Linear Models

Machine Learning
Fall 2023
Checkpoint: The bigger picture

• Supervised learning: instances, labels, and hypotheses
Checkpoint: The bigger picture

- Supervised learning: instances, labels, and hypotheses

![Diagram showing the process of supervised learning](image)
Checkpoint: The bigger picture

- Supervised learning: instances, labels, and hypotheses

- Specific learners
  - Decision trees
  - Adaboost
  - Bagged trees
  - Random forests
  - ...

![Diagram showing the process of learning with labeled data, learning algorithm, and hypothesis/prediction.](image)
Checkpoint: The bigger picture

- Supervised learning: instances, labels, and hypotheses

- Specific learners
  - Decision trees
  - Adaboost
  - Bagged trees
  - Random forests
  - ...

- General concepts
  - Features as high dimensional vectors
  - Overfitting
  - PAC learnability

Questions?
Lecture outline

• Linear classifiers

• What functions do linear classifiers express?

• Least Squares Method for Regression
Where are we?

• Linear classifiers
  – Definition
  – Geometry of linear classifiers
  – A notational simplification
Which is the better classifier?

Suppose this is our training set and we have to separate the blue circles from the red triangles.
Which is the better classifier?

Suppose this our training set and we have to separate the blue circles from the red triangles.
Which is the better classifier?

Suppose this our training set and we have to separate the blue circles from the red triangles.
Which is the better classifier?

Suppose this our training set and we have to separate the blue circles from the red triangles.
Which is the better classifier?

Suppose this our training set and we have to separate the blue circles from the red triangles.

Think about overfitting.

Which curve runs the risk of overfitting?
Which is the better classifier?

Suppose this our training set and we have to separate the blue circles from the red triangles.

Think about overfitting.

Which curve runs the risk of overfitting?
Similar argument for regression

Linear regression might make smaller errors on new points
Similar argument for regression

Linear regression might make smaller errors on new points
Similar argument for regression

Linear regression might make smaller errors on new points
Similar argument for regression

Linear regression might make smaller errors on new points

\[ F(x) \]
Linear Classifiers

Input is a n dimensional vector $x$
Output is a label $y \in \{-1, 1\}$

*Linear Threshold Units* classify an example $x$ using a weight vector $w$ and $b$ (a real number) according to the following classification rule

\[
\text{Output} = \text{sign}(w^T x + b) = \text{sign}(b + \sum w_i x_i)
\]

\[
\begin{align*}
w^T x + b & \geq 0 \implies \text{Predict } y = 1 \\
w^T x + b & < 0 \implies \text{Predict } y = -1
\end{align*}
\]

$b$ is called the bias term
The geometry of a linear classifier
The geometry of a linear classifier
The geometry of a linear classifier

\[ \text{sgn}(b + w_1 x_1 + w_2 x_2) \]
The geometry of a linear classifier

\[ \text{sgn}(b + w_1 x_1 + w_2 x_2) \]
The geometry of a linear classifier

\[ \text{sgn}(b + w_1 x_1 + w_2 x_2) \]

\[ b + w_1 x_1 + w_2 x_2 = 0 \]
The geometry of a linear classifier

\[ \text{sgn}(b + w_1 x_1 + w_2 x_2) \]

\[ b + w_1 x_1 + w_2 x_2 = 0 \]
The geometry of a linear classifier

\[ \text{sgn}(b + w_1 x_1 + w_2 x_2) \]

We only care about the sign, not the magnitude.
The geometry of a linear classifier

\[ \text{sgn}(b + w_1 x_1 + w_2 x_2) \]

In \( n \) dimensions, a linear classifier represents a hyperplane that separates the space into two half-spaces.

We only care about the sign, not the magnitude.

Questions?
Simplifying notation

We can stop writing $b$ at each step using notational sugar:

The prediction function is $\text{sgn}(b + w^T x)$

Rewrite $x$ as $[1, x] = x'$

Rewrite $w$ as $[b, w] = w'$

Increases dimensionality by one

Equivalent to adding a feature that is always 1
Simplifying notation

We can stop writing $b$ at each step using notational sugar:

The prediction function is $\text{sgn}(b + w^T x)$

Rewrite $x$ as $[1, x] = x'$

Rewrite $w$ as $[b, w] = w'$

- Increases dimensionality by one
- Equivalent to adding a feature that is always 1

The prediction is now $\text{sgn}(w'^T x')$
Simplifying notation

We can stop writing $b$ at each step using notational sugar:

