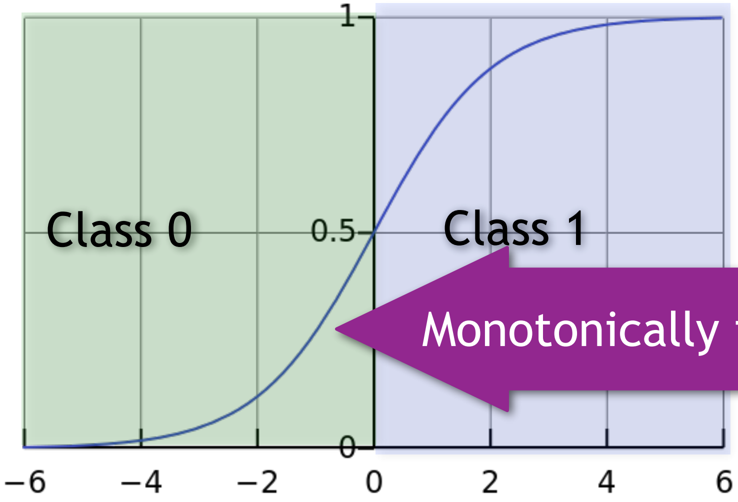
Pair share: Why is the output of σ is always in the interval (0, 1)? Why can't it equal 0 or equal 1? For what value of z does σ(z) = 0.5?

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

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

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

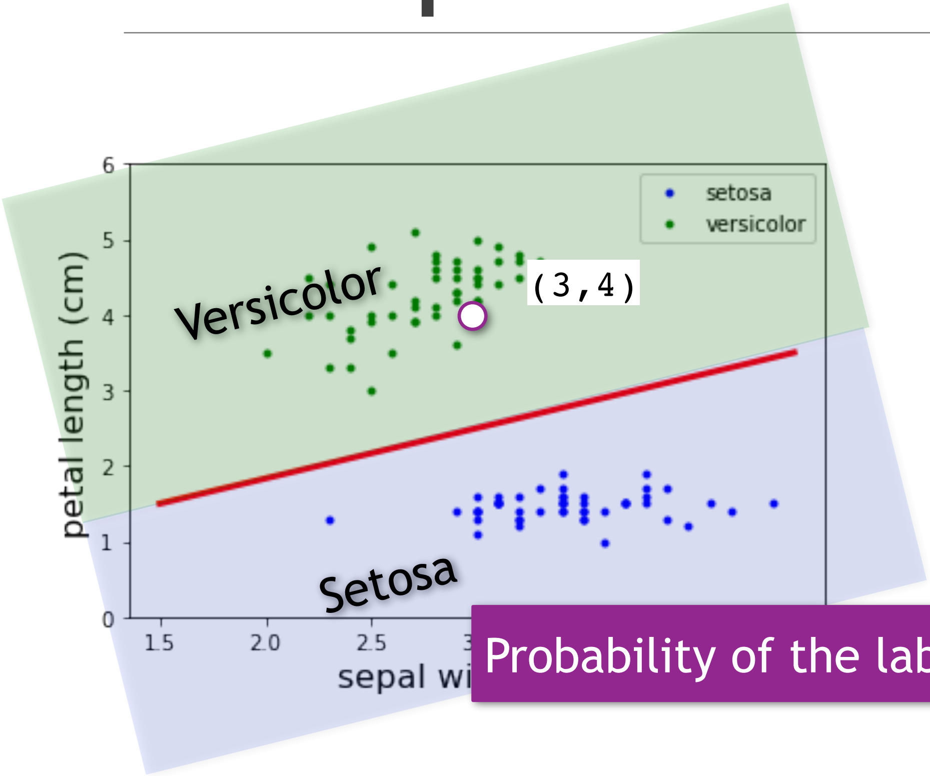
$\sigma(z)$ bounded between 0 and 1
Thus we can interpret as probability