The prediction function is $\text{sgn}(b + w^T x)$

Rewrite $x$ as $[1, x] = x'$

Rewrite $w$ as $[b, w] = w'$

Increases dimensionality by one

Equivalent to adding a feature that is always 1

The prediction is now $\text{sgn}(w'^T x')$

We sometimes fold the bias term $b$ into the input by adding an extra constant feature. But remember that it is there
Coming up (next several weeks): Linear classification

- **Perceptron**: Error-driven learning, updates the hypothesis if there is an error

- **Support Vector Machines**: Define a different cost function that includes an error term and a term that targets future performance (structural risk minimization)

- **Naïve Bayes classifier**: A simple linear classifier with a probabilistic interpretation (generative)

- **Logistic regression**: Another probabilistic linear classifier (discriminative)

In all cases, the prediction will be done with the same rule:

\[ \mathbf{w}^T \mathbf{x} + b \geq 0 \implies \text{Predict } y = 1 \]
\[ \mathbf{w}^T \mathbf{x} + b < 0 \implies \text{Predict } y = -1 \]
Regression vs. Classification

• Linear regression is about predicting real valued outputs

• Linear classification is about predicting a discrete class label
  – +1 or −1
  – SPAM or NOT-SPAM
  – Or more than two categories
Where are we?

• Linear classifiers: Introduction

• What functions do linear classifiers express?
  – Conjunctions and disjunctions
  – m-of-n functions
  – Not all functions are linearly separable
  – Feature space transformations
  – Exercises

• Least Squares Method for Regression
Which Boolean functions can linear classifiers represent?

• Linear classifiers are an expressive hypothesis class

• Many Boolean functions are **linearly separable**
  – Not all though
  – **Recall**: Decision trees can represent any Boolean function
Conjunctions and disjunctions

\[ y = x_1 \land x_2 \land x_3 \]

<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Conjunctions and disjunctions

$y = x_1 \land x_2 \land x_3$ is equivalent to “$y = 1$ if $x_1 + x_2 + x_3 \geq 3$”

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Conjunctions and disjunctions

$y = x_1 \land x_2 \land x_3$ is equivalent to “$y = 1$ if $x_1 + x_2 + x_3 \geq 3$”

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$y$</th>
<th>$x_1 + x_2 + x_3 - 3$</th>
<th>sign</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Conjunctions and disjunctions

\[ y = x_1 \land x_2 \land x_3 \] is equivalent to “\( y = 1 \) if \( x_1 + x_2 + x_3 \geq 3 \)”

<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( y )</th>
<th>( x_1 + x_2 + x_3 - 3 )</th>
<th>sign</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

How about negations?

\[ y = x_1 \land x_2 \land \neg x_3 \]
Conjunctions and disjunctions

\[ y = x_1 \land x_2 \land x_3 \] is equivalent to “\[ y = 1 \text{ if } x_1 + x_2 + x_3 \geq 3 \]”

Negations are okay too. In general, use 1-x in the linear threshold unit if x is negated

\[ y = x_1 \land x_2 \land \neg x_3 \]

is equivalent to

\[ y = 1 \text{ if } x_1 + x_2 + 1 - x_3 \geq 3 \]
Conjunctions and disjunctions

\[ y = x_1 \land x_2 \land x_3 \] is equivalent to “\( y = 1 \text{ if } x_1 + x_2 + x_3 \geq 3 \)”

<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( y )</th>
<th>( x_1 + x_2 + x_3 - 3 )</th>
<th>( \text{sign} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Negations are okay too. In general, use 1-x in the linear threshold unit if x is negated

\[ y = x_1 \land x_2 \land \neg x_3 \]

is equivalent to

\[ y = 1 \text{ if } x_1 + x_2 - x_3 \geq 2 \]
Conjunctions and disjunctions

\[ y = x_1 \land x_2 \land x_3 \] is equivalent to “\( y = \begin{cases} 1 & \text{if } x_1 + x_2 + x_3 \geq 3 \end{cases} \)”

<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( y )</th>
<th>( x_1 + x_2 + x_3 - 3 )</th>
<th>sign</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Negations are okay too. In general, use \( 1-x \) in the linear threshold unit if \( x \) is negated

\[ y = x_1 \land x_2 \land \neg x_3 \]

is equivalent to

\[ y = \begin{cases} 1 & \text{if } x_1 + x_2 - x_3 \geq 2 \end{cases} \]