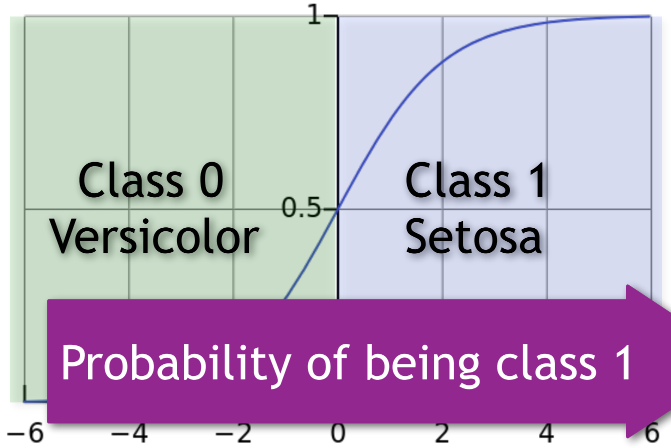


Monotonically increasing

Example

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





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

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

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

"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. "

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

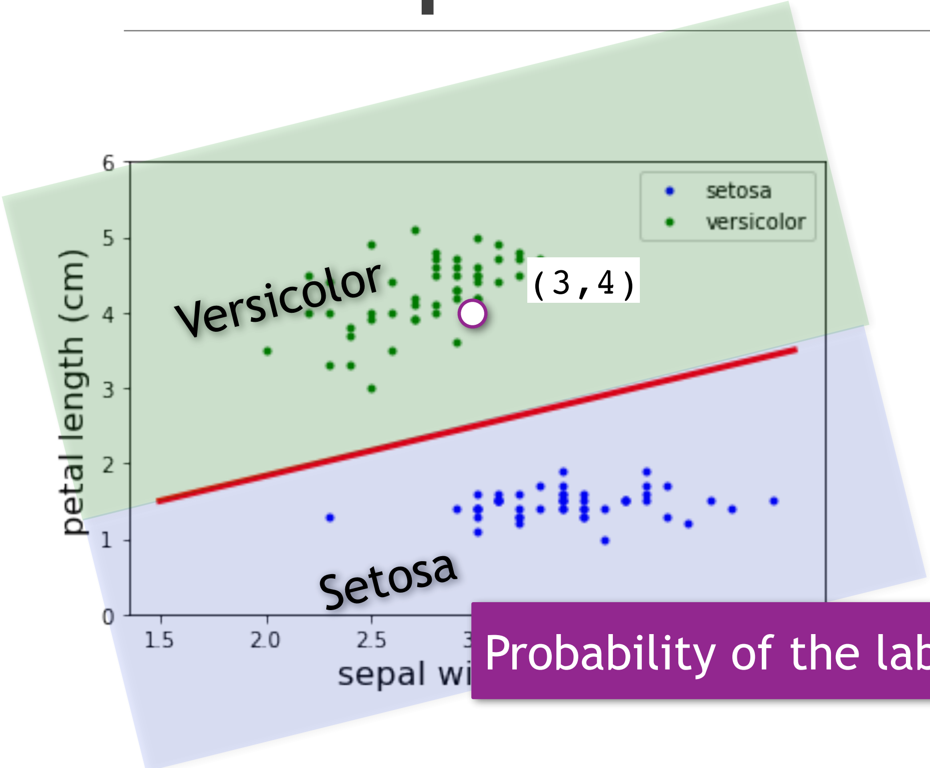
$$p(y = 0 \mid [1,3,4]^T; \mathbf{w}) = ?$$

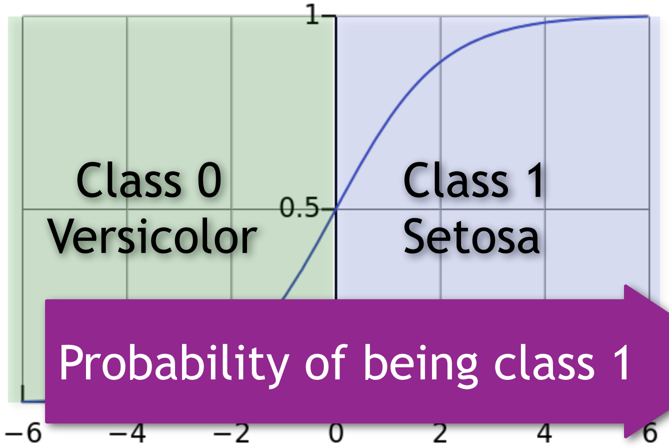
Pair share

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

Example

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





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

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

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

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

$$p(y = 0 \mid [1,3,4]^T; \mathbf{w}) = (.182)^0(1 - .182)^{1-0} = .718$$

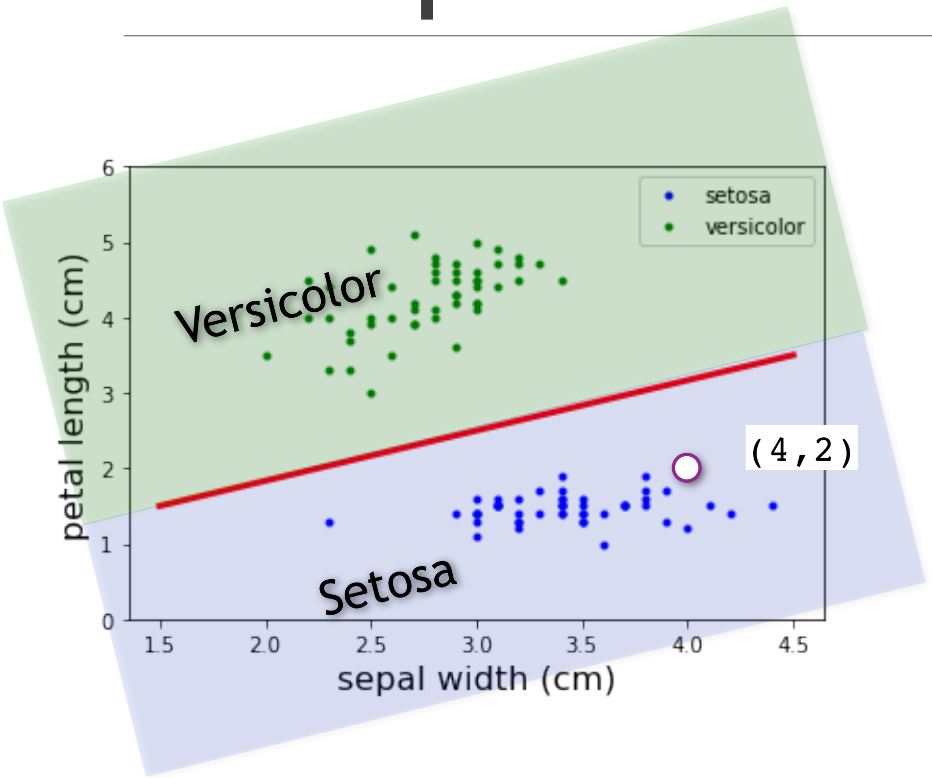
Pair share

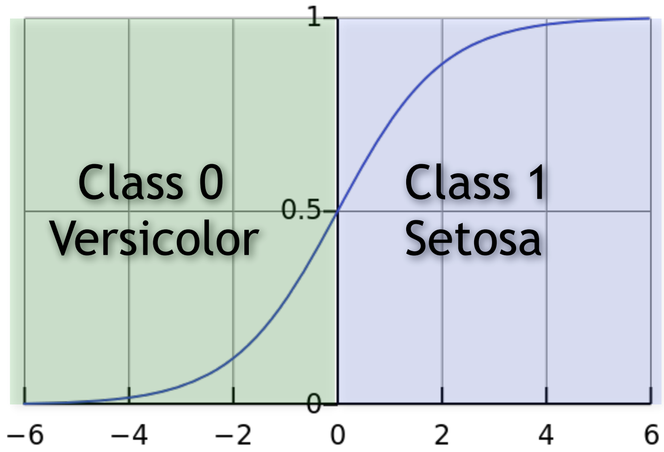
"Notational note: In the expression p(y|x; w) the semicolon indicates that w is a parameter, not a random variable that is being conditioned on, even though it is to the right of the vertical bar. "

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

Example

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





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

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

$$p(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x}^{(i)})^{y^{(i)}} \left(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)})\right)^{1-y^{(i)}}$$

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

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

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

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

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

$$p(y = 0 \mid [1,4,2]^T; \mathbf{w}) = (.763)^0 (1 - .763)^{1-0} = .237$$

Logistic Regression

Data: $(\mathbf{x}^{(i)}, y^{(i)}), i = 1,2,\ldots,N$ 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

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

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

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

Next we will show why this is true And find an optimizer to find the "best" w

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

Multiple linear regression

Outline

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

Which model ?