Exercise: What would the linear threshold function be if the conjunctions here were replaced with disjunctions?
Conjunctions and disjunctions

\[ y = x_1 \land x_2 \land x_3 \text{ is equivalent to } "y = 1 \text{ if } x_1 + x_2 + x_3 \geq 3" \]

<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( y )</th>
<th>( x_1 + x_2 + x_3 - 3 )</th>
<th>( \text{sign} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Negations are okay too. In general, use \( 1 - x \) in the linear threshold unit if \( x \) is negated.

\[ y = x_1 \land x_2 \land \neg x_3 \]

is equivalent to

\[ y = 1 \text{ if } x_1 + x_2 - x_3 \geq 2 \]

**Exercise**: What would the linear threshold function be if the conjunctions here were replaced with disjunctions?

Questions?
m-of-n functions

m-of-n rules

• There is a fixed set of n variables
• \( y = \text{true} \) if, and only if, at least m of them are \text{true} 
• All other variables are ignored

Suppose there are three Boolean variables: \( x_1, x_2, x_3 \)

What is a linear threshold unit that is equivalent to the classification rule “at least 2 of \( \{x_1, x_2, x_3\} \)”?
m-of-n functions

m-of-n rules
• There is a fixed set of n variables
• \( y = \text{true} \) if, and only if, at least m of them are \text{true}
• All other variables are ignored

Suppose there are three Boolean variables: \( x_1, x_2, x_3 \)

What is a linear threshold unit that is equivalent to the classification rule “at least 2 of \{x_1, x_2, x_3\}”?

\[ x_1 + x_2 + x_3 \geq 2 \]
**m-of-n functions**

**m-of-n rules**
- There is a fixed set of n variables
- \( y = \text{true} \) if, and only if, at least m of them are true
- All other variables are ignored

Suppose there are three Boolean variables: \( x_1, x_2, x_3 \)

What is a linear threshold unit that is equivalent to the classification rule “at least 2 of \( \{x_1, x_2, x_3\} \)”?

\[ x_1 + x_2 + x_3 \geq 2 \]

Questions?
Parity is not linearly separable

Not all functions are linearly separable

Can’t draw a line to separate the two classes

Questions?
Not all functions are linearly separable

• XOR is not linear
  – \[ y = (x_1 \lor x_2) \land (\neg x_1 \lor \neg x_2) \]
  – **Parity** cannot be represented as a linear classifier
    • \( f(x) = 1 \) if the number of 1’s is even

• Many non-trivial Boolean functions
  – \[ y = (x_1 \land x_2) \lor (x_3 \land \neg x_4) \]
  – The function is not linear in the four variables
Even these functions can be *made* linear

These points are not separable in 1-dimension by a line

What is a one-dimensional line, by the way?

The trick: Change the representation
The blown up feature space

The trick: Use feature *conjunctions*

Transform points: Represent each point $x$ in 2 dimensions by $(x, x^2)$
The blown up feature space

The trick: Use feature *conjunctions*

Transform points: Represent each point \( x \) in 2 dimensions by \((x, x^2)\)
The blown up feature space

The trick: Use feature *conjunctions*

Transform points: Represent each point $x$ in 2 dimensions by $(x, x^2)$

Now the data is linearly separable in this space!
The blown up feature space

The trick: Use feature *conjunctions*

Transform points: Represent each point \( x \) in 2 dimensions by \( (x, x^2) \)

Now the data is linearly separable in this space!
Exercise

How would you use the feature transformation idea to make XOR in two dimensions linearly separable in a new space?

To answer this question, you need to think about a function that maps examples from two dimensional space to a higher dimensional space.
Almost linearly separable data

$$\text{sgn}(b + w_1 x_1 + w_2 x_2)$$

Training data is almost separable, except for some noise

How much noise do we allow for?
Almost linearly separable data

\[ \text{sgn}(b + w_1 x_1 + w_2 x_2) \]

Training data is almost separable, except for some noise

How much noise do we allow for?
Linear classifiers: An expressive hypothesis class

• Many functions are linear

• Often a good guess for a hypothesis space

• Some functions are not linear
  – The XOR function
  – Non-trivial Boolean functions

• But there are ways of making them linear in a higher dimensional feature space
Why is the bias term needed?

\[ b + w_1 x_1 + w_2 x_2 = 0 \]
Why is the bias term needed?

If $b$ is zero, then we are restricting the learner only to hyperplanes that go through the origin.

May not be expressive enough.
Why is the bias term needed?

If $b$ is zero, then we are restricting the learner only to hyperplanes that go through the origin. May not be expressive enough.