Optimizer

Finding an objective function

Next

Given $D = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(N)}, y^{(N)})\}$,

how can we find the "best" hyperplane, $\mathbf{w}$?

Optimize $\mathbf{w}$

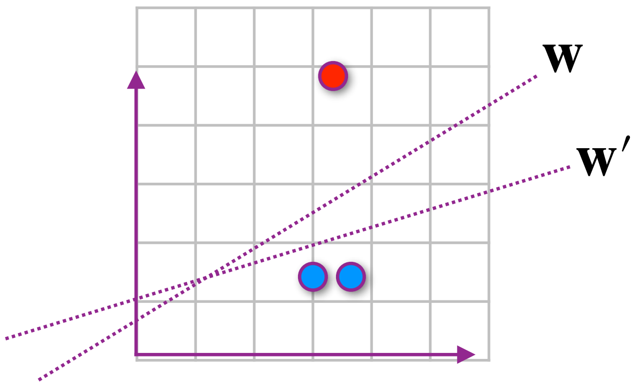
We first need to decide what makes one hyperplane better than another? (i.e. an objective function)

Pair share

Maximum Likelihood Estimation(MLE)

Likelihood of seeing data

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



- Given the following data:

$\mathbf{x}^{(1)} = [1,\ 3.2\ \ 4.7]\ \ y^{(1)} = 0$
$\mathbf{x}^{(2)} = [1,\ 3.5\ \ 1.4]\ \ y^{(2)} = 1$
$\mathbf{x}^{(3)} = [1,\ 3.0\ \ 1.4]\ \ y^{(3)} = 1$

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

$$\mathbf{w} = \begin{bmatrix} 1/2 \\ 2/3 \\ -1 \end{bmatrix} \qquad L(\mathbf{w}) = \left(1 - \frac{1}{1+e^{-(1/2+(2/3)3.2-4.7)}}\right)\left(\frac{1}{1+e^{-(1/2+(2/3)3.5-1.4)}}\right)\left(\frac{1}{1+e^{-(1/2+(2/3)3-1.4)}}\right) = 0.54$$

1-0.11 0.81 0.75

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

Classification Example

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

Our model:

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

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

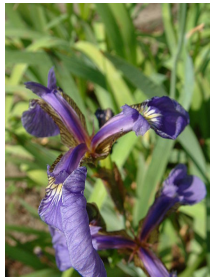
$$p(y = 0 \mid \mathbf{x}^{(i)}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}^{(i)}) = 1 - \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}^{(i)}}}$$

versicolor

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

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

setosa



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

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

$\mathbf{x}^{(3)}$ = [1, 3.0 1.4] $y^{(3)}$ = 1

Classification Example

versicolor:

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

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

setosa:

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

$\mathbf{x}^{(3)}$ = [1, 3.0 1.4] $y^{(3)}$ = 1

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

$$L(\mathbf{w}) = \left(1 - \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x}^{(1)})}}\right)\left(\frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x}^{(2)})}}\right)\left(\frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x}^{(3)})}}\right) = \left(1 - \frac{1}{1 + e^{-(w_0 + w_1 \cdot 3.2 + w_2 \cdot 4.7)}}\right)\left(\frac{1}{1 + e^{-(w_0 + w_1 \cdot 3.5 + w_2 \cdot 1.4)}}\right)\left(\frac{1}{1 + e^{-(w_0 + w_1 \cdot 3 + w_2 \cdot 1.4)}}\right)$$

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

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

The conditional likelihood function

Conditional likelihood function
(conditioned on **x**)
Larger value means more likely

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

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

Here we assume
all the examples are independent

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

$$L(\mathbf{w}) = \prod_{i:y^{(i)}=1} p(y^{(i)}=1 \mid \mathbf{x}^{(i)};\mathbf{w})^{y^{(i)}} \left(1 - p(y^{(i)}=1 \mid \mathbf{x}^{(i)};\mathbf{w})\right)^{1-y^{(i)}} \boxed{= \prod_{i=1}^{N} \sigma(\mathbf{w}^T\mathbf{x}^{(i)})^{y^{(i)}} \left(1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)^{1-y^{(i)}}}$$

How can we find the best $\mathbf{w}$?

Pair share

Can we maximizes this function ?

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

The Log-likelihood function

- We wanted to maximize $L(\mathbf{w}) = \prod_{i=1}^{N} \sigma(\mathbf{w}^T\mathbf{x}^{(i)})^{y^{(i)}} \left(1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)^{1-y^{(i)}}$
- This is the same as maximizing $\ell(\mathbf{w}) = \ln(L(\mathbf{w})) = \ln\left[\prod_{i=1}^{N} \sigma(\mathbf{w}^T\mathbf{x}^{(i)})^{y^{(i)}} \left(1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)^{1-y^{(i)}}\right]$

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

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

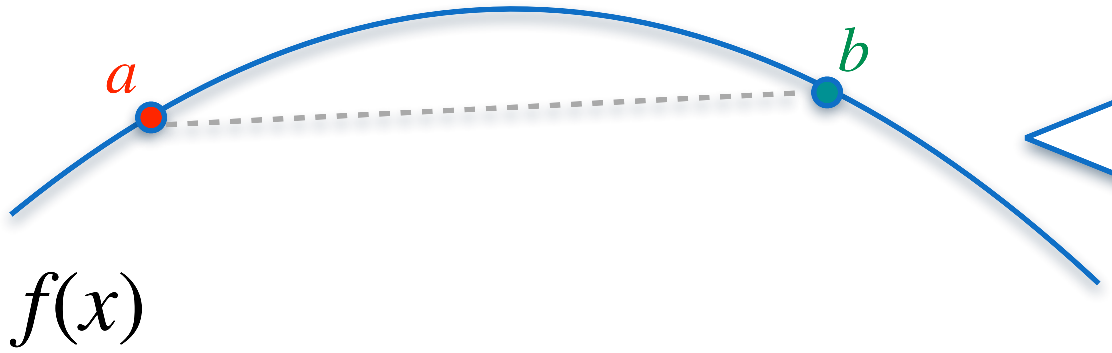
$$= \sum_{i=1}^{N} \left[\underbrace{y^{(i)} \ln \sigma(\mathbf{w}^T\mathbf{x}^{(i)})}_{c \log a} + \underbrace{(1 - y^{(i)}) \ln\left(1 - \sigma(\mathbf{w}^T\mathbf{x}^{(i)})\right)}_{d \log b}\right]$$

Define:

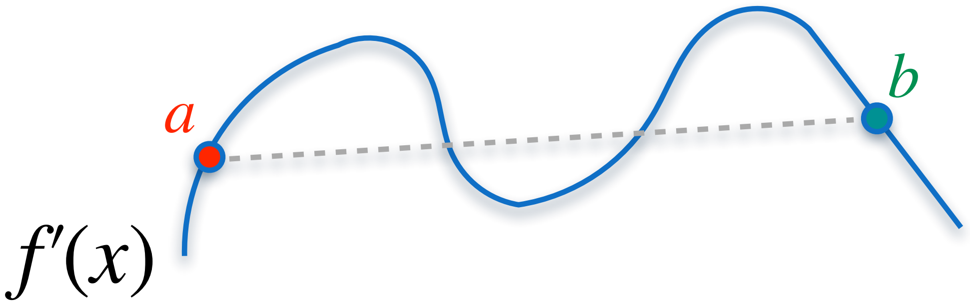
$$p(y = 1 \mid \mathbf{x}^{(i)}; \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}^{(i)}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}^{(i)}}}$$

Concave(Non-Concave) function

https://homes.cs.washington.edu/~marcotcr/blog/concavity



Concave **ONLY** one global maximum value



Non-Concave has **more than one** global maximum value